MATS UNIVERSITY

NAAC GRADE A+ ACCREDITED UNIVERSITY

MATS CENTRE FOR DISTANCE & ONLINE EDUCATION

Business statistics

Master of Business Administration (MBA)
Semester - 1

SELF LEARNING MATERIAL

ODL/MSMSR/MBA/104
Business Statistics

BUSINESS STATISTICS

COURSEDEVELOPMENTEXPERTCOMMITTEE

1. Prof. (Dr.) Umesh Gupta, Dean, School of Business & Management Studies, MATS University, Raipur, Chhattisgarh
2. Prof. (Dr.) Ashok Mishra, Dean, School of Studies in Commerce & Management, Guru Ghasidas University, Bilaspur,Chhattisgarh
3. Dr. Madhu Menon, Associate Professor, School of Business & Management Studies, MATS University, Raipur,Chhattisgarh
4. Dr. Nitin Kalla, Associate Professor, School of Business & Management Studies, MATS University, Raipur,Chhattisgarh
5. Mr.Y.C.Rao,CompanySecretary,GodavariGroup,Raipur,Chhattisgarh

COURSECOORDINATOR

Dr. Premendra Sahu, Assistant Professor, School of Business & Management Studies, MATS University, Raipur, Chhattisgarh

COURSE/BLOCKPREPARATION

Dr. V. SureshPillai

AssistantProfessor

MATSUniversity,Raipur,Chhattisgarh

ISBN-978-93-49954-11-3

March,2025

Printed&publishedonbehalfofMATSUniversity,Village-Gullu,Aarang,RaipurbyMr. MeghanadhuduKatabathuni,Facilities&Operations,MATSUniversity,Raipur(C.G.)

Printedat:TheDigitalPress,KrishnaComplex,Raipur-492001(Chhattisgarh)

Acknowledgement

The material (pictures and passages) we have used is purely for educational purposes. Every effort has been made to trace the copyright holders of material reproduced in this book.Should any infringement have occurred, the publishers and editors apologize and will be pleased to make the necessary corrections in future editions of this book.

MODULE1INTRODUCTIONTOSTATISTICS

Structure

UNIT1.1 Meaningand Definition ofStatistics

UNIT1.2 Scopeand Importanceof Statistics

UNIT1.3 TypesofStatistics(DescriptiveandInferential)

UNIT1.4 FunctionsandLimitationsof Statistics

UNIT1.5 MeasuresofCentral Tendency

UNIT1.6 MeasuresofDispersion

UNIT1.7 Skewness and Kurtosis

UNIT1.8 Index Numbers

OBJECTIVES

- Explainthefundamentalconceptanddefinitionofstatistics.
- Identifythesignificanceandapplicationsofstatisticsinvariousfields.
- Distinguishbetweendescriptiveandinferentialstatisticswithexamples.
- Discussthekeyfunctionsandconstraintsofstatisticalmethods.
- Calculate and assess range, interquartile range, mean deviation, standard deviation, variance, & variation. coefficient
- Define,measure,&analyzeskewnessandkurtosisinstatisticaldistributions.
- Explainthemeaning,importance,types,andapplicationsofindex numbers in real-world scenarios**.**

UNIT1.1MEANINGANDDEFINITIONOFSTATISTICS

1.1MeaningAndDefinitionOfStatistics

A crucial tool throughwhich to capture the shades of complexity in the ever-complexworldoutsideus,and turndataintosomethingyoucanmeaningfully apply. Between the abstraction of the beautiful theorem and the vaguely disordered world of example, there is the data trained on us, on the limits of our upping creation, which makes it simple for us to prove our own deductions. In a nutshell, statisticsis the language of data, is a means used to develop a strategy to quantify uncertainty or rather to make informed decisions under condition that everything is not perfect. It allows us to compresshugevolumesofdataintosmallandinterpretableforms,toidentify significant differences among populations, to model complex interactions between inputs, and to calculate the probability of various outcomes.

Figure1.1:MeaningandDefinitionofStatistics.

Statistics help us to rise above personal testimonies, biases and emotions to help ground our discussions and debates in evidence-based and data-driven arguments. Test their statistical interpretation after learning that statistics are fundamentally about interpreting data, finding patterns or relationships, and predictingdevelopments ortrendsineventsbasedonwhatisindicatedbythe data. Emit error Not only allows a bunch of formulas and calculations,but is also a highly disciplined, logical approach to arrive at a solution based on mathematical principles applied in disciplines such as science, business, economics, social sciences, medicine, engineering and many more. Statistics has everything from simple descriptive measures like the mean andpercentiles to more sophisticated inferential techniques that allow drawing insights about entire populations based on the data of only a sample. Mathematicsisallaboutuncertaintyandmakingsenseofthisuncertaintyto

make better decisions. The field encompasses a wide range of methods, uncertainty. Statistics provides us with tools to quantify the uncertainty we experience in a complex and changing world theworld we find ourselves in.

StatisticsasNumericalData:QuantitativeRepresentationofPhenomena

Relevance in the consideration of the limitations of statistical data, and critically discussing the validity and reliability of the collected and correctly analyzed one. numbers by itself have nocontext in it so as one can understand the story behind it. Statistics need to be understood in context and, critically, they were judgments. That of course, over time, to compare different groups or areas, and to identify trends and patterns." Whilemaking statistical inference, we can use our quantitative reasoning skills and search for something beyond gutfeelingsandprescriptiveargumentationandmakethemthebasisfora clear and objective data-driven story about our world. An excellent introduction to statistics as numerical dataare important, thesetell us aleaves a wayforthem. These statistics can takeseveral forms, such as student enrollment, graduation rates or standardized test scores. In all these cases, you have objective and quantitative data points about the events being studied (that deaths, and treatment effectiveness statistics. For example: Findingseducational statistics and GDP could be cross walked. Medical statistics, on the other hand, include diseases, also casting decisions you make. From the point of view of economics, these economic statistics can also be processed simply and objectively, such as inflation rates, unemployment rates, and some characteristics of a phenomenon. Measurements are be expressed as counts, measurements, percentages, ratios, or rates. They can also summarize and compare diverse information.

DefinitionsbyEminentStatisticians:DiversePerspectivesontheDiscipline

Manystatisticianstriedtodefinewhattheydidovertheyearstotheirparticular viewpoint andfield. It shows the different roles of statistics in variousfields and its transformation till now.

Figure1.2:StatisticsasNumericalData:QuantitativeRepresentationofPhenomena.

- **A.L.Bowley:** "It can be rightly said "Statistics is the science of average. This all sounds familiar, we have had similar exposure to a data definition: Averages are a basic concept from statistics, but this is a somewhat narrow definition anddoesn't capture the entirety of the field.
- **YuleandKendall**: "Statistics are numerical statements of facts inany department of inquiry placed in relation to each other." As such this definition places importance on context and relationships in a statistical analysis. Statistical data is not just a number abstracted from all the others, ratherit becomes meaningfully when put into comparison with other data.
- **Croxton&Cowden**: "Statistics is science of collection, presentation, analysis and interpretation of numerical data." This definition envisions you statisticallyasyoureacheverysingleend-user processstartingfrom extraction of data to finally prediction. Now it is considered to be a more accurate and more representative definitionof the discipline.
- **R.A.Fisher**: "Statistics may be regarded as(i) populations, study (ii) study variability, (iii) study of the reduction of data. Statistics is a science concernedwith populations, variability, as well asdata reduction, accordingto Fisher. He was widely regardedas one of the founding fathers of statistics due to his contributions to the field.
- **C.R.Rao:** "Statistics is a branch of science dealing with the collection, analysis, interpretation and presentation of empirical data and providing

- methodsformakingrationaldecisioninthepresenceofuncertainty.Rao's definitionfocuses on decision making and uncertainty.
- **MauriceKendall**: "Statistics is the branch of scientific method which deals with the data obtained by counting or measuring the properties of populations of natural phenomena, and which develops methods for the collection, classification, analysis and interpretation of such data." This definition emphasizes the methodology and the importance of the accumulated data and thesaurus.

Each of these definitions offers a varying perspective of the same thing alongside the numerical data itself, statistics also encompass the methods we use to analyzethese data and the techniques we and apply to derive meaning fromthedatathatwehavecollected.Theyemphasizetherelevanceofcontext, relationships and uncertainty to statistical analysis. Each definition brings anew flavor in explaining the use of data to provide insight or informed decisions.

EvolutionoftheDefinition:AdaptingtoModernApplications

Statistics is broadening in its application, and, as our understanding of the discipline has evolved, so has the definition. In the early days, statistics primarily involved the collection and summarization of numerical data, primarilyforgovernmentalandadministrativeaid purposes.However,thefield of statistics has been extended remarkably as better statistical tools have come up along with the increasing data available. Statistics are everywhere these days, from scientific experiments and business analytics to public policy and health care. There have been changes in the field itself with the introduction of big data and machine learning, where new statistical methods are being developed to cope with large datasets and to identify complex patterns. Therefore, statistics is a vast domain and still has a redefinition ofstatistics. Recent definitions include computer and computational methods, the ability to manage large, complex data sets, and also theemphasis placed on prediction and decision making. recognized as a fundamental lens for understanding and addressingthe complexities of today's world.

UNIT1.2SCOPEANDIMPORTANCEOFSTATISTICS

Statistics, the field that deals with collecting, organizing, analyzing,interpreting and presenting data, is embedded in virtually every part of modern life. It goes far beyond numbers, trends,graphs, aggregated for dataset-based decisions and innovations. Statistics is a fundamental tool used in nearly every aspect of life, from scientific research to business and government operationsto navigate the uncertainty and find meaningful patterns in thelarge amountsof data generated. It leverages the raw data to create information that enablesus to perceive, comprehend, predict patterns and trends, and to evaluatewhether the actions we take are working or not.

1.2ScopeAndImportanceOfStatistics

ScientificResearchandExperimentation: Scientific research and experimentation, which becomes significant statistical significance and hypothesis testing. Hypothesis generation and statistical analysis of experimental data and determination of statistical significance of the resultant effects. Researchers can apply methods such as hypothesis testing, multivariate regression, and analysis of variance, at least to support the objectivity of their interpretations and to quantify the uncertainty in the results. Essentially, statistical analysis is essential to furthering understanding and formulating evidencebased practices across all domains, from medicine to biology, physics and the social sciences. Like for statistical analyses that are conducted in clinical trials of new drugs, or treatments and ecological studies of statistical models that assess the population dynamics and environmental changes. In other words, Science Statistics (Stats) does something else: it challenges the core (implict) dogmas, and then: science becomes harder to manipulate and tendentious, it becomes more robust and repeatable.

BusinessandEconomics: In the competitive world of Business; Statistics forms the backbone of taking the right decisions, across market analysis and enhancing operational effectiveness. Statistical tools help companies forecast sales, analyze customer behavior, order inventory and analyze financial risk. Theycanalsoincludemarketresearchbasedonsamplingtechniquesand

statistical surveys asused by businesses to study consumer preferences, market trends and competitive landscapes. Econometrics, stands out as a powerful tool that aids economists in applying statistical theories to economic data, thereby establishing economic relationships, forecasting potential changes in financial markets, and evaluating the impact of economic policies. SPC techniques are applied in manufacturing for quality control of the products, reduction inproduct defects and increase in productivity. Furthermore, banks and other financial institutions utilize statistical modeling toassess the credit risk of loan applicants, to fine-tune investment portfolios, and for detecting fraudulent activities. Statistics is a very useful method applied in many areas, such as business and economics.

GovernmentandPublicPolicy: Statistics are crucial for governments at all levels so they can make evidence-based decisions while assessing policies, distributing resources, and tracking the status of their citizens. Population Statistics National statistical agencies are responsible for the collection and dissemination of data on the demographics of the population, economic indicators, health statistics, and social trends. These data inform the assessment ofthe success of public programs, highlight areas of need, and is help produce evidence-based policies. Census data, for instance, are critical to redistricting,the distribution of federal funding and the planning of infrastructureconstruction. A statisticalof the disease which they track to help monitor that vaccination rates and assess the impact of public health interventions. Next we use GDP, zero unemployment, and inflation etc. Withoutpolice or crime data, crime statistics are used to analyze Crime and law enforcement patterns and trends, evaluate law enforcement strategies and that identify programs for the prevention of crime. Statistical data is important for the government and public policy asit helps to enable the government and its activities by increasing the accountability and transparency in how government administers its business which ultimately leads to better governance.

SocialSciencesandHumanities: Statistics is also an important aspect of studying human behavior, social interactions, and cultural phenomena in the socialsciences.Statisticaltechniquesareappliedtosurveydata,experiments,

and hypotheses concerning social and psychologicalmechanisms. Sociologists use statistical techniques to conduct studies about social stratification and inequalityanddemographic trends.Psychologistswithstatistics mean distilledpsychologystudies.UnlikeTomClancynovels,votersarestatistically analyzed and modeled like any other scientific variables political scientists' model in their political, social, and scientific models. Statistical methods are now being wielded more sharply in thehumanities to make sense of large data sets of texts, images and other cultural objects. Historical subfields synthesize data through statistical methods (e.g., text mining, network analysis), and digital humanities initiatives consume large amounts of data from historical documents, literary works, and artwork. Researchers apply statistics to the social sciences and humanities, using quantitative methods toreveal trends in the data that are hidden from plain view, to test theoretical models, and to deepen our understanding of the human experience.

HealthcareandMedicine: Statistics is vital tomany aspects of healthcare and medicine such as clinical trials and epidemiology. Statistical methods are central to the design and analysis of clinical trials, evaluation of the efficacy and safety of new treatments, and identification of risk factors for many conditions for medical researchers. Epidemiologists specializingin infectious diseases study how these health-related events are distributed across populations as well as the determinants of health and disease, and we track the spread of infectious diseases, examining the effectiveness of public health interventions. Biostatisticians also provide statistical expertise to hospitals and research institutions, helping to analyze clinical studies, data and quality improvement projects. Healthcare administrators use statistics for monitoring patient outcomes, enhancing healthcare providers' efficiency, and controlling healthcare costs. When used correctly, statistics enhance patient care, advance medical knowledge andpromote evidence-based public health.

EngineeringandTechnology:Statisticsisusedinengineeringandtechnology for quality control, reliability analysis, process optimization and many others. Engineers use statistical methods as the foundation for experimental design, dataanalysis,aswellasproductandprocessoptimization.Manufacturingof

more brands S0F SPC techniques are dominant products quality and defects in data analysis and the designed quality engineers at the design process of manufacturing. In reliability analysis, statistical models are used tocharacterize the failure likelihoods of engineering components and systems. Some techniques basically based on statistical-based methods, like machine learning and data mining areused to get information from certain large number of datasets and the aforementioned techniques are called data-driven methodsto predict complex issues in various engineering processes or systems. Hereare the few sentences to explain this concept Statistics in Civil Engineering If statistics be used in civil engineering, statistical methods are used to analyze structural data for safety of bridges and buildings. In computer science,network traffic analysis, these statistical techniques are applied on Cyber security as well data compression. Statistical Techniques in Business and Industry: Enhance Quality, Boost Productivity and Promote Innovation.

Environmentalscience&ecology: Environmental scientists and ecologists use statistical methods to examine the effects of human activity on the environment and to monitor changes in the environment and in ecosystems. Statistical methods may be used to process environmental data, emulate ecological phenomena, and ascertain the effectiveness of conservation efforts. Statistics Development of probabilistic models (e.g. weather), analysis of climatedata,model forclimatechangeimpacts. Ecological Statistical methods are used by ecologists tostudy population dynamics, species interactions, and biodiversity. Statistical sampling techniques are also applied in environmental monitoring programs measuring air and water quality as well as pollutionlevels and the effects of regulations.Wu, B. All of these statistics play an important role in thefields of environmental science and ecology, as they will help understand the detail of the ecosystems and move towards potential decisions about environmental policy.

Statistics has been the backbone of the data scienceand artificial intelligence revolution that is reshaping large parts of the tech and business landscapetoday. Using outliers from statistics and extracting data from large datasets, datascientistsdesignpredictivemodelsanddiscoveractions.Supervised

learning algorithms, grounded in the statisticalproperties of data, are used in applications including image classification, natural language processing and fraud detection. Data visualization,data cleaning, or feature selection alsousestatisticaltechniques.But,inaworldwherethecreationofdataisatodds, we need the skills of capturing and transferring knowledge. Statistical MethodsforBigDatainDSAIandHands-onworkRationale:Theintegration of statistics with data science and artificial intelligence has driven radical innovation in healthcare, finance, transportation, entertainment, and elsewhere.

Finally, the essence of statistics is the quasi-parametric recognition art. It encompasses a widerange of domains and applications. It is fundamental in thatit transforms raw data into computable knowledge that underpins sound decision making, the resolution of complex problems, and advancements in scientific understanding. In an increasingly data-driven world, the need for statistical proficiency is on the rise, Statics is crucialand amongst the most requisite skills across virtually every domain. Reading science,data scienceis being trained to hunt, analyze and chew data, it is17 important to organize the randomness of life, realize science, technology and society is very important, the meaning of the 21st century.

UNIT1.3TYPESOFSTATISTICS(DESCRIPTIVEAND INFERENTIAL)

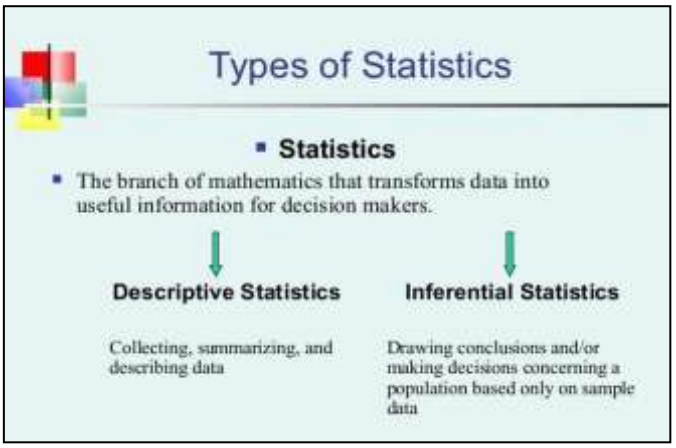

Figure1.3:TypesofStatistics(DescriptiveandInferential).

1.3TypesOfStatistics

DescriptiveStatistics:SummarizingandPresentingData

Descriptive Statistics is a set of methods in which information is summarized based on an overview of the raw data. This branch focuses on just characterizing a dataset's key characteristics without taking inferences and extrapolating beyond the dataset or sampling unit. Descriptive statistics inherently is the tool used to summarize large amounts of data into usable summaries that help researchers and analysts understand the fundamental characteristics of a sample or population. Central tendency refers to the value that is in the center, for instance, the mean (average), the median (middle value), or the mode (mostfrequent value) of a data set. Mean is sensitive to extremevaluesandworkswellforsymmetricdistributions,whilethemedianis resistant to extreme values and is better suitedin skewed distributions. Mode gives the most occurred value so it is very useful in Categorical data. Additionally, measures of dispersion, in particular, range (the difference between the highest and lowest values), variance (the mean of the squares ofthe differences between each data, and mean) and standard deviation (the square root of the variance) give an insight into how much variability (or spread)thereis around the central tendency. A small standard deviationmeans thatyournumbersclusteraroundthemean,andabigonemeansthatyouhave

a more spread-out bunch of numbers. Whereas, percentiles and quartilesdividethe data into equal portions and have us understand how individualdata values are situated in relation to the entire distribution. These areknown as histograms, bar charts, pie charts, box plots etc., and such visual representations help to understand the distribution of data and patterns involved therein. Histograms are used for continuous data (frequency distribution), bar charts are usedfor categorical data, pie charts are used for portions of a whole, and box plots are used for summary of statistics of distribution such as quartiles and outliers. That brings us to the third part of Descriptive statistics also known as shape measures (skewness: symmetry of the distribution; and kurtosis: peaked Ness of the distribution) giving the whole entire spectrum of the data in terms ofits shape. Skewness indicatesthe symmetry of the distribution of data (or lack thereof), while kurtosis indicates data is concentrated around the mean where heavier or lighter tails lie. Inessence, it

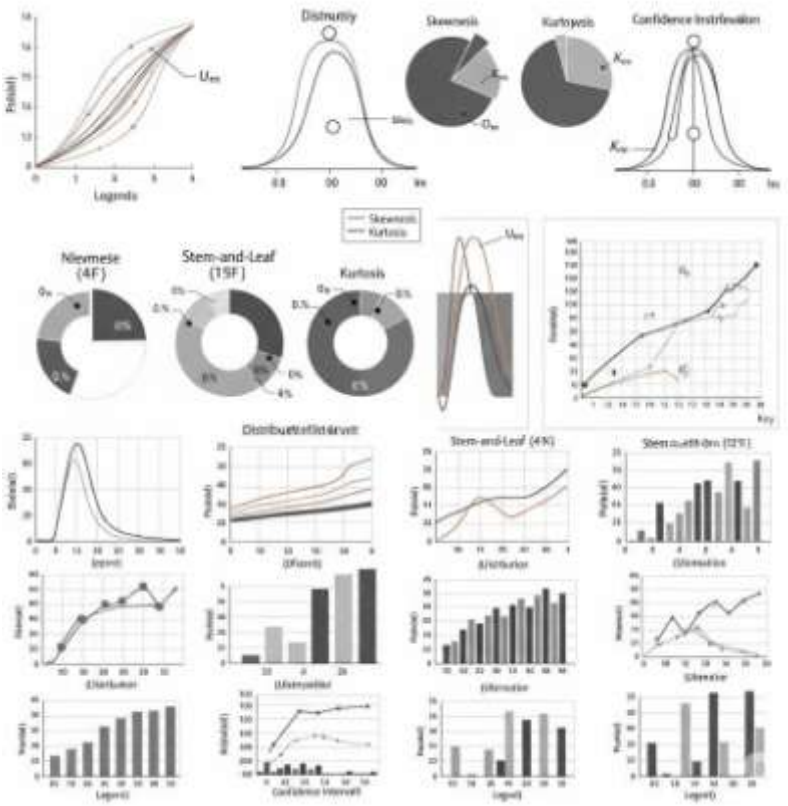

Figure1.4:descriptivestatistics

provides the data filler for deeper analyses and meanings. Descriptivestatistics provide researchers with methods to describe their raw data in various ways in order to find patterns and outliers within the data set so that they can derive conclusions to inform their understanding of the phenomenon theyarestudying. Ifyou areinterested,you would get toknowsomeofthese in these post 3 Exploratory Data Analysis(R/W) This use of EDA is meant to find the patterns, that enables to proceed from EDA to other more sophisticated statistical analysis. When the process of descriptive statistics is performedto the fullest extent possible, it sets a strong analytical foundation for subsequent operations, all of which can be resting on firm knowledge of the basic characteristics of the data. This allows for the identification of potential issues with the data that has been collected, such as outliers or inconsistencies that can be corrected before performing more advanced analyses. While it is one thing to demonstrate that you have the skills to analyze the data, it is another thing to provethat you can communicate the insights you have from your descriptive statistics - you will want to sharewhat you have found to as many people as you can, and not just other statisticians.

InferentialStatistics:

After description, the need for inferential statistics comes into play, not to mention how statistics is derived from the complexity of data between which first seem uncorrelated or unrelated, and acts by inferring, and hypothesize over data from samples that it is intended to represent more extensive and unique populations until it reaches the workplace. If you have no prior knowledge about the entire population then you can still derive the inferences through samples, in case you conduct the study and interpret them using inferential statistics. The idea behind inferential statistics is that if you draw a sample and that sample is a proper representative of that population (properly selected), you would have an idea of the characteristics of the population. The methods usedin inferential statistics include but are not limited to hypothesis testing, confidence intervals, and regression analysis. The null hypothesis (statusquoornodifferencebetweentwogroups)andthealternative

hypothesis (the opposite of the null hypothesis) are just initial assertions of hypothesis testing. Statistical Tests (T-tests, chi-square tests, ANOVA, etc.) can be used to confirm whether or not we have sufficient evidence to rejectthe null in favor of the alternative. Using sample data, confidence intervals provide an interval in which the true population parameter will lie. The terminology that is often used is that a 95% confidence interval means the following: If the sampling process were repeated many times, there is 95% chance that the 95% confidence intervals will sweep through the value of the true population parameter. This simplest form of analysis is the regression analysis where the the dependent variable is established based on the dependent variables. A linear regression is, for instance, a straight line with morethan two variable relationships. Inferential statistics are underpinned by probability theory, which enables researchers to quantify uncertainty andmakeprobabilisticinferencesaboutpopulationparameters.Sampling(random sampling, stratified sampling, cluster sampling) is important to make the sample representative of the population. Data collection methods depend on the research question, the characteristics of the population, and available resources.Sampling technique best suited to population characteristics.

The validity and reliability of inferential statistics depends on how good the sample is from which we are drawing a conclusion, and how appropriate the tests are for our data. Assumptions onthe distribution of the population must, like any such normality, be used and tested with caution. Inference based on data science for making data-driven decisions and advancing scientific knowledge exists in various fields of life: like biology, psychology, economics, social science and so on, hence inferentialstatistics is ubiquitous. To give a better real-world example, you use inferential statistics when runningclinical trial to find out whether anewdrug is effectivein comparison to placebo. For example, inferential statistics are used inmarket research to make predictions about consumer behavior and preferences. Social Sciences examine social trends and patterns (including by means of inferential statistics). This allows researchers to draw conclusions about a broader populationbasedontheinformationgatheredfromthesample.

variableisonemorebenefitofinferentialstatistics.Thisabilitytopredict allowsforbetterplanningandresourceallocation.Relativeconfidenceof predictionshelpsresearcherstomakemoreinformeddecisionsandavoid some risks.

UNIT1.4FUNCTIONSANDLIMITATIONSOFSTATISTICS

At its core, basic statistics makes it possible to describe and summarize data, turning raw numbers into meaning with measures of central tendency (mean, median, mode), measures of dispersion (variance, standard deviation), and graphical methods. This theoretical concept gives us an idea to understand the dataset at a higher level byidentifying important features and helps us to find the phenomena hidden in the raw data. Statistics balances, align, sorts and scales so complex information can be communicated effectively and efficient. Data analysis and interpreting thedata is possible throughstatistics and various techniques like hypothesis testing, regression analysis and variance analysis, and can be used to derive inferences and understand the relationship between variables. Analytics enables us to identify cause-and-effect relationships, predict future behavior or condition, and assess the significance of differences inthe data we arepresented with.Whatcomes nextisnotmere description but rather generalizations and theory testing. The latter lays the foundation for making decisions and shaping policies with evidence-based findings that influencedecisionsinvariousfields.Businessesusestatisticalanalysistomake decisions, forecasting future circumstances and risk assessments, while governments rely on statistical information to form policies on public health, education, and economic progress. Applying statistical modeling and forecasting enables companies to predict the trend before others do and make necessary adjustments methodology is also fundamental in scientific investigation, where it guides experiment design, data collection and analysisto reach valid conclusions. From clinical interventions to ecological studies, statistics provides the rigorous framework necessary to test hypotheses and discover new knowledge. Lastly, statistics is used in quality control and improvement to measure and improve the consistency and reliability of processes and products. Consequently, statistical methods are alwaysapplicable to the variations, their sources of error, thus enabling production to be optimized, defects diminished and quality enhanced.

LimitationsofStatistics: Thestatistics mayoffer you sometools, but it isalso important to recognize what the limitations of the statistics. Statistics is inherently biased for two main reasons,the first of which, is that the entiredata selection, collection, and interpretation process is completely in the hands of the researcher and is subject to his/her views and preferences. For example, biased sampling can lead to unrepresentative data and flawed conclusions. Moreover, statistics have the limitation of quantifying data, so they can never capture qualitative modalities such as subjective experience, opinion and emotion. Qualitative data can be abstracted into quantitative representations — but doing so losesnuance and detail. Second, statistics relies on assumptionsofnormality or independence that do not hold in the real world. The reason is the assumptions (mentioned above) which, if any one of them holds, the statistical results are not valid and therefore any conclusions can bemisleading. Moreover, statistics can be biased or misapplied, and statistical evidence may be manipulated or employed selectively to promote particular interests. Furthermore, the power of statistics is limited by the accuracy and validity of data; errors in data collection, measurement or documentation can propagate through analyses, producing erroneous results. As the saying goes, garbage in, garbage out; statistical output is ultimately constrained by the quality of input data. Averaging, however, can obscure crucial individual differences. But you have to remind yourself that statsonly can tell trend and pattern; they do not explain trends and patterns. And statistical analysiscannot make any inference about causality, much less reverse causation. The key to causalinference is design and confounding. And statistics is a time-sensitive discipline because data and trends can change rapidly, with potentiallyoutdated analyses. It is most applicable in such fast-changing fields as economics, finance and the social sciences. Generally speaking, forecasts and statistics-based models need to be constantly updated to reflect, as accuratelyas possible, the current state of affairs. Third, statistical methods arecontextual, meaning that they may not work in other disciplines, cultures, and settings nor be interpretable in them. A statisticallysignificant finding in one context is not necessarily meaningful in a different context. Another problem with sole reliance upon statistical significance is that this may place emphasis onstatisticallysignificantresultsattheexpenseofpracticallysignificantones.

UNIT1.5MEASURESOFCENTRALTENDENCY

Central Tendency this is a very basic statistic that indicates a representative value of the dataset i.e. the typical or central value of a dataset. These give a quick way to find out where mostof the data are, which is useful in making comparisons and inferences. Chapter 3 describes a number of measures (arithmetic, geometric and harmonic means, median, mode, and quartiles) in terms of their calculation, use, and advantages and disadvantages.

1.5MeasuresofCentralTendency

Mean(Arithmetic,Geometric,Harmonic): The arithmetic mean (The average) is calculated by adding all the values of all the data points togetherand dividing the sum by the number ofdata points It is extremely sensitive to outliers, so a symmetrical distribution without extreme values is ideal. E.g. daily sales for a week for a small bakery: [20, 25,30, 28, 32, 22, 26] So the averageDailySalesforAisArithmeticmean(20+25+30+28+32+22+26)/7= 26.14 So if there were high sales on one day (say 100) the mean would be highly skewed and would notreflect sales accurately. It's used more with data that expands in multiplicative or exponential manners, such as financial return or patterns of growth in a community. It's calculated as the nth root of the product of n individual data points. Since the geometric mean considers the product of stock returns, to account for compounding, for three years of stock returns5%, 10% and 15% the calculation to find geometric mean return is (1.05 x 1.10 x 1.15)^(1/3) − 1 ≈ 9.98% corresponding to compounded average growth. It is less affected by extreme values than the arithmetic mean, but can only be applied when all values are positive. Harmonic Mean: Used in situations involving rates or ratios. So you can calculate that value as the number of datapoint divided sum of the inverse of the data point. E.g., if we travelled a distance of 100 km with a constant speed of 40 km/h and then travelled the same distance with a speed of 60 km/h in the end, the average speed for the entire trip = (2/(1/40 + 1/60)) = 48 km/h (harmonic mean speed) Thisis particularly something very different when the denominator is constant and it can be said the harmonic mean is more appropriate than standard mean that time.

Median: The median isthe middle value in an ordered data set. In the case of even number of values in the dataset, the medianis the average of the two center values. Whereas the arithmetic mean is less robust when dealing with outliers, simply because of how individual valuesaffect the mean, the median is less influenced by outlying values, and as such, a robust measure, usually when the population is skewed. To illustrate this, imagine that you have the salaries of employees of a small company: [30000, 35000, 40000, 45000, 100000] Even though the arithmetic mean salary is 50000, skewed by the outlier 100000, the median salary 40000 is a much more accurate representation of the average salary. Tofind the median, we first arrange the arrayinincreasingorder[30,000,35,000,40,000,45,000,100,000].The middle value is 40,000. If the list was even, e.g. [30,000, 35,000, 40,000, 45,000], the median would be (35,000 + 40,000)/2= 37,500.

Mode: Themodeis the numberwith themost common occurrenceofany data set.Adatasetisunimodalifithasonemode,bimodalifithastwomodes,and multimodal if it has multiple modes. This is useful for categorical and discrete numerical data. Atrivial example: thecolors ofcarsin aparkinglot: [red, blue, red, green, red, blue, yellow]. The mode the most common color isred. In the caseofanumericdatasetlike[1,2,2,3,4,4,4,5],themodalitywillbe4.In other words, for the list [1, 2, 2, 3, 4, 4], themodes are 2 and 4, so it isbimodal distribution. Although the mode is best used at classifying the dominant category or number, it cannot reflect if the exceptional number is not cited via themedian.

Quartiles: Quartiles are metrics that divide a dataset into a lower 25%,second 25%, third 25% and upper 25%. The first quartile or Q1is the median of lower half of the data whereas the second quartile or Q2 is the median ofthe dataset (which is also the median) and the third quartile or Q3 is the median of upper half of the data. In conjunction with the median, they help gauge the spread and distributionof data. scores: [50, 60, 65, 70,75, 80, 85, 90, 95, 100] First, we essentially findthe quartiles and order the data (that is already ordered).Median (Q2)= (75+80)/2 =77.5 LowerhalfforQ1 =[50, 60,65,70,75]somove2termsupanddivideby2.Q1=(60+65)/2=62.5

The top half is [80, 85,90, 95, 100] thus Q3 = 90 The quartiles tell you the location of the middle 50% of data (interquartile range, IQR = Q3 - Q1),whichinthiscaseisbetween90 -65 =25. Even better,theinterquartilerange (IQR= Q3− Q1) isa more robustmeasureof spreadthantherange (Gibbons, 1974; McGill et al., 1978). Quartiles are often used to visualize these data pointson box plots.

All three measures of central tendency provide slightly different perspectiveson the center of a dataset. Therefore, it can be good average for symmetric distributions, but, very sensitiveto outliers. For multiplicative data, we use the geometric mean, and the harmonic mean in case of rates. The median is resistant to outliers, thus its suitable for skeweddata. The mode tells youwhich value appears most frequently, whereas quartiles show how the data splits into equal quarters, providingyou with a sense of spread. The measure chosen will vary based on the data type of the analysis along with theanalysis objective. The analysts then are empowered with the right knowledge and with the right skills to interpret the data and come to conclusivelyhelp understand the data in much simpler terms.

UNIT1.6MEASURESOFDISPERSION

1.6MeasuresofDispersion

Central tendency summariesthe mean, median and mode provide a glimpse into what a typical value looks like in a data set, but don't capture the full picture. We also need to look at the distribution of the data to capture what lies behind the data. Variance is a measurement of how far data points are spread out from their average value. It is an important idea in many fields, from finance, where it is a measure of risk, to quality control,where it is a measure of consistency. Finally in this section, a couple of important measures of dispersion like range, interquartile range, mean deviation, standard deviation, variance and coefficient of variation and significance is discussed by giving suitable examples.

1. 6.1RangeandInterquartileRange:SimpleYetInsightful

I also encourage you to play around with measures of spread like range (Max–Min) and the interquartilerange (Q3 – Q1) these are so simple to compute but can give you clear insight into the spread of your data. The range is the simplest measurement of dispersion, it's just the difference between thelargest and smallest number in a set of data. 1 Easy to compute, it is quite sensitive to outliers, providing a very bad indication of global variability. For example, if this is the daily high temperature for a week {25, 27, 26, 28, 30, 26, 45} (in degree Celsius): This is because the range is45 - 25 = 20 degree. But those 45 outliers really stretchthe range. The interquartile range (IQR) is a measure of spread that looks at the middle 50% of the data and is less affectedbyoutliers.Thisisalsoknownastheinterquartilerange(IQR),which is the difference between the third quartile (Q3) and the first quartile (Q1). Quartiles canbe used to split a data set into four equal segments. Using the same temperature data, however, sorted: {25,26,26,27,28,30,45}, so the and Q1: Q1: 26 while Q3 is similar to 29 (approx) Hence IQR = 29 − 26 = 3 degrees. This metric is more resistant to outliers and thus a better representation of the spread ofthe central entries. In your analysis ofincome distribution, consideration of IQR might provide

Information on the extent of middle-class wealth without being skewed by extreme affluence or poverty, for instance.

1.6.2MeanDeviation:AverageAbsoluteDeviation

MD: Meanabsolute deviations of each observation from mean. It provides a more comprehensive image of dispersion thanrange or IQR, as it considers all of the data. The formula for MD is:

$MD=\Sigma|x_i-\mu|/ n$

where x_i refers to each individual data point, μ is the mean, and n is the total number of datapoints.

Let'ssayyouhaveafewtestscores:{70,80,90,60,100}.Themeanis80. Theabsolutedeviationsare|70−80|=10,|80−80|=0,|90−80|=10,|60−80|=20, |100−80|=20. The sum of these absolute deviation is 60. 60/5= 12 the mean deviation this imply, on average, 12 points away from the mean have test scores. Mean deviation is a very intuitive measure, but it is less commonlyused than one would think, because its mathematical computation isintractable.

1.6.3StandardDeviationandVariance:TheCornerstonesofDispersion

The SD is also themostcommon measure of dispersion (or variance), where it is defined as theaverage distanceadatapoint is to themean. Then thestandard deviation, which is the square root of the variance here. Variance isthe mean of the squared deviation from the mean. The formulas are:

$Variance(\sigma^2)=\Sigma(x_i-\mu)^2/n$(forpopulation)or$\Sigma(x_i-\bar{x})^2/(n-1)$(forsample)
$StandardDeviation(\sigma) = \sqrt{Variance}$

Withthe same testscores{70, 80, 90, 60, 100}, the variance:

$[(70-80)^2+(80-80)^2+(90-80)^2+(60-80)^2+(100-80)^2]/4=[100+0+100+400+400]/4=1000/4=250$.TheStandardDeviationis$\sqrt{250}=15.81$ (approximately).

A higher standard deviation means greater diversity, while a lower number means the data points cluster closely to the mean. In finance, greater standard deviation ofstock returns mean greater risk. For example, in manufacturing,by showing the lower standard deviation of the product dimension indicates more uniformof the product dimension that leads to a higher product quality.

1.6.4CoefficientofVariation:RelativeVariability

The CV is relative measure of dispersion expressed as a percentage. It's calculatedas the ratio of the standard deviation to the mean:

$CV=(\sigma/ \mu) * 100\%$

The CV deals with the variability of multiple datasets which could havevarying units and very different means. Standard deviations, as a matter of convention, are completely irrelevant when comparing: e.g. natural comparisons, like the variability of stock prices (in dollars) and the variability of temperature (in degrees Celsius), are meaningless. However, the CV makes for adecent comparison.

Supposetwodatasetshavethefollowing properties:

- DatasetA:Mean=50,StandardDeviation= 10
- DatasetB:Mean=200,StandardDeviation=20

ThestandarddeviationofDatasetB ishigher,but theCVsare:

- CV(A)=(10 / 50) * 100%=20%
- CV(B) =(20 / 200)* 100%=10%

In Dataset A, we have more relative variability but less absolute variability (standard deviation). The CV is significant for finance and quality control,since it is needed to comparethe relative risk nor process variation

ChoosingtheRightMeasureforInsightfulAnalysis

Without understanding the measure of dispersion, overall analysisabout data remainsincomplete.Althoughtherangeand IQRlist alldatapoints(nonewere included in this example), these options quickly summarize total spread and typical variability. Mean deviation measures the average absolute deviation, whereas the standard deviation and variance are the building blocks for measuring squared mean deviation. Finally, this property enables comparisonof relative dispersion among different data sets by the coefficientof variation. Which measure is appropriate and in which case depend on the nature of the dataand data context. And having an in-depth understanding of these metrics helpsanalyststoworkwithadeeperunderstandingofhowmuchdatacanvary, makingthem lead to better decisions and right conclusions.

UNIT1.7SKEWNESSANDKURTOSIS

1.7.1UnveilingtheShape:MeaningandInterpretationofSkewnessandKurtosis

The basic concepts of statistics are concerning the central tendency and variation of the data. These measures alone, however, often do not express enough about the underlying distribution. Skewnessand kurtosis look deeper intotheshapeandsymmetryofdatasets.Inlaymanterms,skewnesstellsyou about the asymmetry of a distribution. A perfectly symmetric distribution (such as the bell-shaped normal distribution) has zero skewness. Havinglonger or fatter tail to the right denotes positive skewness: The mean of the distribution is higher thanthe median. This suggests thatthere are some very high values that are affecting the average. Conversely, in negative skewness (left skewness), the left side has a longer or thicker tail so that it has a mean lower than the median byextreme low-value.

Kurtosis, on the other hand, is analytics of tailenders or peaked Ness of a distribution. It measures how closely data points cluster arounda mean and how heavy tails are. Leptokurtic: high kurtosis sharp peak heavy tails adding to thetail extremism Platykurtic distributions have low kurtosis and a lower peak with thinner tails and fewer extreme values. In particular, a normal distribution, the reference, has moderate kurtosis and is called mesokurtic. These properties of data are incredibly revealing in exposing the profound features of data to a level much deeperthan basic characteristics of meansand spread. Data on the risksideofthe distribution tail, such as financial data that are influenced by extreme events, tend to have a high kurtosis. We willget normal distribution fordata from stable process.

1.7.2MeasuringAsymmetry:DelvingintoMeasuresofSkewness

In order to measure the skewness, it needs to bequantified. An eternal method ofmeasuringtheskew,wouldbetouse(D1)thefirstcoefficientofskewness,

(Pearson), which is calculated between the mean vs mode. This metric is Computed as: (Mean-Mode) / Standard Deviation If the value is positive, it will have a positive skewness, if it is negative, it will have a negativeskewness & if the value is very close to 0., then itis symmetric. However,this measure issensitive to the mode that is not always reliably determined. Another popular measure is Pearson's second coefficient of skewness basedon mean and median. Formula to calculate Skewness: 3(Mean – Median) (or) 3(Median – Mode) / σ This is slightly better of a measure compared to the first, since the median is more robust against extreme values than the mode. The sign showsthe direction of skewness and its absolute value, the force. A more subtle and routine technique uses the third moment of thedistribution. This approach calculates the standardized third moment, resulting in a numerical score that reflects the degree of asymmetry. Typically, this is calculated throughsoftware. For example, we have a data set of scores for an examandweusedsomestatisticalsoftwaretofindoutthep-values.Anet
+0.7 would suggest a “fairly positively skewed” distribution; that is, many of the scores are below the average, such that the higher scores “pull up” the mean. Where a slight negative skew would be -0.3All this skewness is measure that give a little bit different insightsinto the nature of the data, it gives researchers and analytics to choose the kind that is better for them.

1.7.3GraspingtheTails:TheKurtosisIndexandItsSignificance

Kurtosis,as mentioned earlier, describes the tailenders of a distribution. This property is measured with a number called the kurtosis index(kurtosis) Now, the above formula of kurtosis has the fourth moment of the distribution asits initial part, normalized to the degree that sets up for differences in scale. (Just know that the most common way software packages report this is as “excess kurtosis,” which is kurtosis − 3.) This is done so the normal distribution has excess kurtosis of 0 (the kurtosis of the normal distribution is 3).

- **Leptokurtic(positiveexcesskurtosis):** It has pointy peakand heavy tails (known as leptokurtic). Thisindicates that data points are clustered near the center, and a broader distribution of tail chances. Leptokurtic distributions arecommoninfinancialmarkets,particularlyinstockreturns,indicating that

- extreme positive or negative outcomes are more likely than what a normal distribution might imply. For example, a kurtosis index of 5 would imply a leptokurtic distribution while analyzinga data set of hourly stock price changes.
- **Platykurtic(negativeexcesskurtosis):** A platykurtic distribution has a lower, flater peak with thinner tails (indicating more evenly dispersed data, suchthatextremevalues arelesslikely).max isnearto1Anormaldistribution ends at 3 std dev so this is more probably a special condition of Less squares or More squares condition where data is limited or controlled.
- **Mesokurtic(withexcesskurtosisnear-zero):** Mesokurtic distributions(forexample,thenormaldistribution)have intermediate tailsand a moderate peak It is what you were trainedto measure against.

From Kurtosis index you can have a hypothesis test of thetailedness of the distribution (how far it is from normality). Such data is vital for risk assessment, statistical modeling, and decision-making.

1.7.4PracticalApplicationsandInterpretiveNuances

Skewness and kurtosis are not just themselves abstractions; they bear great practical meaning in various fields. Even in finance, most ofthese measures represent the risk of investments. Positive skewness in returns would mean that you have a higher chance of having higher returns while high kurtosis indicates a higher chance of lower returns or downside risk. Skewness in production, for example,may show bias in the manufacturing process, while kurtosis can show variation in the dimensions of parts produced. In the social sciences, such measures help facilitate understanding of how income, test scores and other such variables are distributed. Skewnessand kurtosis make sense given context, however. In small-sized samples, the utility of skewness and kurtosisestimates can be questionable. Thus, confirm sample size, and use best practice.Additionally, histograms and box plots fordata visualization act as ExtraM/minimum to the numeric outcomes. In short, if Researchersunderstand skewness and kurtosis, they will get more insightsand eventually will also make better and informed decisions.

UNIT1.8INDEXNUMBERS

1.8.1IndexNumbers:MeaningAndImportance

Index numbers are a very efficientstatistical tools to measure the variation in a variable (that is a con Shared Attribute) or a set of related variables overtime or from them in different locations. In short, they distill data into a number that communicates a lot with little explanation. Instead of working with raw data, which can be unwieldy, index numbers provide a comparative measure of change; an index number uses a base period or location as a reference point. The base is typically set at a value of 100 and the relative amounts are described in percentage terms in comparison to this base. The consumer price index (CPI) is an index number in which a number indicates how muchpriceshaveincreasedfromabaseperiod(100).Theyare important as they indicate trends and patterns not readily discernible otherwise. They are relied on by economists,policymakers, businesses, and researchers who seek to understand and analyze economic phenomena. Index numbers are a statistical measure that enables ongoing quantitative comparisons over time by recognizing that prices, outputs and other variables are always in flux. They assess the impact of economic policy, determine the cost of living, monitor inflation and guidebusiness decisions.

Say, we want to compare the wheat production of a region over a decade. Rather than measuring in raw tonnage which would be misconstrued to larger variables such as area of land, Item of weatherand many more, we may take index number. The concept is quite simple, we take a base year, we cansay 2010 and index it at 100. This means here if Wheat Production in 2020 = 125In 2010 we had a Wheat production of 100 and we observe 25% growth with comparedfiguresofearlieryear.Simplifieditmightbe,butitmakesfor rapid, useful comparison. Index numbers also enable comparison across time and space. (For example, where you compare the CPI between countries to measure relative inflation differences. In business, they track salesperformance, market share and productivity. Index numbers also aid in summarizing the changes, making informed decisions and strategic planning.) Thisreductionisnotonlyusefulforfunctionssuchasdevelopment,

entrepreneurship, and innovation (among many others), but also provides important insights due to them being compressed with the exploration of the resulting economic-and socio-historical vectors. Butindexnumbers,yousee, also allow you to deflate nominal values into real values. Nominal GDP may rise, but that rise may simply reflect inflation or it may reflect an increasein production. Real GDP measures the value of output produced in an economy while controlling it for inflation and using a price index to deflate the nominal GDP. Therefore, realGDP is adjusted for the price level in the economy.

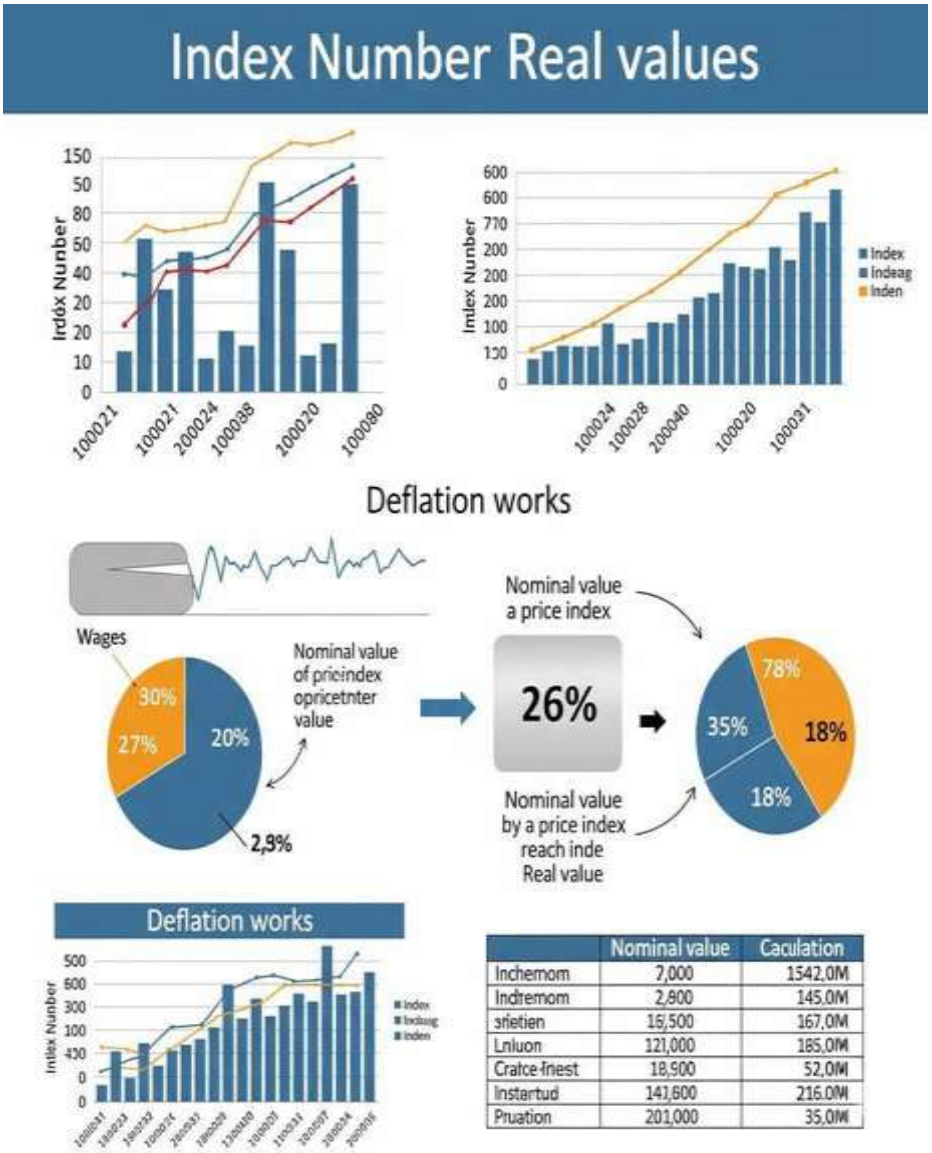

Figure1.5:IndexNumberRealValues

1.8.2 *TypesofIndexNumbers*

Broadly, index numbers can be classified on the basis of the variables measured, the methods of construction. Understanding these differenceswould helpus select a relevant index for that use case.

1. **PriceIndexNumbers:** Price index numbers are most commonly used index numbers, as they measure changes in the general price level. The Consumer Price Index (CPI) is classic example, which seeks to measure average change over time in prices paid by urban consumers for market consumer basket goods &services. The WPI is a measure that tracks the prices of goods sold in bulk as well as inwholesale markets. Anotherinflation measure is the Producer Price Index (PPI), which looks at theaverage price increases domestic producersreceive for their products.

Example: the CPI for a country could demonstrate an increase from 100 to 110 from 2020 to 2023, which means that consumer prices rose by 10%over the course of three years.

2. **QuantityIndexNumbers:** Volume/quantity of goods & services producedor consumed. Tomonitor this and arrive at a better assessment of the health of the industry, economists use a number of metrics, one available on a monthly basis Most importantly, the Index of Industrial Production (IIP), which measures growth in the physical volume of production across sectors in the economy

Example: If IIP goes up from 100 in one quarter to 105 in the next, it means that industrial output has expandedby 5%.

3. **ValueIndexNumbers**: Index numbers, which indicate the aggregate value of a variable determined by a combination of price and quantity. They combine bothprice and quantity movements.

Example: Value Higher prices and increased selling volume could lift value index retail sales.

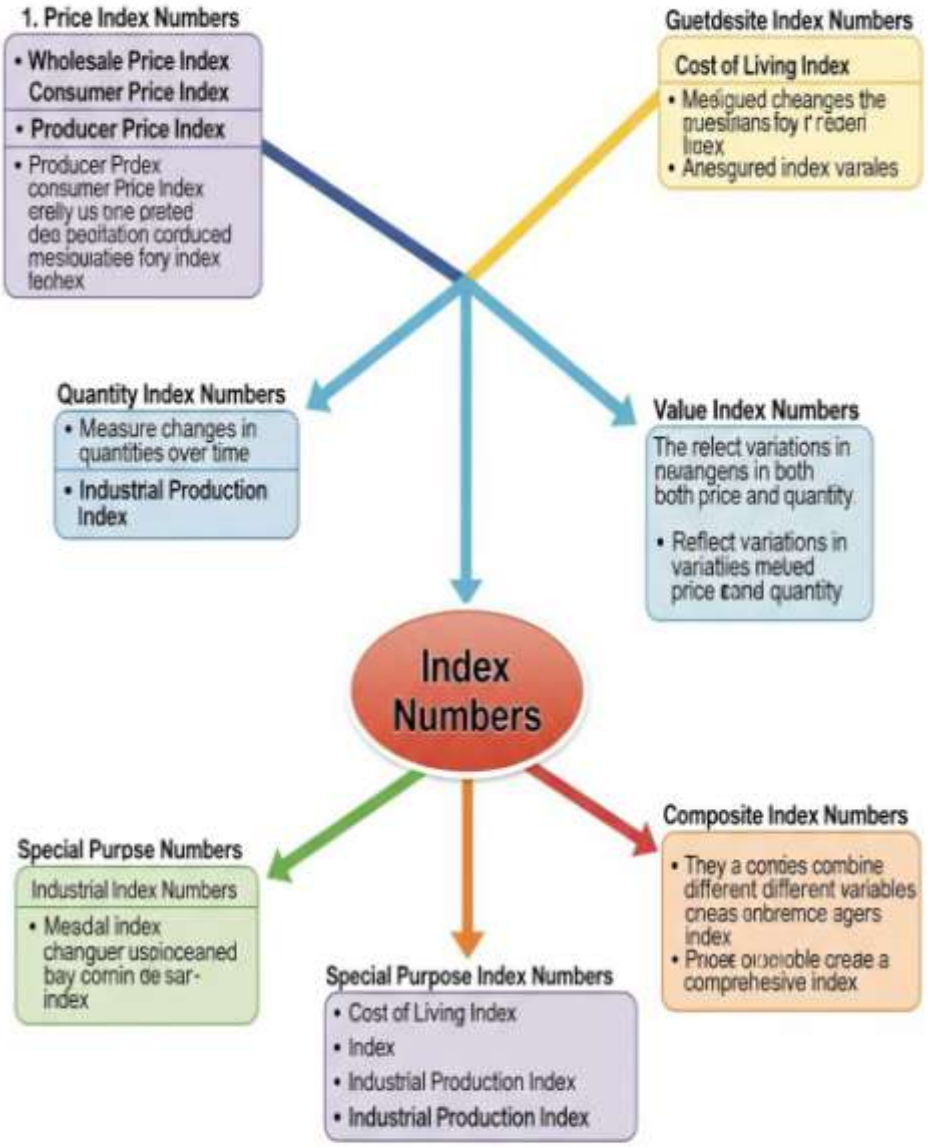

Figure1.6:TypesofIndexNumbers

ThesuppliedtableisasynthesisofIndexNumbers,wherethepurposeand the way of construction is clearly arranged. The nucleus of this idea is "Index Numbers", which can be further classified into five principal varieties : Price Index Numbers, Quantity Index Numbers, Value Index Numbers, Special Purpose Index Numbers, and Composite Index Numbers.

4. **SpecialPurposeIndexNumbers:** These are constructed to represent specific phenomena of change. An instance in thiscategory are stock market indices, the S&P 500 being an example: this index tracks changes in stock prices; indices associated with agricultural production, exports, or imports also fall under this category.

- **Example:**It would be similar to saying that the index of stockmarket grow by 15%, means the value of listed stocks increase exponentially.

5. **CompositeIndexNumbers**: Custom Email Manager You can configure a filter for your emails, and Custom Email Manager will wait them in your inbox all the same. For example, one possible composite economic index also would have production, employment and price indices.

- **Example**, a number of individual indicators can be aggregated to create an index of economic sentiment, e.g. consumer confidence, business confidence and financial market indices.

Furthermore,index numberscanbe constructedusingdifferentmethods,such as:

- **SimpleAggregativeMethod:** This simply sums up prices/quantities of all items for a given period and compares to from the base period.
- **WeightedAggregativeMethod**: Use this method, where you needto assign weight to each object based on their importance level. Indexing methods are commonly standardized using Laspeyres, Paasche and Fisher idealindex weights.
- **AverageofRelativesMethod**:Foreveryitem,wecalculateadjusted price or quantity relatives (ratios) and average them.

Whichindextypetouse,andhowtobuilditwoulddependonthespecific research question, as well as the properties of the data being analyzed.

1.8.3 UsesofIndexNumbers

Index numbers are used in many different fields, so they are an essential toolfor analysis anddecision-making.

1. **EconomicPolicyFormulation:** There are few notable applications of Index Numbers, they are listed as follows− Economic Policy Formulation Government and policy makers use the index numbers to keep track of the trends in economy and formulate the policies accordingly. The CPI, for instance, is a vital measure in gauging inflation, and adjusting monetary and fiscalpolicy. IIP assists toincrease industrial growth and formulate plans for enhancing production.

- **Example:** A central bank mayraise interest rates to curtail a rise in inflation based on CPI numbers.

2. **BusinessDecision-Making:** Companies use index numbers to identify sales, expenses, and productivity. They assist in predicting demand, pricing goods and making investment choices.

- **Example:** Using asales index to detect seasonal trends and guide inventory adjustments.

3. **WageandSalaryAdjustments:** Many wage and salary agreements are linked to the CPIto ensure that workers' purchasing power is maintained in the face of inflation.

- **Example:** sales index to detect seasonal trends and guide inventory adjustments.Using a.

4. **InternationalComparisons:** It is often used in an index for wage and salary adjustments: Many of theagreements for wages and salaries are tied to the CPI to maintain the purchasing power of workers in the event of inflation.

- **Example:** in many labor contracts cost-of-living adjustments (COLAs) are based on changes inthe CPI.

5. **MarketAnalysis:** In financial markets, stock market indices provide a snapshot of overall market performance and help investors make informed decisions.

- **Example:**A risein theS&P 500indicates an overall increase in the value of listed stocks, which can influence investment strategies.

- **DeflatingEconomicData:** Inflation adjustment is done using index numbers so that nominal economic data reflect real changes. Nominal GDP, for example, can be deflated by a price index to get real gross domestic product.
- **Example:** GDP growth is merely 2%.if nominal GDP has grown by 5% and CPI has gone up by 3%, the real

6. **SocialAnalysis:** They are also used in social analysis to measure a change in social indicators; for instance, poverty rates, health indicators, educational attainment, and health insurance--also referred to as an index number.
   - **Example:** An index ofhuman development maybe constructed fromlife expectancy, education and income indices to gauge overall social progress.
7. **Forecasting:** Index numbers serve in time series analysis to discern trends and patterns, thereby facilitating the forecasting of future values.
   - **Example:** In the IIP context, it is used to predict future industrial production levels through analysis of potential upcoming trends.

Last but not the least, index numbers are being powerful instruments for analyzing and interpretingeconomic and social statistics. This ability to take complex information and distil it down into a simple, stand raised form thatcan be absorbed and understood has made them a must have weapon in the arsenal of policy making, business decision, social analysis and forecasting.

Problems in using Index numbers

Index numbers are powerful statistical tools used to measure changes in economic and business variables like prices, quantities, and values over time. However, their use comes with several problems and limitations:

1. Selection of Base Year

Choosing an appropriate base year is difficult.

If the base year is abnormal (e.g., inflation, recession), the index may give misleading results.

2. Choice of Commodities

Deciding which goods or services to include in the index is complex.

Excluding important items or including irrelevant ones can distort the results.

3. Changes in Quality of Goods

Over time, the quality of goods improves (e.g., technology, packaging).

Index numbers may not accurately reflect real price or quantity changes if quality variations are not adjusted.

4. Problem of Weights

Assigning appropriate weights to different items is subjective.

Wrong weights may bias the index, especially if consumer preferences change.

5. Changes in Consumption Pattern

People’s tastes, fashions, and preferences change frequently.

A fixed basket of goods may not represent the actual pattern of consumption.

6. Data Collection Issues

Reliable and accurate price/quantity data is difficult to collect.

Incomplete, outdated, or manipulated data can lead to misleading index numbers.

7. Regional Variations

Prices of commodities vary across regions.

An index number prepared for one area may not reflect the situation in another.

8. Problem of International Comparisons

Different countries use different methods, base years, and baskets of goods.

This makes global comparison of index numbers difficult.

9. Static Nature

Most index numbers are based on a fixed base year and do not adapt quickly to economic changes.

They may fail to capture structural shifts in the economy.

10. Misinterpretation

Index numbers are averages and may not reflect extreme variations.

Policymakers, businesses, or the public may misinterpret them without understanding their limitations.

Great question

For index numbers to be reliable and meaningful, certain basic requirements should be fulfilled. These are:

Basic Requirements of Index Numbers

1. Purpose Clarity

The objective of constructing the index should be clearly defined (e.g., cost of living, price changes, production trends).

2. Selection of Base Year

The base year must be a normal year, free from abnormal events like wars, famines, or pandemics.

It should be relevant and not too far from the current period.

3. Selection of Commodities/Items

The items included must represent the purpose of the index.

For example, cost of living index should cover essential consumer goods and services.

4. Selection of Price/Quantity Data

Data should be reliable, accurate, and comparable across time and regions.

Prices should be taken from standard sources like government reports or official agencies.

5. Choice of Average

A suitable average (arithmetic mean, geometric mean, etc.) should be chosen to combine data.

The choice should depend on the purpose and nature of the index.

6. Appropriate Weights

Weights should reflect the relative importance of different items.

For example, food gets higher weight in consumer price index than luxury goods.

7. Consistency of Data

The units of measurement (kg, litre, meter, etc.) must remain consistent over time.

Standardization ensures comparability.

8. Scientific Method of Construction

Index should be based on a well-defined formula (Laspeyres, Paasche, Fisher, etc.).

The method chosen must minimize bias.

9. Periodical Revision

The basket of goods, weights, and base year should be updated regularly to reflect changing consumption patterns.

10. Comparability

Index numbers should be comparable across time, sectors, and (if needed) countries.

SELFASSENMENTQUESTION

Multiple-ChoiceQuestions(MCQs)

1. Whatistheprimarypurposeofstatistics?

a. Tomanipulatedatarandomly

b. Tocollect,analyze,andinterpretdata

c. Tocreateunnecessarydata

d. Toavoiddecision-making

2. Whichofthefollowingisanexampleofdescriptivestatistics?

a. Predictingnextyear'ssalesbasedonpast data

b. Calculatingtheaverage marksofstudentsinaclass

c. Testinghypothesesaboutpopulationparameters

d. Drawingconclusionsaboutapopulationfroma sample

3. Inferentialstatisticsinvolves:

a. Summarizingdatawithoutmaking conclusions

b. Drawingconclusionsaboutapopulationfroma sample

c. Listingallobservationsina table

d. Measuringonlyqualitativedata

4. Themeasureofcentraltendencythatismostaffectedbyextremevaluesis:

a. Mean

b. Median

c. Mode

d. Quartiles

5. Whichofthefollowingcorrectlydefinesthemedian?

a. Themost frequently occurringvalueina dataset

b. Themiddlevaluewhendata isarrangedinascending order

c. Thesumof allvaluesdividedby thetotal number of values

d. Thedifferencebetweenthehighestand lowest values

6. Whichofthefollowingistrueaboutquartiles?

a. Theydividedatainto threeequal parts

b. Theydividedataintofourequal parts

c. Theyarealwaysequaltothemean

d. Theyarethesameaspercentiles

7. Standarddeviationmeasures:

a. Thedifferencebetweenthehighestand lowest values

b. Thespread ordispersionofdata aroundthe mean

c. Themost frequently occurringvalueina dataset

d. Themiddle valueof adataset

8. Component:CoefficientVariance(CV)Usethecoefficientofvariation(CV)to:

a. Assessthelevelofrelativevariabilityacross solutions.

b. Haveitsdatarangeonly.

c. DeterminetheMostCommon ValueinaDataSet

d. Findtheaverageof aset of data.

9. Skewadatasetisdefinedas:

a. Thesharpest or the largest datadistribution

b. Degreeand directionof distributional asymmetryin thedata

c. Noneof theabove averagefor a dataset

d. Extent:Thedifference of maximumandminimum values.

10. Whatisthewordforhowpointyorflatacurveis?

a. Standard deviation

b. Skewness

c. Kurtosis

d. Range

ShortQuestions:

1. Whatisstatistics?Explainitsscope.
2. Distinguishbetweendescriptiveandinferentialstatistics.
3. Explain The Mean, Median and Mode and Give at Least Three Illustrative Examples.
4. Whatarequartiles?Explain theirsignificance.
5. Expressthemeaningofstandarddeviationanditssignificance.

LongQuestions:

1. Discusstheutilityofstatisticsanditslimitations.
2. Explainthevariouscentraltendencies.
3. Distinguishbetweenmean,median,andmode.
4. Explainthesignificanceofthedispersionmeasuresinthestatistics.
5. Describetheimportanceofstandarddeviationandvarianceexpressed from the data.

MODULE1

Glossary:IntroductiontoStatistics

Term	Definition
Statistics	Thebranchofmathematicsconcernedwithcollecting,organizing,analyzing, interpreting, and presenting data to solve real-world problems.
Population	Thecompletesetofindividualsoritemsthatarethefocusof astatisticalstudy.
Sample	Asubset ofthepopulationselected for analysis.
Variable	Acharacteristicorattributethat cantakeondifferent values.
Parameter	Anumericalcharacteristicdescribinga population.
Statistic	Anumericalcharacteristiccalculatedfromsample data.
Descriptive Statistics	Methods that summarize or describe the main features of a dataset, such as measuresofcentraltendency(mean,median,mode)anddispersion(variance, standard deviation).
Inferential Statistics	Techniquesthatusesampledatatomakegeneralizationsorpredictionsabouta population, including hypothesis testing and estimation.
Data	Numericalorcategoricalvaluescollectedforanalysis.
Probability	Ameasurequantifyingthelikelihoodthatagiveneventwilloccur,rangingfrom 0 (impossible) to 1 (certain).
Sampling	Theprocessofselectingasubsetofapopulationforanalysistoinfer characteristics about the whole.

SUMMARY

Statisticsis a branch of mathematics that deals with the collection, organization, analysis, interpretation, and presentation of data. It helps in converting raw data into meaningful information for decision-making.

Key Concepts:

1. Definition:

Statistics is both a science and an art of dealing with data. It involves methods for handling quantitative information to draw useful conclusions.

2. TypesofStatistics:

Descriptive Statistics: Summarizes data using measures like mean, median, mode, range, standard deviation, etc.

Inferential Statistics:Makes predictions or generalizations about a population based on asample using probability theory.

3. ImportanceofStatistics:

Aidsinresearchacrossvariousfieldslikeeconomics,business,medicine,psychology,and social sciences.

Facilitates informed decision-making.

Helpsinforecastingandidentifyingtrends.

4. Functions of Statistics:

Simplifies complex data.

Providestoolsforcomparison.

Helpsinhypothesistestingand estimation.

Supportspolicymakingand planning.

5. LimitationsofStatistics:

Cannot provide exact answers; only estimates.Canbemisusedtomisleadifnotappliedcorr ectly.
Datainterpretationdependsonthequalityandhonestyofdata collection.

MultipleChoiceQuestions:

1. What is the primary purpose of statistics?

Answer:b.Tocollect,analyze,andinterpretdata

2. Whichofthefollowingisanexampleofdescriptivestatistics?

Answer:b. Calculating the average marks of students in a class

3. Inferentialstatisticsinvolves:

Answer:b. Drawingconclusionsaboutapopulationfromasample

4. Themeasureofcentraltendencythatismostaffectedbyextremevaluesis:

Answer:a. Mean

5. Whichofthefollowingcorrectlydefinesthe median?

Answer:b.The middle value when data isarranged in ascending order

6. Whichof the following istrue about quartiles?

Answer:b. They divide data into four equal parts

7. Standard deviation measures:

Answer:b. The spread or dispersion of data around the mean

8. Thecoefficient of variation (CV) isused to:

Answer:a. Compare the relative variability between datasets

9. Skewnessinadatasetrefersto:

Answer:b. The direction and degree of asymmetry in data distribution

10. Whichmeasuredescribesthe"peakedness"or"flatness"ofadistribution?

Answer:c. Kurtosis

MODULE2

PROBABILITYANDPROBABILITY DISTRIBUTIONS

Structure

OBJECTIVES

- Explaintheconceptandsignificanceofprobabilityinstatistical analysis.
- Digestclassical,empirical,andsubjective probability.
- Applytheadditiveandmultiplicativelawsofprobabilitytoproblem solving.
- Understandandemployprobabalisticdecision-makingprinciples.
- Usebasicresultsfromprobabilitytheoryinstatisticalcalculation operations.
- Beawareoftheapplicationsandmethodsofsamplinginstatistics.

UNIT2.1 INTRODUCTIONTOPROBABILITY

2.1IntroductiontoProbability

Atthecoreofeverythingweexperienceisprobability,influencingourlivesin the obvious and the subtle. It's an edifice that rules our paltry uncertainty, forming a vestigial lung we breathe in each day, a vast model of how to make senseofaworldinwhichcompletecertaintyisararity.Atheart,probabilityis a measure of the extent to which different things could happen in situations of uncertainty. Consider the weather forecast and the fact that there is a 70% chance of rain, or when a doctor explains the percentage success rate of a medical procedure – these are probability in practice. Although most people think of probability only in terms of a game of dice, and cards, this concept applies to various other spheres of life from gambling to science, to medicine, insurance to financial industry, and right to the way we make decisions in our dailylivesirrespectiveofallrationalconsiderations.Probabilityhasitsorigins in antiquity, and 16th-century Italian mathematicians Gerolamo Cardano and Pietro Cataldi are among the first to write of it. But during the 17th century, formal probability theory emerged in correspondence between French mathematicians Blaise Pascal and Pierre de Fermat, while working on gambling problems brought to them by a nobleman called the Chevalier de Méré. Their work introduced the concept of how to systematically compute probabilities of different outcome. From these simple origins grew probability theory, which over the centuries became an elegant part of mathematics with fundamental application in the real world. In everyday life we base a myriadof decisions, from the conscious to the automatic, on probability. When we look at the weather before deciding whether to take an umbrella, we'remaking a decision with probability. When we buy insurance, we're in effect paying to protect ourselves from rare but potentially catastrophic events. Our medical interventions are frequently formulaic, delivered based on statistical evidence of what works with masses of patients. Even a seemingly simple decision such as which route to take to work might entail a back-of-the- envelope calculation of which option is likely to experience less congestion. One of the more interesting things about probability is how it defies our intuition. Human intuition about chance

eventsisnotoriouslyunreliable,leadingtomany commonmisconceptionsand biases in our thinking. For example, after seeing five heads in a row when flipping a fair coin, many people intuitively feel that tails is "due" to appear next. This is the “gambler’s fallacy”: that if something happens more often than normal in one period, it will happen less than normal in the next period, or vice versa. When you boil it down, every single coin flip is it’s own event, so the odds of heads tails will always be 50/50, despite the events before. Learning to think like a probability can teach us to recognize and thwart these cognitive biases.

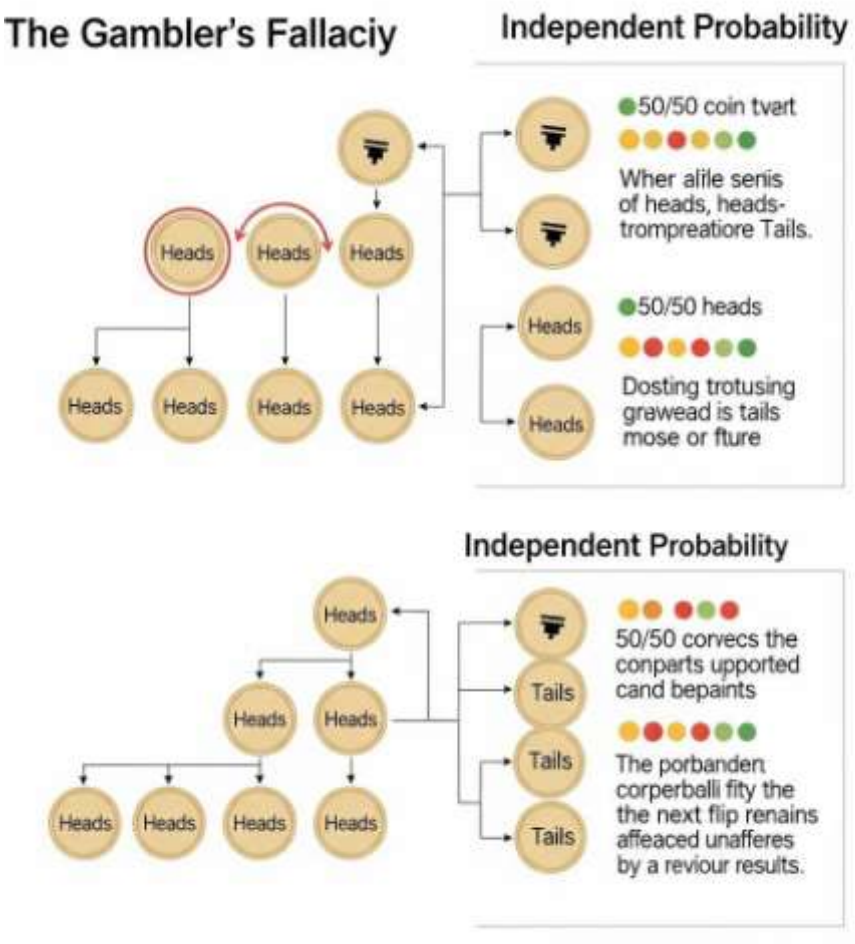

Figure2.1:TheGambler’sFallaciy

The language of probability gives us precise ways to discuss uncertainty. We represent these probabilitiesas a value between 0 and 1 (or0% and 100%). Zero probably event is impossible-the event can't happen, under no condition. The probability 1 is certainty, it's going to happen, guaranteed. All else is a matter of different sorts of likelihood. For example, a fair six-sided dice has equal chance of 1/6 (0.167 or 16.7%) of landing on any of the numbers. This system of numbers provides a means to measure uncertainty and to compare different cases. We can classify the boiling of dice as well: so we can make sense of which events occur together, which don't, and if so, which happened first. Two independent events are two events such that if the first happens, it doesn't change the second's probability from whatever it was before we knew that the first event occurred – like independent coin flips. In contrast, dependent events have an influence on each other — such as when you draw cards from a deck without replacing them, and each successive draw affects the constituent cards that remain. If two events are mutually exclusive, then they both cannot happen at the same time - such as a die showing a 3 and a 4 in a single roll. In other words, complementary events are opposites – if one event doesn't happen, the other one must. These kinds of labels help us to choose the right rules when computing probabilities in complicated situations. Probabilities are spread out over the various possible outcomes according to a probability distribution. The easiest type of distribution is the uniform one, in which all results are equally probable, like for a fair die or coin. Unfortunately, many real world processesare not normally distributed. The normal distribution (or "bell curve") is common in nature and social processes, from the sizes of people's weights and heights to errors in scientific measurements. Other widely used distributions include the binomial distribution (for cases with two possible outcomes, such as success or failure) and the Poisson distribution (for counting rare events on time or space). In probability, we often look for the probability of combined events. The addition rule enables us to determine the probability that one event or another event will occur. For mutually exclusive, together, we just add the probabilities of the individuals. For events that can occur simultaneously, we need to account for the overlap by subtracting the probability of both events occurring together. The multiplication rule helps us find the probability of two events both occurring.

For independent events, we multiply their individual probabilities. For dependent events, we multiply the probability of one event by the conditional probability of the second event given that the first has occurred. Conditional probability addresses how the likelihood of an event changes based on additional information. For example, the probability of a randomly selected personhaving a certain diseasemight bequitelow. But ifweknowthat subject possesses a particular symptom, then we might raise that probability considerably. We can write conditional probability as "the probability of A given B." It's a fundamental concept in probability, and it's used in many advanced probability ideas, such as Bayes's theorem, which provides a method for systematically updating probability estimates as new evidence or information comes to light. Bayes's theorem is one of the most influential and general ideas in probability. Named for the 18th-century English statisticianand minister Thomas Bayes, this theorem offers a mathematical formula for updating beliefs when new evidenceis received. It's especially useful when we are interested in knowing the probability of a cause given an observed effect. For example, if someone tests positive for a disease, Bayes' theorem can be used to compute the probability that the person actually has the disease, accounting for how accurate the test is and how common the disease is in the population. This approach can be applied in medical diagnosis, spam filtering, criminal investigation and machine learning classifiers. If the random process were repeated many times, the expected value is the mean (average) value of the random process. In probability theory, it is the product of all possible outcomes with their respective likelihoods and then summed.

Inotherwords,inagameinwhichyouwin$10withprobability0.2andlose $2withprobability0.8,youcanexpecttowin(10×0.2)+(−2×0.8)=$2− $1.60=$0.40.Thisimpliesyouwouldaverage a gainof40centsperplay over a large number of such plays. The notion of expected value is central in decision theory, insurance, gambling, investments, and a multitude of areas in which the long run is more important than individual outcomes. Statistics derives from probability theory and is concerned with the collection, analysis, interpretation, and presentation of data.

These statistical tricks give us the tools to make very strong inferences about an entire population on the basis of only samples, and a quantification of how uncertain we are in an estimate,hen to test hypotheses, to say something about whether one variable causes another, we now have a way of thinking aboutthose questions. Statistical analysis is indispensable in various fields, such as medical research, quality control in the production industry, development of public policy, and research in the social sciences. One of the general principles for relating theoretical model to reality is the law of large numbers. This is expressed that the average of the result grows close to the expected value as the number of the tasks is grown. For example, if you were to flip a fair coin only10 times, you could get 7 heads and 3 tails — something far from the expected 50-50 split. Yet ifyou flip it 10,000 times,thepercentageofheads will probably be a lot closer to 0.5. This is why casinos make a steady profit and arefinancially successful over time, since individual big wins can be averaged outto the casino's theoretical advantage. That randomness and unpredictable nature doesn't mean there is no pattern, not necessarily. In truth, random processes frequently show interesting and uniform phenomena when repeated for many iterations. Stochastic processes are used to model systems that advance in time according to an element of randomness. This is the case with stocks, the flow of particles in a fluid, or the spread of diseases within a population. These processes, however, may behave in nontrivial ways even though they arelargely governed by probabilistic rules. Knowledge of these patterns enables scientists and analysts to model and predict systems that would seem at first to chaotic or unpredictable to analyze. Probability is essential for science when using the concept of "statistical significance". Scientists who carry out experiments must decide whether the results they observe reflect an actual effect — an experimental value, such as the speed of light — or is just the result of random chance. Tests of statistical significance allow us to estimate the likelihood of the observed data under the assumption of no genuine effect (the "null hypothesis"). Ifthisprobabilityislowenough(usuallyunder5percent or1percent),scientists call the results statistically significant: The data suggest we're seeing something more than random chance at play.

This framework has been the foundation of the scientific method through the generations, though it is worth emphasizing that statistical significance doesnot equate to practical importance. Probability concepts are highly stressed in terms of risk assessment and management. Risk is such a thing for which we could express the probability of an eventual come upon of something bad and themagnitudeofthe bad thing was going to happen. In addition, insurersapply complex probability models to determine premiums that equilibrate infrequent large pay-outs against continuous income in the form of premiums. Engineers add in that margin of safety when they design systems. Risk assessments are employed by health care providers to identify those patients who are likely to be most in need of preventive interventions. Even personally, our intuitive senseofriskguidescountlessdailydecisions,fromhowfastweshoulddrivein various conditions to which investments might be appropriate for ourretirementportfolios.Probabilistictheorem hasdevelopedextraordinarilysince the advent of the computer, and computational techniques have crafted new horizons. Methods such as Monte Carlo simulation take random samples to approximate solutions to problems that are hard or impossible to solve analytically. For instance, a financial analyst could simulate thousands of potential future market realities to evaluate investment risks, or a physics researcher could use a random sample to estimate complicated multidimensional integrals. Most machine learning algorithms are based on probability theory, and learn without algorithms being specificallyprogrammed to do so through statistical patterns in data used for decision or prediction making. These computational methods have transformed everything from climatemodeling to AI. Probability is thehero ofgamesofchance."Card games and dice games and roulette and the lottery all provide those rules of probability. Knowing these rules isn't a guarantee of winning (the house edgeis designed to ensure that casinos never lose over the long term), but it allows players to make more informed decisions and avoid common misconceptions. One such example might be the basic strategy while playing blackjack is described by probability and can be used to lower the house edge. Poker is a game that is part probability and part psychology – the players need to assess the probabilities of different hands, as well as their opponents' likely tactics. Even basic children's games feature probability through dice or card draws.

In the early 20th century quantum mechanics brought probability to the verycore of our understanding of physical reality. While classical physics is deterministic, quantum physics is fundamentally probabilistic. The well-known Schrödinger's wave equation doesn't describe the exact position or momentum of a particle, but a probability distribution — of where the particle might show up when you measure it. This probabilistic character of quantum systems is nota feature of our measuring tools or of our state of knowledge, but rather of the reality itself, at the level of the quantum world. This in itself was revolutionary: it overturned centuries of Determinism and still provokes questions among philosophers about the nature of reality. Genetic transmission occurs in a probabilistic manner so that probability theory is inherently important to both genetics and evolutionary biology. The laws discovered by Mendel are the ones that clarify how characteristics get from parents to the next generation and not simply by accident but in a predictable ratio. For a simple cross between two heterozygotes for a trait, each child would have a 25% chance of both of its inherited alleles being recessive for the recessive trait if the alleles are independent. Population genetics employs probabilistic models to follow the evolution of gene frequencies across generations as a result of forces such as natural selection, genetic drift, mutation and migration. Such models help to account for whythe features of species are relatively fixed and why changeover time occurs. Decision theory provides a formal structure for optimal decision under uncertainty, where probability and utility (a measure for the value or satisfaction) are combined. When one has to make a decision with uncertain consequences, according to the expected utility hypothesis, one ought to choose the option with the greatest expected utility--the sum of the utility of each possible outcome (albeit weighted by its probability). This model can explain a lot about how humans make choices, from decisions about money to choices about health. But in behavioral economics, we have research showing thatpeople frequently do not adhere to this model of rational behavior, typically because of cognitive and emotional biases, or because their subjective assessments of probability don't match the actual probabilities. Information theory, established by Claude Shannon in the latter part of the 20 century,creates deep links between probability theory and the notion of information and entropy.

In this context, the information carried by message is a function of its unpredictability (rare messages carry more information than common ones). For example, getting a message that "the sun rose today" is at least almost uselessbecauseitisso likelyandhardlysurprising.Ontheotherhand,learning “your lottery numbers actually won” contains a huge amount of information, precisely because it’s so unlikely. These factors have applications in data compression, communication systems, cryptography and more recently, we have begun to understand its consequence in systems of biophysical interest suchasneuralnetworks, DNA,etc. Ithappensthatprobabilisticreasoning goes far beyond mathematics to affect in what mode we understand knowledge and certainty in everyday life But hold. The project of the Bayesian philosophy of science is to set rational acceptance on the firm bases of probability theory applied to questions of what is known, shown, or believed at any given time, where rational belief people think should behave according to the laws they have come to recognize for everyday life. From this perspective, beliefs should be constantly revised as new evidence occurs, in accordance with Bayes’ theorem. This is very different than the classical ”yes or no” approach to knowledge and treats knowledge as a matter of degrees of belief and respective confidences. This probabilistic generation of knowledge fits nicely with the way science works, which is to draw tentative conclusions tempered by an openness to new evidence. There are many situations where Probability meets ethics and fairness. However, when resources or opportunities are allocated according to some probabilistic assessment, insurance premiums, loan applications or predictive policing, questions about fairness and discrimination can comeinto play.Forexample, pricing insuranceon thebasis ofpostal codes might discriminate indirectly against some demographic groups that areheavily represented in certain neighbourhoods. Likewise, machine learning algorithms which predict future outcomes using past data, may end up replicating the existing biases. These challenges have created increasinginterest in “algorithmic fairness”, creating techniques to ensure that, for example, a probabilistic decision system treats people fairly while still making statistically accurate predictions. There are some interesting facts about human cognition in there. Years of science reveal that people are prone to systematic errorswhentheyreasonaboutprobability.Wethinkthatdramaticevents(a

plane crash) are more likely to happen than they are, while events that are less dramatic but more likely to befall us (car crash) are less likely to happen than they are. We see patterns in truly random sequences and fail to appreciate the role of chance in many outcomes. As our thinking drives those we consult to frame probabilities (the same medical procedure described as having a "90% survival rate" is more attractive than the one with "10% mortality rate") we become influenced by that framing. Knowing about these cognitive biases can help us to make better decisions in situations of risk and uncertainty. In the modern age, reading of probability has become even more important.Probability information about health risks, financial investments, weather forecasts,andelectionpolls,tonameafew,isconstantlypresentedtothepublic.

Misinterpreting probabilities can result in bad decisions with widespread ramifications. The interpreting badly of the results of medical screening can cause unnecessary anxiety or unwarranted courses of treatment. Similarly, failure to understand the margin of error in opinion polls can also produce confidence in the results of an election that may not exist. Better probability education could assist people in making more informed decisions about everything from personal health choices to policy preferences on complex societal issues. The idea of probability distributions generalizes to multivariate probability distributions, which cover situations in which multiple random variables are of interest at the same time. These joint distributions reflect not onlytheprobabilityofspecificoutcomes,butalsothedegreesofassociationthat may exist between variables. The correlation coefficient ranges from 1 to −1, indicates the strength and direction of a linear relationship between two variables, and 0 indicates no linear relationship. But correlation does not mean causation - this is a fallacy. And just because two variables Scaffidi discussesare correlated does not mean one is causing the other; it could be that both are affected by a third factor, or that the relationship is spurious. Appreciating these differences is important for correctly interpreting results of statistical analysis. Probability theory is still developing and new problems and applications are being addressed. One active area is the development of strategies for responding to extremely rare occurrences that, when they occur, can have huge effects — "black swans," in the metaphorpopularized by the finance expert Nassim Taleb.

Another frontier involves complex systems with many interacting components, where emergent behaviors can arise that are difficult to predict from individual elements. Yet another theme is the natural extensions of probability theory to describe structure and dynamics in networked systems (social networks, transportation systems, biological networks). These advances have only served to extend the range and relevance of probability. The probability theory is inthe center of more and more complex and larger applied AI systems. Most machine learning algorithms employ probabilistic models to cope with the uncertainties in data and to predict. Language processing systems for natural languages use this probability to decide which sense a word has in a given sentence. Computer Vision systems score the likelihood that a potential object is what it has been trained to detect. Learning from reinforcement is guided by probability to strike a balance between exploring unknown strategies and exploiting established effectiveones and is used to powersystems thatlearn by trial and error. These are some of the most advanced and useful applications of the theory of probability in operation today. As the manner in which societies have perceived chance, randomness, and uncertainty has changed, so has probability theory. In past cultures it was common to attribute casual events to the Gods or to "Fate." The evolution of probability in diverse cultures has stimulated early interest in the study of probabilities. Classical period During the Renaissance, scientists such as Leonardo da Vinci sought to understand the mathematics of probability, but it was Stevin who put it on a firm theoretical basis. The 20th century brought transformative extensions through links with statistics, physics and computer science, among other areas. This evolution endures to this day, and with it probability has wormed its way deeper and deeper into ways we perceive and interact with our complicated world. Objective and subjective interpretations of probability present key dividing lines in philosophy. The frequentist interpretation identifies probability as the relative frequency of the event occurring in a large number of trials, conducted in the same or over similar circumstances, in the long run. This view views the probability as a real property in the world that operates irrespective of human knowledge or belief. The posterior Bayesian perspective, in contrast, views probability as a degree of belief, which can differ between people given what they know beforehand and how they interpret the evidence.

This subjective view permits us to make objective probability statements about one-off occurrences which can not be reproduced (e.g. The chance that it will rain tomorrow"). Both views have their merits and utility, and contemporary probability theory is inspired by elements of both traditions. Probability theory offers a set of important tools for reasoning under uncertainty, but it has some very real limitations and can be abused.Statistical measures can create a false sense of precision or certainty if their limitations aren't understood. Probability calculations are only as good as the assumptions and data that go into them. Bastard models are endearing when they perform well, but catastrophic when they do not. And even perfect probability knowledge does not dispense with value judgments, if we actually knew what the precise probability of different outcomes was, we'd still have to decide which outcome we want. These limitations emphasize the need for complementing probabilistic reasoning with criticalthinking,domain knowledge, andethicalconsiderationswhen faced with crucial decisions. Finally, probability is one of the most potent intellectual tools available to humanity to make sense of, and macro-navigate, our lightning-strikingly uncertain world. Developed from a course for students of statisticsandpsychology,thisbookisrelativelyeasytoreadforanyonewithhigh-school- level math. It includes a variety of problems with numerical answers. It allowsus to interpret randomness, measure risk, update our beliefs in the face of evidence and make better decisions. At the same time, probability confronts us with the limitations of certainty and prediction. In a world in which we're constantly confronted with incomplete information and unknown outcomes, however, probability literacy provides the route toward a more rational, nuanced and effective engagement with life's essential vagaries. In embracing probabilistic thinking, we are not sacrificing certainty for uncertainty, butsimply offloading some of the complexity into a framework better designed to deal with it. It seems to me that the yield is not complete and utter certainty (which may, in any case, be a mirage), but something just as valuable: a systematic way of navigating through the uncertainty that is essential to our personal and collective futures.

2.1.1 PracticalApplicationsofProbabilityinDailyLife: Probability concepts permeate our everyday lives, often in ways we don't immediately recognize. Take weather forecasts, for instance, which we consult almost daily. When meteorologists predict a 30% chance of rain, they're indicating that, based on current atmospheric conditions, similar weather patterns have historically resulted in rainfall about 30% of the time This likelihood information informs our practical decisions – to take an umbrella, rearrange outdoor plans, be ready for interruptions. The more we know about these probability statement, the better poised we are to make sense of them and to take measures without overreacting or underreacting to the forecast. Another field in whichprobability ideas are tangible is in the realm of personal finance. Capital allocation choices always come with an element of unknown associated with future return. Diversification, or spreading out investments among different types of assets, mitigates risk specifically because it's unlikely for all investment categories to perform poorly at the same time. Likewise, decisions with insurance are also a kind of intuitive probability reasoning. We buy insurance to guard against scenarios that are unlikely but potentially catastrophic, such as house fires or the diagnosis of a serious illness. The insurance firm charges premiums against the odds of these events and consumers agree to the protection depending on how much they care about the risks how much they are willing to pay. Even basic budgeting incorporates probability as we budget for variable items that vary and we cannot predict from month to month. Many healthcare decisions need to make judgements using probability (although often implicitly, not explicitly). The trade-off in deciding whether or not to undergo a screening test include our prior probability (pretest probability) of thecondition inquestion, the sensitivity(the pretestpositiveprobability)ofthetesttodetecttheconditionifpresent,andthe specificity (the pretest negative probability) of the test to determine that the tested person does not have the condition if the condition is truly absent. Understanding these probabilities can help patients and doctors make decisions about testing and treatment. And behaviour, such as the decision to drink ornot, like diet, exercise and smoking, means assessing trade-offs between probabilities of health states and immediate benefits/ convenience. Althoughwe do not perform these probability computations in a conscious manner, such

intuitive estimates underlie many health behaviors. Transport and travel planning are using probability in different shapes of forms. When we're deciding what time to leave for an important meeting, we naturally take into consideration the possibility of delays - adding some buffer time if we're on the road during rush hour, say, or when the weather is bad. Those GPS navigation apps that have always given estimated arrival times now show ranges of times, to account for uncertainty in conditions. Airlines overbook flights based on the expectation that some people won't show up, weighingthe costs of occasionally having to pay for passengers they have to bump against the extra money they make by flying with fuller planes. The same holds true for connections between flights or trains, as when travelers withhalf a brain plan these transfers they factor in buffer time based on the likelihood of delays, knowing that tight connections raise the likelihood of missing a subsequent departure. Social life is full of probability calculations, even if we don't normally consciously think of it like that. When we read that someone commented on something, and that comment was sincere or sarcastic, we make a probability judgment based on the context, and possibly the tone, our knowledge of the person, and so on. Choosing whom to date and whom to lay are estimates of compatibility and long-term success derivedfrom available information. In a professional environment, we could also strive to maintain connections with individuals most likely in the future to offer opportunities, or offer information. See even routine decisions about what you can and can't bring up in small talk amount to lightning fast assessments of what the other person will and won't tolerate. Consumers decisions often rely on judgements about probability. For consumers, the decision to buy an extended warranty comes down to how likely a product is to fail and the cost of the warranty. When we decide whether, say, to buy a name-brand product instead of a cheaper alternative we haven't tried, what we're often doing is making intuitive probabilityestimates both of qualityand of how satisfied we'll be with the decision later. Deciding how much fresh food to buy is a matter of what you think the likelihood of eating it before it spoils. Purchasing decisions in online shopping involve assessments of the trustworthiness of merchants, the truthfulness of product descriptions and the chanceofreceivingtimelydelivery.Thesearenotnecessarilynot

computations of formal probabilities, but they represent probabilistic reasoning. In reality, tasks around the house use probability in various applied forms. Homeowners have to determine which preventive maintenance steps are a good value — in part, based on the likelihood and expense of problems that could otherwise arise. For example, one's choice of frequency of gutter cleaning is responsive to this person's risk of water damage resulting from clogged gutters. And the same is true of these decisions on when to replaceold appliances; it's a trade-off between the increasing chance of death and the cost of a new one.

But basic prudent acts in the home, like having extra light bulbs, batteries or pantry staples on hand, acknowledge some probability that a need will ariseoneday,evenifyoucan'tknowforsurewhenyou'llneedthem.Choicesabout education andcareers requirea varietyof sophisticatedprobabilityjudgements. When students pick a major or a course of study, they consider their relative "likelihood of success" in different fields, the availability of jobs in the future, and potential earnings. When it comes to deciding whether to switch jobs or careers, workers weigh the likelihood of positive outcomes against the risks they face in making a move. The choice to further your education or training is partly based on your estimate of the investment in your future in return for a higher-paying job or more personal fulfillment. Even if such estimates are neververyprecise,they areatleastoneexample ofprobabilisticthinkingabout the uncertain future. Social media and knowledge sharing are based on probability judgments of accuracy, and relevance. At a time when we're all overloaded with information, and disinformation, those who read, view and listen to the media should always be questioning how good the source really is, and what the likelihood is that what it's presenting is correct. Multi-sourcing is an - if one independent source confirms, the likelihood of truth increases. Likewise, when we choose which news stories to open or which videos to watch, we are gambling very rapidly on the likelihood that this content will be most valuable or most entertaining as we make a lightning-fast probabilistic calculation from titles or previews and our past experience with similarcontent. Pleasant pastimes frequently involve challenges that are presented in a probabilisticcontext.Mostboardandcardgameshaveanelementofluck,

where good strategies require a good assessment of probabilities. Fantasy sports participants choose players based in part on probability evaluations of future performance. Gardeners “zone plant,” using hardiness zones to determine which plants are likely to survive in different climates. Weather forecast determines what the outdoor-activity enthusiast does. Probability is also at play when we watch TV as we predict if we’re likely to like a new series before we hit play. These uses of probability thinking during leisure time enrich and are enjo- yable. There are about a gazillion probability judgments in cooking and cooking-like activities. Similarly, experiencedcookshavean almostintuitivesenseforhowlikelyitisthatcertaintechniques willachievetheirdesired results.Atthetimeofmealplanning,itisdifficultto estimate if there will be enough time and energy to execute a planned meal on a certain day.

Good food storage is a judgment of the likelihood of needing somethingversus the risk of it going to waste. Probability enters recipe following, too, as cooks manipulate technique in light of the likely behavior of their specific ingredients and equipment. These problem solving situations with food emphasize the ubiquitous nature of probability thinking in ordinary life. Probability is used in energy use and conservation. Thermostat setting decisions trade-off comfort against energy cost, and programmablethermostats can be used to have different settings depending on the likelihood of occupation. With investments in energy-saving appliances or home-strengthening upgrades, it's a matter of gauging whether you'll save enough over time to make it pay. If nothing else, even little behaviors, like switching off lights when you leave rooms, convey a probabilistic computation of the oddsofreturninashortterm.Withworriesabouttheclimateontherise,more andmore consumers are takingpersonalresponsibilityfortheir energy choices — from thecarstheydriveto thelight bulbs intheirlamps. Being aparent is a constant risk assessment of child safety, development and well-being. Thetrick for parents is balancing the fact that it's very unlikely their child will be seriously hurt running around at the playground with the developmental value of letting the child take measured risks and experience some independence. Thejudgmentsconfrontedalsoincludewhenchildrenarejudgedcapableof

new privileges or responsibilities, and these are probabilistic judgments also. There are times when even simple decisions, such as how much food to cook or at what point during the day to set out for a day at the beach, depend to at least some degree on conjectures based on past experience of the likelihoods of various outcomes. This is also an inherent part of good parenting – adjusting these probability estimates as children age and acquire new skills.

Evidence of probability thinking in daily life are time management strategies. When making to-do lists or schedules, we already take into account the likelihood of finishing things according to schedule. Decisions about what do to first are often made not merely as a matter of importance, but as a function of how bad things will get if a task is held off. Padding time between patients recognizes the likelihood that things don't go as planned. Some even involve choices about when to multitask and when to instead attend to one activity, with a consideration of the likelihood of errors or inefficiency when attention is fragmented. For all of this to work, and to use inferential perspective, one would need to make good judgements of the probabilities of both how long tasks will take, and how likely they are to be completed. Depending on many things to which they can't subscribe to probability, and which, if they could, would result with deterrence-which are to say, lives. Speed limits are established in part based on the likelihood and severity of accidents atdifferent speeds. Defensive driving strategies aim to lower the risk of such collisions by properly educating and understanding the dangers associated with driving. The "three second rule" to maintain distance from the vehicle in front makes driving safer, and takes into account the fact that vehicles in front might suddenly stop. Probability is even used in the design of highway systems, as can be seen in such features as merge lanes, traffic circles, and signal timings to reduce the probability of collisions. Local routing decisions, timesofdepartureand arrivalallseemtobeattemptstocompromisesbetween the time that we spend travelling and the odds that we're going to have acrash. Totting up includes delicate probability judgements about what the recipient would like and how they would react. People who aregood at giving gifts tend to be good at predicting the likelihood that an individual will liketheparticularthingonebuys.Giftreceiptsarerecognitionthatjudgesthese

questions and allows them to be revisited if the guess work reflected by them is invalidated in reality. Price ranges on gifts are generally based on a measuring of the importance of the relationship in compromise with the likelihood that items falling within a particular price range may be found. Even choices of when to give a gift card versus a specific item are probability judgments about what the recipient wants, and what the giver knows about what the recipient wants.

2.1.2 Foundations:DefiningProbabilityanditsCoreConcepts

At its most fundamental, probability is measureof how likely an event is to occur. This frameworkallows measuring uncertainty and decision making in thepresenceof randomness. It's,inaway, amathematicallydistilled knowing numberthat tells you how likely it is that something will happen, which is to saysomewherebetween0(impossible)and1(certain).Probabilityisinvolved inallthingsin our lives,such aspredictingthe weather, diagnosinga person's disease,andeventhewinningscoreofgamesandtheclosingpriceof the stock market. To talk about probability, we first needto establish some fundamental concepts. An experiment is simply a method or action that produces an observableoutcome. Thecollection of all possible outcomes of an experiment is called sample space & is usually denoted as S. An event is subsetofsamplespacethatdescribesasingleoutcomeoroutcomes.collection to givean example, consider flipping of a coin. The sample space is {Heads, Tails}.Forexample,thissecondevent"gettingheads"isdefinedastheset, {Heads}. P(A)= fraction of favorable number outcomes divided by the total number of possible outcomes when all things are equally likely. In case,P(A)= n(A)/ n(S); where n(A) is number of events in event -A, & n(S) is number events number in sample space S. That is classical definition of probability which assumes that all possible outcomes an experiment havesame chance of happening, regardless of how likely they areto occur. Instead, weusetheempiricaldefinitionofprobability(or relativefrequency approach) in situations where probabilities of outcomes are not equal. This is like establishing probability of an event based on empirical data. Thus, the empiricalprobability,accordingtotheempiricaldefinitionofprobabilityis

given as: If an experiment is repeated 'n' times & event 'A' occurred 'm' times, then empirical probability of A is approximated as P(A) = m/n. As 'n'becomes large, the empirical probability of A converges to the true probability. To demonstrate this, let us take theexample of rolling a fair 6- sided die. Classical probability is given by ratio of number favorableoutcomes to the total number of possible outcomes, such as the statement, (numberoffavorableoutcomes/rolling4)/(numberoutcomes/1,2,3,4,5,6)
=> number of favorable outcomes = 1, number of outcomes = 6 and, thus, probability of rolling a '4' = 1/6., when we roll the die 100 times and get '4' 18 times then the empirical probability= 18/100 = 0.18, and it is pretty close to classical probability 1/6 (≈0.1667). So, we raise the number of rolls say to 1000, and observe 1000 rolls. We wanted the empirical probability to becloser to 1/6. The code simulates this process. You have now mastered conditions andloops now let's write a code, that simulates 1000 rolls of a die and tells you the empirical probability of an even number being rolled. And sure enough, the result of the run (for example 0.505) is quite close to the theoretical probability of the outcome of 0.5 (i.e., three even digits of six possible outcomes). This illustrates that classical probability can approximate empirical probability, withmany high numbers of trails.

andpro videsa general indicati onofho wthelik elihood ofanev entalter swith

2.1.3 ConditionalProbabilityandIndependence:In numerous real-world scenarios, events are interconnected rather than isolated.Conditional probability refers to the likelihood of an event (A) occurring, contingent upon the occurrence of another event (B).This allows us to modify our predicted odds as new information emerges.P(A|B) denotes the conditional probability of event 'A' occurring provided that event 'B' has transpired.P(A|B) = P(A ∩ B) / P(B)For instance, drawing two cards from a regular deck of 52 cards without replacement exemplifies a straightforward scenario.The probability that the second card is a king, given that the first card was a king, is defined as follows: let A represent "the second card is a king" and B represent "the first card is a king".In the first scenario, there are 4 kings in a deck of 52 cards, hence P(B) = 4/52.Assume we select a king.Among the remaining 51 cards, only 3 are kings.Thus, P(A|B) = 3/51.The latter refers to the preceding event

the occurrence of prior events.Conversely, independent events are occurrences whose consequences do not affect one another.Events A and B are considered independent if P(A|B) = P(A) or, equivalently, P(B|A) = P(B). Mathematically, P(A ∩ B) = P(A) * P(B).We can commence by flipping a coin twice.The outcome of the initial flip does not influence the outcome of the subsequent flip.The critical inquiry is the result of the second flip, which is entirely independent of the first flip's conclusion, whether heads or tails, despite the game's total being 1/2.Let A represent the event of obtaining heads on the first flip, and let B denote the occurrence of obtaining heads on the second flip.Therefore, P(A ∩ B) = P(A) * P(B) = (1/2) *The likelihood of achieving heads on both flips is (1/2) × (1/2) = 1/4.The law of total probability asserts that if the occurrences B1, B2, …, Bn constitute a partition of S (being mutually exclusive and collectively exhaustive), then for each event A, the equation P(A) = P(A|B1)P(B1) + P(A|B2)P(B2) + … + P(A|Bn)P(Bn) is valid.This will assist us in deconstructing the problem into smaller components. To illustrate,consider a factory that has two machines, M1 and M2, that make light bulbs. Let the machines be M1, M2, M3.Machine M1 makes 60% of the bulbs it produces,which has a 3% fault rate. Machine M2 makes 40%of the bulbs, 5% of which are defective. If a light bulb isselected at random, what is the chance that it will be defective? We will let A be the event thatyou get a faulty bulb. We are given P(M1) =0.6, P(M2) = 0.4, P(A|M1) = 0.03, and P(A|M2) = 0.05. Using law of total probability:P(A)=(0.03*0.6)+(0.05*0.4)=0.018+0.02=0.038.

Therefore, the probability of a randomlydrawn bulb being defective is 0.038 or 3.8%.

2.1.4 RandomVariablesandProbabilityDistributions:ModelingRandomPhenomena

We introduce random variable to formalize the Manera of handling and analyzing random phenomena. A random variable is set of values whose valuesarethe numerical outcomesofstochasticevent. Itis gottenon:sample space real numbers. Random variable is either discrete or continuous. This termtypicallyreferstoacountablyinfiniterandomvariablewithvaluesthat

might include, for example,the number of heads flipped after tossing coin n times, or the number of bits of a broken part produced by a machine. In the case of continuous random variable, it can take infinitely many values in certain range (x (e.g., height of a person, temperature of a room, etc.). Each random variable is associated with probability distribution that describes likelihoods of its possible values. In common, the chance distribution for discrete random variable is defined via a chance mass serve as (PMF), asmanyprobabilitiesassigned to everypotential value. Takethe simpleexample of flippingfair coin three times. Let us say that number of heads,say X, is random variable. As a result, X can take on values 0, 1, 2, 3. The random variableXhasprobabilitymassfunction(PMF):P(X=0)=1/8;P(X=1)=3/8; P(X=2) = 3/8; P(X=3) = 1/8. In case of a continuous random variable, the probability distribution is defined by a probability density function (PDF) which describes relative likelihood of the random variable taking on a given value. Between two points under the PDF curve lies the probability that our randomvariable belongs to that interval. Itrepresents one of the most widely used continuous probability distributions, commonly known as the normal (or Gaussian) distribution and represented with statistics favorable curve. Normal distribution is commonly used to approximate certain distributions; for example, weight, height, and exam scores. E(X): Expected Value of aRandom Variable Expectation or mean of random variable E(X) represents expected value of random variable, which we can define as a variable that takesonrandomvalueaccordingtosomeprobabilitydistribution.

UNIT 2.2
CONCEPTSOFPROBABILITY(CLASSICAL,EMPIRICAL,ANDSUBJECTIVE)

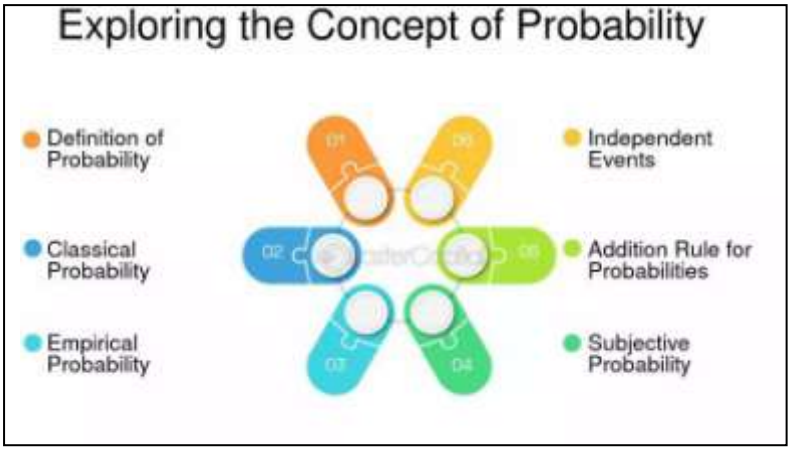

Figure2.2:ConceptsofProbability

2.2.1ClassicalProbability:TheRealmofEquallyLikelyOutcomes

Classical probability, alsoknown as a priori probability, is founded on basisof equal likelihood of alloutcomes of an experiment. Thisworks onlyin very specific situations such as coin tosses, dice rolls, card draws. The definition states that probability of an event (A) is number ratio of positive outcomes (n(A)) to total number of possibleoutcomes (n(S)

Mathematically,thisisrepresentedas:

P(A)=n(A)/n(S)

Classical probability works because of its simplicity, its logical foundations. However, its limitations should be appreciated. It depends on our perfect fairnessand symmetry, neither of which necessarily exists in the real world.

NumericalExample1:RollingaFairDie

Considerstandardsix-sideddie. Whatisprobabilityof rollinganeven number?

- **TotalPossibleOutcomes(S):** {1,2,3, 4,5,6} =>n(S)=6
- **FavorableOutcomes(A):**{2,4,6}=>n(A) =3
- **ProbabilityofRollinganEvenNumber:**P(A)=3/6=1/2or0.5or50%

NumericalExample2:DrawingaCard

What is probability of drawing an Ace from a standard deck of 52 playing cards?

- **TotalPossibleOutcomes(S):**52cards=>n(S)=52
- **FavorableOutcomes(A):**4Aces=>n(A) =4
- **ProbabilityofDrawinganAce:**P(A) =4/52= 1/13

Explanationextension: When we are learning these terms there is other one term that we have to understand that is SAMPLE SPACE. In probability theory, sample space is set of all possible outcomes in a stochasticexperiment. So, in the dice problem above, the sample space would be {1, 2,3,4, 5, 6}.So, the sum of all possibilitiesin thesamplespacemustbeequal to

1. A die with six faces stands a 1/6 chance of falling on any one of the facsimiles on its six sides. For example, by adding 1/6 6 times, you obtain 1. One may next consider the case of classical probability. Classical probability has a nice property when it comes to those things where we would expect true random outcomes, like many games of chance.

2.2.2EmpiricalProbability:LearningfromObservations

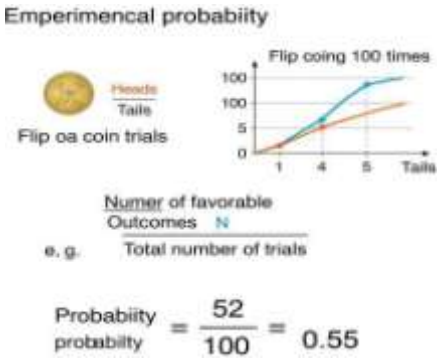

Figure2.3:Empiricalprobability

Empirical probability: It is based on observed data and previous experience; alsoknownas relative frequencyprobability. Itis abouthowlikely aneventis based on how oftenit appeared in trails.

Theformulafor empiricalprobability is:

P(A)=Number oftimeseventAoccurs/Total numberof trials

This is convenient for instances where an application of classical probability cannot be applied due to the fact that there isn't anequally likely outcome. Such as predict weather patterns, predicting failure rate from manufactured products, analyzing customer behavioretc.

NumericalExample3:CoinTossExperiment

Assume you flip a coin 100 times & record 53 heads. What is empirical chance of obtaining heads?

- **NumberofTimesHeadsOccur:**53
- **TotalNumberofTrials:**100
- **EmpiricalProbabilityofHeads:**P(Heads) =53/100=0.53or53%

NumericalExample4:ManufacturingDefects

A factory produces 10,000 units of certain product. Upon inspection, 250units are found to be defective. What is empirical probability of a product being defective?

- **NumberofDefectiveUnits:**250
- **TotalNumberofUnitsProduced:**10,000
- **EmpiricalProbabilityofDefect:**P(Defect)=250/10,000=0.025or 2.5%

Explanationextension: This is avery usefulmethod to analyze theoutcomes of events for which equal probability of all outcomes is not possible and classicalprobabilityisnotapplicable.Example:Weatherpatternprediction,

Failure rate predictionof manufactured products, Customer behavior analysis etc.

2.2.3 SubjectiveProbability:TheRoleofPersonalBeliefs

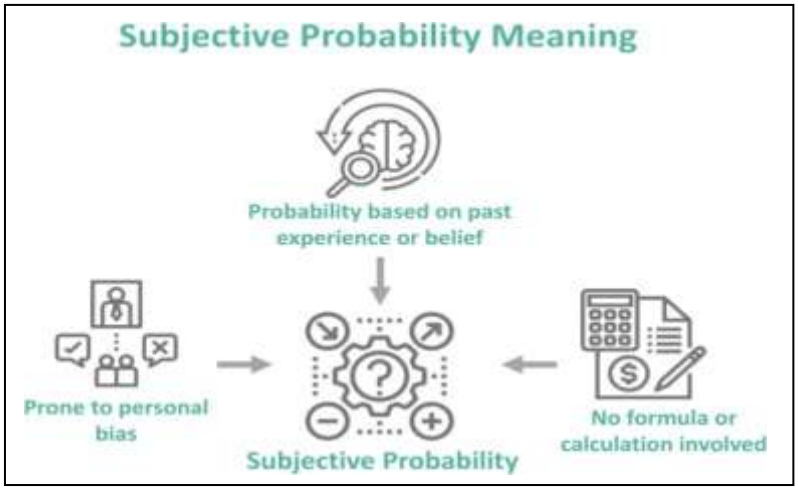

Figure2.4:SubjectiveProbability:TheRoleofPersonalBeliefs.

And this is especially true for rare or unprecedented events for which objective data are scarce or nonexistent. Subjective probability is estimating theprobabilityofsomethingbasedonhowpeoplefeelandwhattheyknow.It is and often in context such as predicting the success of a new businessventure or the outcome of a political election, or the likelihood of a rare medical condition.

NumericalExample5:StartupSuccess

An entrepreneur thinks that their startup will be successful70% of the time due to their market research, experience, and instinct. This is a subjective probability assessment.

- P(StartupSuccess) =0.70 or 70%

NumericalExample6:MedicalDiagnosis A doctor decides that there isa 10%chance,basedonapatient'ssymptoms,medicalhistory,andhow

common the disease is, that the patient has a rare disease. Notethat this is a subjective probability estimate.

- P(RareDisease) =0.10 or 10%

Explanationextension: Ofthethree,subjectiveprobabilityisthemostpoorly defined (and therefore the most contentious), because it is so dependent on individual bias. Two very different people who have access to different informationmight determine very different levels of probability for the exact same event, and be correct. Thus, we often use subjective probability, when objective facts cannot be established. Though individual opinionsvary, they remain helpful in risk assessment, and decision making. We, in a lot of different professions, rely on experience, and judgement to make decisions about likelyoutcomes.

2. InterplayandApplications:BlendingtheApproaches

Conditional ProbabilityWhen discussing thedifferent typesofprobabilities,it is worth mentioning that in many ordinary life situations classical, empirical andsubjectiveprobabilitiesareusedsimultaneously.Forinstance, supposean insurance company wants to calculate risk of its clients to get in a caraccident: It could use classical probability example to measure the probability of accidents, use empirical probability to assess historical claim data and use subjective probability to accounts for individualsrisk profile.

Requiring knowledge about and application of these perspectives of probability is critical to making informed choices in many domains,including:

- Finance:Pricingfinancialinstruments,evaluatinginvestmentrisks.
- Medicine:Diseasediagnosis,treatmentefficacy assessment.
- Engineering:Studyingsystemsreliability,safetydevelopment.
- Business:Salesprediction,marketingcampaignoptimization.
- Science: Statistical analyses,interpreting experimental results

However,doyouknow whatispowerfultoolthatallowsyouto betterdeal with uncertainty andmake sound judgment in a dynamic world by mastering theconcepts of classical, empirical, and subjectiveprobability? It is one of the basic corner stones of statistical analysis,and its principals are useful in ourdaily life.

UNIT 2.3 PROBABILITYLAWS

2.3 PROBABILITYLAWS

ProbabilityLaws:NavigatingtheRealmofChance

1. TheAdditiveLaw: The Additive Law The additive law of probability is critical to calculating the probability of one event or another event. This theorem applies significantly to the cases of mutually exclusive events and non-mutually exclusive events. Mutually exclusive events cannot occur simultaneously, while nonmutually exclusive events can. Disjoint Events (or mutually exclusive events) If A and B are two events which cannot happen at the same time P(A or B) = P(A) + P(B). Mathematically, we interpret this as:P (A or B) = P (A) + P (B)

This fits with what we'd expect to happen according to common sense. In cases where two events cannot both happen at the same time, the probabilityof either occurrence is just the sum of their probability as separate events.

Illustrative Example: Utilize a standard six-sided die.Let event A denote the occurrence of rolling a 2, and let event B denote the occurrence of rolling a 5. The occurrences are mutually incompatible, as it is impossible to roll both a 2 and a 5 simultaneously in a single throw.

P(A) = 1/6 (probability of rolling a two) P(B) = 1/6 (probability of rolling a five)

Applying the additive law: P(A or B) = P(2 or 5) = P(2) + P(5) = 1/6 + 1/6 = 2/6 = 1/3

Consequently,the likelihood of rolling eithera2 or a5 is1/3.

When events are not mutually exclusive, meaning they can occur simultaneously, the addition law must be adjusted.NOTICEDue toinstances where both events occur, it is necessary to eliminate them to avoid double counting.The equation is expressed as: P(A or B) = P(A) + P(B) -P(A & B),P(A ∩ B) denotes the intersection of occurrences A and B, representing the probability that both events occur simultaneously.

NumericalExample:

Imagine you are drawing a card from anormal 52-card deck. Let A beevent of drawing heart, & B be event of drawing king. [Because one can drawthe king of hearts.

- P(A)=13/52=1/4(probabilityofdrawing heart)
- P(B)=4/52= 1/13(probability ofdrawing king)
- P(AandB)=1/52(probabilityofdrawingkingofhearts)

Using the additive law for non-mutually exclusive events:

P(A or B) = P(heart or king) = P(heart) + P(king) - P(heart and king) P(A orB) = 1/4 + 1/13 - 1/52 = 13/52 + 4/52 - 1/52 = 16/52 = 4/13

So,the chanceofdrawing aheart oraking= 4 /13

The additive law isindispensable from figuring out the chances of winning a lottery to assessing the odds of contracting a disease. It helps us to create scenarios and calculating the possibility of joint events happen that than the foundation of our informed decisions.

Figure2.5:additivelawofprobability

2. TheMultiplicativeLaw:DeterminingtheProbabilityof"Both/And"Events

The multiplicative law of probability concerns probability of simultaneous occurrence of two or more events. This is especially important whencalculating independent and dependent events. Dependent Events: An event that has the property that the prediction of one event affect another event.

IndependentEvents:

For independent events that involve A & B, thenchances for both the events to happen will be simply the multiplication of probabilities of A & B. Mathematically,this is expressed as:

P(A&B) =P(A)* P(B)

Theideaisthat =totalprobabilityofajointeventisproductofprobabilitiesof its component events which occur independently of each other.

NumericalExample:

Example1: Tossing a fair coin twice Let A be the event that we getheads on first flip, &B be event that we get heads on second flip. The result of one flip does not affect the next; theseevents are independent.

- P(A)=1/2(probabilityofheadsonfirstflip)

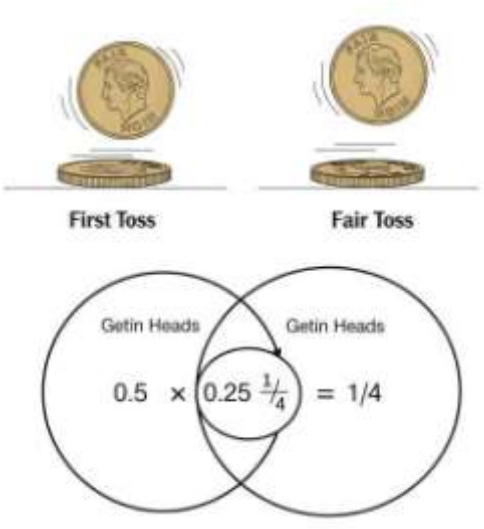

Figure2.6:MultiplicativeLawofProbability

- P(B)=1/2(probabilityofheadsonsecondflip)

Using the multiplicative law:

P(A&B)=P(heads&heads)=P(heads)*P(heads)=1/2*1/2=1/4 Therefore, probability of getting heads on both flips is 1/4.

DependentEvents:

For dependent events, where one event has an impact onprobability of other. The multiplicative law is based on conditional probability P(B|A), The probability of event B occurring, given that event A hasalready happened.The equation is expressed as::

P(AandB)=P(A)* P(B|A)

Thisformulationaccountsfordependencyamongtheevents,adjustingthe likelihood of the second event given the first.

NumericalExample:

Let us think about drawing two cards from a 52-card deck without replacement. Let event A be that we draw a king on the first draw, and event B be that we draw a queen on the second draw. But they are dependentevents, because the result of your first draw directly (albeit indirectly) determines the contents of the rest of the deck.

- P(A)=4/52 =1/13 (likelihoodofselecting akingoninitial draw)
- P(B|A)=4/51(thelikelihoodofdrawingqueenonseconddraw, contingent upon a king being drawn first)

Usingthemultiplicativelawfordependentevents:

P(A & B) = P(king & queen) = P(king) * P(queen king) P(A and B) = 1/13 * 4/51 = 4/663

So the probability of drawing, without replacement, a king followed by a queen would be 4/663.

One of the most important laws in standalone form is known as the law of multiplication,itisappliedinmanyofthesciencefieldslikegenetics,

finance, engineering, etc. It allows us to deduce probabilities of complicated events by breaking them up into simpler, subsequent stages. Understanding whether events are dependent or independent is essential to wisdom of the appropriate implementationof this law.

3. IntegratingAdditiveandMultiplicativeLaws:Real-WorldApplications They are not exclusive laws and most of the time you use them inconjunction tosolveaproblemoncomplexprobabilitysolving.Therearetypicallytwo halves of real-world cases “either/or” and “both/and” “conditions” that should be reconciled.

Example:QualityControl

Let us consider an example of such a situation we have a manufacturingprocess where two machines, M1 and M2 produce items: Machine M1occupies 60% of the product and have defect rate = 2%Machine M2 occupies 40% of the product and have defect rate = 3%.

We are interested in getting the probability for randomly chosen item being defective.

Let:

- A=itemproducedby M1
- B=itemproducedby M2
- D=itemisdefective

We have:

- P(A)=0.60
- P(B) =0.40
- P(D|A)=0.02(probabilityofdefectivegivenitemfromM1)
- P(D|B)=0.03(probabilityofdefectivegivenitemfromM2)

We need to find P(D). We can use law of total probability, which combines the additive and multiplicative laws:

P(D)=P(DandA)+P(DandB)P(D)=P(A)*P(D|A)+P(B)*P(D|B)P(D)
=(0.60* 0.02)+(0.40*0.03)P(D) =0.012+0.012 P(D) = 0.024

Therefore, the probability that a randomly selected item is defective is 0.024 or 2.4%.

This will give a you an example of how the additive and multiplicative laws come together. By knowing and understanding these basic laws that willallow us to record andanalyze uncertainty and make smart decisions. More specifically these probability laws underlie complex probabilistic models and statistical analyses that are employed to better understand the inherent randomness in the world around us.

UNIT 2.4 DECISIONRULEINPROBABILITY

2.4 DecisionRuleInProbability

Deciding under uncertainty is a fact of human existence. Whether it is adoctor diagnosing a patient, a financial analyst predicting prices and future market trends, or a weather forecaster estimating the chance of rain, having to decide (for those responsible for the decision) the right option out of a limited (or vague) amount of information isa fundamental task. Inorder to measure and handle this uncertainty, we turn to math: probability. In effect, a decision rule is a rule-based assumption used to make a decision based onprobability of the occurrence of certain events. It bridges subjective probabilities with tangible actions; less-than probabilities translate into objective choices. Probabilistic reasoning in fact giving numeric values, of probability, towhatis to happen. These probabilities provide an idea on the basis of available information or based upon previous experiences or deduction. As an example, flipping fair coin, we would saythat event heads have a probability (0.5 or 50%) and the event tails (0.5). Reality is not always so convenient. Thatmeans therearefrequently situations whereprobabilitiesareunknown,orthey varywith newinformation.Andthen enterdecisionrulesandthe mechanistic way of making decisions even when faced with ambiguity.

A decision rule usually involves four components: (a) a description of the possible states of the world, (b) a description of a probability distribution over those states, (c) a set of possible actions, and (d) a description of a criterionfor selecting the preferred action (decision rule). This criterion is usually expressed in terms of minimizing expected loss or maximizing expected utility.TheExpectedutilityisanassessmentofhowattractiveacertainactis, and it can be how likely its sorted outcomes will appear, and the worth of those outcomes. Expected loss, on other hand, serves as an indicator of how much downside risk we are taking onby taking an action. So, let’s consider a simple example: A retailer needs to decide how many units of a perishable product to order. What they have to sell is unknown and excess product at the end of the day must bethrown out. The retailer can use historical transaction datatopredicttheprobabilityofvariousdemandlevels.Forexample,they

would consider a 30% probability of low demand, a 50% probability of medium demand, and a 20% probability of high demand. They can then compute the expected profit for different stocking levels and choose one that yieldshighestexpectedprofit.Thisishowyoucanimplement adecisionrule in real world.

And decision rules use thresholds(or some cut-off point). For example, a test for a medical condition might have a threshold probability over which a positive test would be clinically significant. If the probability exceeds this threshold,thedoctor mightrecommend furthertestingortreatment. Thisrule is a decision criterion that minimize false positive risk (treat a non-sick patient)against falsenegativerisk(missadiagnosis).Choosingthisthreshold is critical because anything in context and relative costs of errors matter.

2.4.1 BuildingRobustDecisionRules:ExpectedValue,BayesianInference,andRiskAssessment

Sound decision-making requires sound knowledge of probability theory and statistical methods. Beneath it all, one revolves around expected value. For every possible valueof X, one multiplies it by the probability of X being that value, and then they sum all the products to compute the expected value of X. It calculates the average outcome ofa random event over long period of time. Consider,for example, a lottery ticket that costs $1 and has a 1% chance of payingoff$100.Itwillhaveanexpectedvalueof(0.01*$100)+(0.99*-$1)=$1 −$0.99 =$0.01. That is to say, fortheaveragepersonwhobuyslots of tickets, they'll lose $.99 for every ticket they buy. Sure, some hypothetical someonecomesoutontopandwins,butintermsofexpectedvalue,thelong- term picture is bleak.

Bayesian inference is another strong way to use to create decision rules. It gives us the ability to update our beliefs about thelikelihood of events based onnewinformation.Thisisparticularlyusefulfor fieldswithknowledgethat is constantly changing. So, for example, a self-driving car might have initial beliefs about how likely a person will cross the same street and it could use informationcollectedfromsensorstoadjustthosebeliefsusingsomething

like Bayesian inference. For demonstration purpose let us take a numeric example. Considercase of a diagnostic test for a rare disease. The test is 95 percent sensitive (correctly identifies 95 percent of people with the disease) and 90 percent specific (correctly identifies90 percent of people without the disease). The disease affects 1% global population. If person tests positive, how likely is it that a theyactually have the disease?

EmployingBayes'theorem,wemayget theposterior probability:

- Priorprobabilityofhavingdisease(P(D))=0.01
- Priorprobabilityofnothavingdisease(P(¬D)) = 0.99
- Probabilityofapositivetestgivenhavingdisease (P(+|D)) =0.95
- Probabilityofpositivetestgivennothavingdisease(P(+|¬D))=0.10 The

posterior probability of having disease given positive test (P(D|+)) is: P(D|+)

= [P(+|D) * P(D)] / [P(+|D) * P(D) + P(+|¬D) * P(¬D)]

P(D|+)= (0.95* 0.01)/ (0.95 * 0.01+0.10 * 0.99)

P(D|+)=0.0095/ (0.0095+0.099)

P(D|+)=0.0095/0.1085

P(D|+)≈ 0.0876

And this means that even if you get positive test result, probabilitythat you actually have disease is roughly 8.76%. This highlights the delicatebalance between prior probabilities and test characteristics that must be struck when considering test results. Decision rule development is really a risk assessment process. This involves the process of identifying potential risks, assessing the probability and consequences of those risks, and developing strategies to mitigate those risks. This can be done using oneof many popular methods used for risk assessment, such as sensitivity analysis, scenario analysis, or decisiontreeanalysis.Sensitivityanalysisexamineshowvariationintheinput ofadecisionruleimpactsitsoveralloutput.Infact,scenarioanalysisenables

to scope out different scenarios while decision tree analysis provides a diagrammatic aid displaying the different pathways taken to arrive at a decision along with the probability and the payoff associated with each. Such techniques add more stability and caution tothe decision rules.

2.4.2 ImplementingandEvaluatingDecisionRules:PracticalConsiderationsandEthicalImplications

Youdo not train on data past said date,so you havereal business decisions to maketo train therules that matter. Useof poor-quality datacan neverbefixed by even well-trained algorithms, and in the absence of accurate and complete data, poor decisions are bound to be made. Some decision rules are computationally hard and require specialized algorithms and software. Furthermore, human judgment is often critical in the interpretation of probabilistic informationand final decision-making. Finally, the first instance in finance trading we can identify are algorithmic trading systems that are systems of decision rules that are programmed with the ability toautomatically execute trades based on market data and parameters fed inahead of time. Propelled by large datasets and sophisticated algorithms free of human bias, these systems can sniff out profitable trading opportunities. However, these systems still need an overseer, in the form of human traders,to be ableto monitor their performance and make adjustmentswhen required.

Performance assessment of decision rules is a fundamental issue for thereports ofviolence.Methodologieslikebacktesting,simulationand empirical experimentation are used to make this possible. Backtesting means applying a decision rule to past data to check how well it would have performed.Thatinanycasesimulationisawayofliterallymodelingasystem and then using that model you wrote to enter all kinds of various decisionrules into the model you just wrote. We call this approach real-world experimentation: an effort to implement a decision rule, under controlled circumstances, in the real world, and measure its impact. The creation and application of decision rules also raises ethical dilemmas. Some of thedecision rules devoured by AIs could have pernicious consequences or could entrenchbiasesalreadypresentinsociety. Forinstance,decisionrulesthatare

implemented in criminal justice systems can have unequal impacts on subpopulations. Decision rules have to be fair, transparentand ethical.

Furthermore, the increasing use of artificial intelligence (AI) and machine learning in decision making raises new ethical concerns. And while that its true, an AI algorithm can learn incredibly complex patterns from data; it can just as easily learn to amplify existing biases present in the data set. The challenge for us to ensure that algorithmic decision systems are fair, transparent and explainable. The takeaway: decision rules are a big-picture approach for dealing with uncertainty and making low-regret choices. What differentiates us is the ability to derive valid decision rules to optimize these outcomesthrough theuseofprobabilisticreasoning,statistical methodologies, and ethical constraints. As far as new trends in data science and AI are concerned, decision rules willbe an evolving pun.

UNIT 2.5 PROBABILITYDISTRIBUTIONS

2.5 PROBABILITYDISTRIBUTIONS

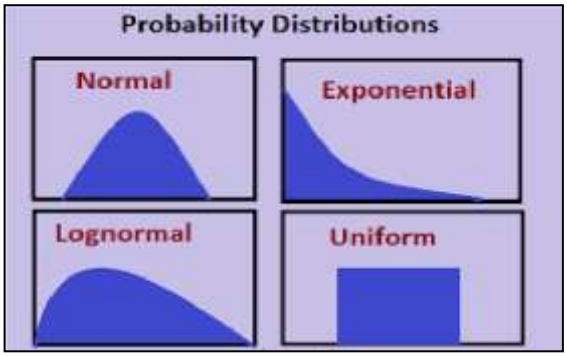

Figure2.7:ProbabilityDistributions.

2.5.1TheFoundation:UnderstandingProbabilityDistributions

Probability distributions form the bedrock for statistical inference and predictive modeling. They offer a mathematical structure for characterizing variousprobabilityoutcomes instochasticevent.Everypossibleoutcomeofa random variable has probability mass assigned to it by probabilitydistribution. The occurrence of random phenomena is an event whose fate is absolutely impossible to predict, yet this concept, albeit a little confusing, corresponds to the mathematical field of random variable, which isa variable amountthatvariesinaccordance withtheoutcomeoftherealevent.Thereare two types: discrete & continuous random variables. In contrast, discrete random variables have finite or countably infinite domain different values (e.g., the number of heads of coin tosses, the number of defects). The simple answer is that we are ultimately trying to get a better understanding ofthe uncertainty, and nothing captures the uncertainty better than the probability distribution. Instead of simply stating this event might happen, we canprovide a pros and cons of it happening. This enables us to take action and make predictions based on likelihood of different outcomes. PMF indicates probability corresponding toevery actual value of PMF. Discrete Stochastic Variables For continuous random variables, PDF (probability densityfunction)describesprobabilitydistributionofthecontinuousrandomvariable

and indicates relativeprobability that that random variable will equal a given true value. Knowing that CDF is found through integration of probability densityfunction.

One of major tools is cumulative distribution function (CDF). It represents probability that a random variable is nogreater than some specified value. The cumulative distribution function (CDF) generalizes to both discrete & continuous random variables. This is useful because predictive distributions only make sense if you understand what every type of parameter represents, so having a mental map of how they act and influence predictions will allow you to more easily navigate their practical functioning. The mean, or expectedvalue E(X) or μ, measures average value of the random variable, and the varianceσ^2=Var(X) measuresthe spread ofvaluesaroundthat mean. Assuch,

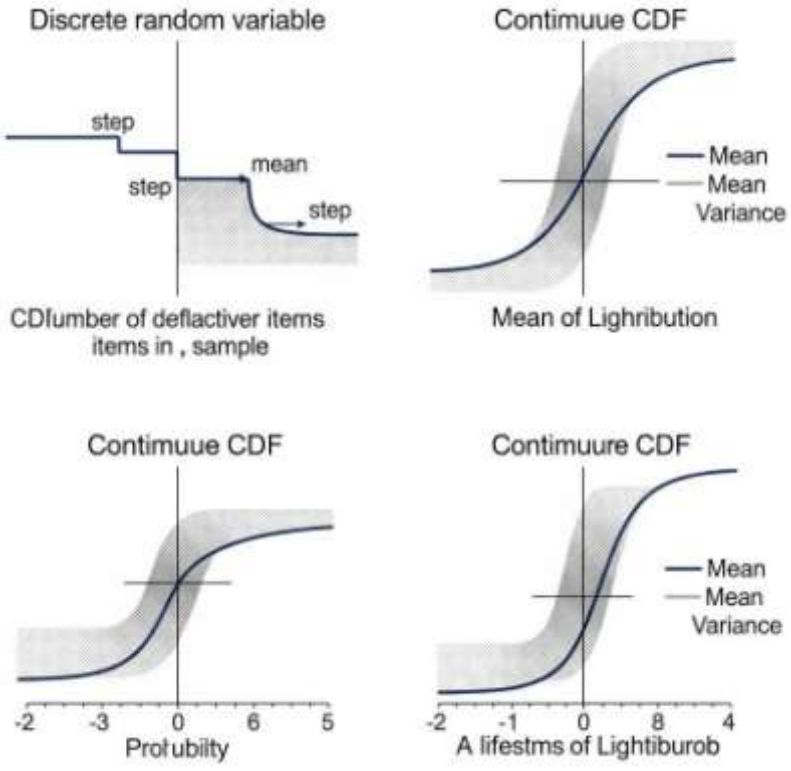

Figure2.8:CumulativeDistributionFunction(CDF)

these properties offer a complete picture of the distribution'sshape and whereit lies.

2.5.2DiscreteDistributions:BinomialandPoisson

BinomialDistribution:TheProbabilityofSuccesses

The Binomial probability distribution is type of probability distribution that describesnumberofsuccessesinfixedexperimental numbertrials.ABernoulli trial is a stochastic experiment (such as flipping a coin) that results in a binary outcome, with each possible outcome being assigned either the label of success or failure. These experiments are independent: The outcome of one trial does not influence outcomes of any other experiment. The fastest method is to take advantageoftheBernoullidistribution,whichreflectsaconstantprobabilityof success (p) on every trial. There are two key components of the binomial distribution, number of trials, n, & success probability, p.

ThePMFof binomialdistributioncan bewrittenas:

The probability formula they provided is probability mass function(PMF) of binomial distribution:

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

Rewritingitwithfactorialnotation:

$$P(X = k) = \frac{n!}{k!(n-k)!} p^k (1-p)^{n-k}$$

where:

- Xdenotesrandomvariablethatsignifiesquantityofsuccesses.
- krepresentsquantityofsuccesses (0,1,2,...,n))

- The binomial coefficient, denoted as (n choose k), signifies number of methods to select k successes from n trials. The calculation is expressed as n! / (k! * (n - k)!).
- pdenotesprobabilityofsuccessinone trial.
- (1-prepresentslikelihoodof failurein singular trial.

NumericalExample:

Consider fair coin being tossed ten times. What is likelihood of obtaining precisely 6 heads?

- n =10 numberof trials)
- k=6(quantity of successes)
- p=0.5(probability ofgetting head)

P(X=6)=(10choose6)*(0.5)^6*(0.5)^4P(X=6)=(10!/(6!*4!))*
(0.5)^10P(X=6)=210*0.0009765625 P(X=6) ≈ 0.2051

Thelikelihoodofobtainingprecisely6headsin10throwsisroughly0.2051. The

mean (expected value) of binomial distribution is expressed as:

The equation:

$$\mu = n \cdot p$$

Theformulaforvariance inabinomialdistribution is:

$$\sigma^2 = np(1-p)$$

PoissonDistribution:TheProbabilityofRareEvents

Its proof is beyond the scope of the present discussion; in a few instances,some authors employ some distributions, for example Poisson. The Poisson distribution is used to model events that are rare innature.

ForthePoisson distribution,there isoneparameterthatweneedtoconsider, λ (lambda),or averagenumberofoccurrencesingiven interval.

So, the probability mass function (PMF)of the Poisson distribution is givenby:

where:

- Xdenotesrandomvariablethatsignifiesquantityofoccurrences.
- kisnumberof events(0,1,2, ...).
- λis averagenumberofeventsingiven interval.
- eis
- baseofnaturallogarithm (approximately2.71828).

NumericalExample:

Forexample,ifcallcenterreceivesanaverageof5calls/min.λ=5(average numberof callsper minute)

- k =3 (numberof calls)

P(X=3)=(e^(-5)*5^3)/3!P(X=3)=(0.006737947*125)/6P(X=3)≈ 0.1404

Hence, The probability of getting exactly 3calls in minute is approximately0.1404.

Themean&varianceofPoissondistributionarebothequivalenttoλ: $\mu = \lambda$

$\sigma^2 = \lambda$

2.5.3ContinuousDistributions:NormalDistribution

NormalDistribution:TheBellCurve

Normal Distribution Also known as a Gaussiandistribution, it is continuousprobabilitydistributionthatissymmetricaboutitsmean,givingitabell-

shaped appearance. This makes normal distribution one of most important distributions in statistics because many natural phenomena and empirical data are often normally distributed. It is defined by two parameters, average (μ) & standard deviation (σ). The mean gives center of distribution and standard deviationgives distribution.

Thenormaldistribution isdefinedby itsprobabilitydensity function:

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

where:

- xisrandom variable.
- μismean.
- σisstandard deviation.
- πisapproximately3.14159.
- eisbaseofnaturallogarithm(approximately2.71828).

NumericalExample:

Let's say heights of the adult males in particular community are normally distributed with average = 175 cm & standard deviation = 8 cm. Finally, wecan standardize thevalue 190 cm using z-score formula:

First,weneedtostandardizethevalue190cmusingz-scoreformula: z = (x

- μ) / σ z = (190 - 175) / 8 z = 15 / 8 z = 1.875

Then we want P(Z > 1.875), with Z a standardnormal random variable with mean 0 & standard deviation 1. So, by looking at the regular normal distribution table or calculator, we see that:

P(Z>1.875)≈ 0.0304

Therefore, probability that randomly selected male is taller than 190 cm is approximately 0.0304.

UNIT 2.6 THEOREMSOFPROBABILITY

2.6FoundationsofProbability:TheoremsandApplications

2.6.1TheFundamentalPrinciples:DefiningProbabilityandBasicTheorems

The Central Limit Theorem states that sampling distribution of mean tends to benormal, no matter what initial sample distribution looks like, as samplesize gets sufficiently large. This theorem underlies many themes of statistical procedures hypothesis testing, estimation of confidence intervals, etc.

DefiningProbability:

- Probability is represented as a numerical value ranging from 0 to 1, inclusive. A probability of 0 signifies that an event is impossible, whereas probability of 1 denotes that an event is certain.
- The probability of an occurrence A, represented as P(A), is mathematically defined inside sample space (S) that encompasses all possible outcomes.:
  - P(A) = Number of good results in A divided by total number of outcomes in S)

- It is important to understand that sample space must contain all possible outcomes.

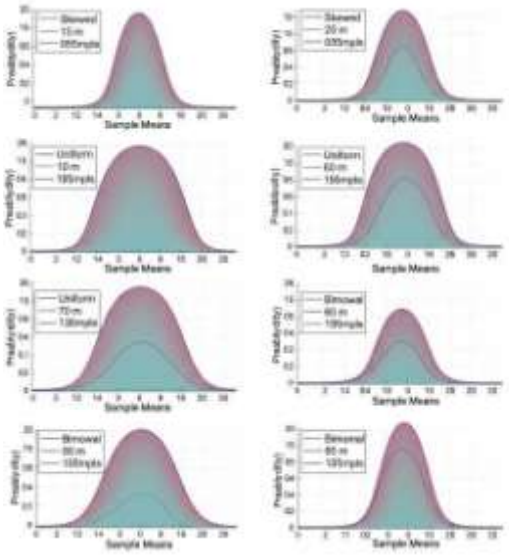

Figure3.1CentralLimitTheorem(CLT).

NumericalExample:

- Examineanequitablesix-faceddice.ThesamplespaceisS={1,2,3,4,5, 6}}.
- Theevent ofrollingan even numberisA={2, 4,6}.
- Therefore,P(A)=3/6= 1/2.

BasicTheorems:

- Theorem1:Probabilityofthe Impossible:
- §Ifsomethingcannothappen,theprobabilityis0.
- §P(Ø) =0,withØbeingtheempty set.
- Theorem2TheProbabilityofaParticular Event:
- §Ifanevent iscertaintohappen thenitsprobabilityisone.
- §P(S) =1 (Here,S issamplespace).
- Theorem3:The complementrule:
- §TheprobabilityofaneventNOToccurringis1 minustheprobabilitythat the event does occur.
- §P(A')=1-P(A)whereA'isthecomplementofeventA.

Numerical Example:

- §Forthedie above;probabilityofnotobtaininganevennumber(A')is:
- §P(A')=1−P(A)=1−1/2=1/2.
- Theorem4ProbabilityRange:
- §ForanyeventA,0≤P(A)≤1.Thisimpliesthatriskprobabilitieswillbe setinbetweenthisrange.

2. 6.2 TheAdditionTheorem:CombiningProbabilities

The addition theorem is essential for determining probability of occurrence of either event.

MutuallyExclusiveEvents:

- Twooccurrencesaremutuallyexclusiveiftheycannothappenatsame time.
- IfA&Baremutuallyexclusive,thenP(A∩B)=0,where∩denotesthe intersectionof events..

AdditionTheoremforMutuallyExclusiveEvents:

- P(A∪B) =P(A) +P(B),where∪denotesunionof events.

NumericalExample:

- Contemplateselectingonecard froma regular52-card deck.
- LetAbethe eventof drawing heart,and Bbeeventof drawing spade.

- Theseeventsaremutually exclusive.
- P(A)=13/52= 1/4, andP(B) =13/52 = 1/4.
- Theprobability of drawingaheart oraspadeis:
- P(A∪B) =1/4 + 1/4 =1/2.

Non-MutuallyExclusiveEvents:

- Twoeventsarenon-mutuallyexclusiveif theycanoccur simultaneously.

AdditionTheoremforNon-MutuallyExclusiveEvents:

- P(A∪B)=P(A) +P(B)-P(A∩B)

NumericalExample:

- Considerdrawingasinglecardfrom astandard52-card deck.
- LetAbeevent ofdrawing aking, andB be theeventof drawingaheart.
- These events are not mutually exclusive because you can draw the king of hearts.
- P(A)=4/52 =1/13, P(B)=13/52 = 1/4,and P(A∩B) = 1/52.
- Theprobability ofdrawing king oraheart is

- P(A∪B) =1/13 + 1/4 -1/52 = (4 +13 -1)/52 = 16/52 = 4/13.

2.6.3 TheMultiplicationTheorem:IndependentandDependentEvents

The multiplication theorem facilitates the computation of probability of simultaneous occurrence of two or more occurrences. It distinguishes between independent and dependent occurrences.

IndependentEvents: Twooccurrencesareindependentifoccurrenceof one event does not influence occurrence of other.

MultiplicationTheoremforIndependentEvents:

- P(A∩B) =P(A)* P(B)

NumericalExample:

- Considerflippingafaircointwice
- Let A denote event of obtaining heads on initial flip, & B denote event of obtaining heads on the subsequent flip.
- Theseoccurrencesareautonomous.
- P(A)=1/2&P(B) =1/2.
- Thelikelihoodofobtainingheadsonbothflipsis:
- P(A∩B) =(1/2)×(1/2)=1/4.

DependentEventsandConditionalProbability:

Two events aredependent if the occurrenceofoneaffects causethe occurrence of the other.

ConditionalProbability:

- The probability of event A happening is expressed as P(A)§ The probability of event B happening, if A has already occurred is known as P(B|A).
- P(B|A)=P(A∩B)/P(A)ifP(A)>0

- **MultiplicationTheoremforDependentEvents:**
  - P(A∩B) =P(A)* P(B|A)
  - **NumericalExample:**

  - Considerselectingtwocardsfromnormal52-carddeckwithout replacement.
  - LetArepresenteventofdrawingakingoninitialdraw,andBdenote event of drawing a king on subsequent draw.

  - Thesefouractionsare interrelated.
  - P(A)=4/52= 1/13.
  - If a King is drawn on the first draw, there are 3 more Kings remaining inthe other 51 cards.
  - P(B|A)=3/51=1/17.
  - Thechanceatpicking outtwokingsis:
  - P(A∩B) =(1/13)*(1/17) = 1/221.

3. AdvancedTheoremsandApplications

Beyond the fundamental principles, Probability theory encompasses sophisticated theorems that are crucial for addressing intricate situations and practical applications.

Bayes'Theorem:

- Bayes' Theorem delineates likelihood of an event, contingent upon priorknowledge of conditions potentially associated with the event.
- Itisgivenby:P(A|B)=[P(B|A)*P(A)]/P(B)

Where:

- P(A|B)isposteriorprobabilityofeventAoccurring,contingentupontruth of event B.
- P(B|A)representstheprobabilityofeventBoccurringcontingentuponthe truth of event A.
- P(A)denotespriorprobabilityofeventA

- P(B)denotespriorprobabilityofeventB.

NumericalExample:

- A medical test has a 95% accuracy rate. 1% of population has the disease. If person tests positive, what is probability they have disease?
- LetD=havingdisease,& +=testing positive.
- P(D)=0.01,P(+|D)=0.95,P(+|D')=0.05.
- P(+) = P(+|D) * P(D) + P(+|D') * P(D') = 0.95 * 0.01 + 0.05 * 0.99 =0.059.
- P(D|+) = (0.95 * 0.01) / 0.059 = 0.161 (approximately). Therefore, even though test is 95% accurate, because occurrence of the disease is so rare, there is only a 16.1% chance the person has the disease if they testpositive.

2.6.4 LawofTotalProbability:

- This theorem provides a way to calculate the probability of an eventthat can happen in more than one way.
- IfA1,A2,…,Anismutuallyexclusive&exhaustive&Bisanevent, then:
- P(B)=P(B|A1)P(A1)+P(B|A2)P(A2)+...+P(B|An)P(An)

2.6.5 Applications:

Thesetheoremsarevital innumerous fields:

- **Statistics:**Hypothesistesting,confidence intervals.
- **Finance:**Riskassessment,portfoliomanagement.
- **Medicine:**Diagnostictesting,epidemiological studies.
- **Computerscience:**Machinelearning,artificialintelligence.

By mastering these fundamental and advanced theorems, one gains the ability to navigate the complex world of probability and apply its principles effectively to solve a wide range of real-world problems.

UNIT2.7 CONCEPTOFSAMPLING

2.7CONCEPTOFSAMPLING

2.7.1 UnveilingtheNeedforSampling:FromVastPopulationstoManageableInsights

Some populations (like the entire country of China, for example) are simply too large or too complex to beable to study head to toe allowing researchers and analysts to cherry pick a smaller, manageable sample to draw conclusions. Imagine trying to parse the sentiment of every citizen in a country, catalog the quality of every good coming off production line ormodel growth of every tree in giant forest. Such efforts would be far tootime consuming and costly not to mention, logistically impossible. This is where the concept of sampling comes into play.) Sampling is the technique of assessing a part or sample of a bigger population torepresent the features of the whole population.

So rather than trying to take on the entire population, we are dealing with a few, more tractable entities,to extrapolate from them to the larger whole. The reasoning goes that as long as a sample is representative of population, we can get useful information without needingto look at every single case. Not onlyis sampling practical, it is also efficient. Focusing our attention on a single sample allows us toconserve a great deal of resources: time, money, people. Mindthat this timeliness is critical in disciplines like market research, where time-to-insight is crucial for business decisions. So, for instance, a company launching a new product might create atest event featuring a select audience of target customers to gauge interest in the product before committing to a full production run. Similarly, in the medical domain the clinical trials mostoften refers to a sequence of testing new pharmaceutical or treatment on a subset of patients in order to validate efficacy and safety before large scale deployment in patient population. Generalizability, the ability to apply knowledge derived from a sample to all of (or some relevant portion of) a population, is the cornerstone of scientific discovery and the evidence-based policymaking that drives much of the contemporary world.

The effectiveness ofsampling, however, depends upon how representative the sample is. Assuming sample is representative of population findings will be valid, but ifit turns out to bea biased sample, the resultingconclusions will be incorrect. Sampling aims to eliminate biasby making sure that sample reflects diversity and community. Characteristics This means being intentional about how the sample is drawn, how many people to sample, and what potential sources oferror exist. But numerous sampling methodshave been developed, each with distinct advantages and disadvantages. The selection process canalso be different based onthe requirements of research, characteristics of the population studied, and available resources at play. Thus, a proper sampling strategy is vital in order to verify the research results

NumericalExample:

For example, a producer produces 100K lampsa day. They what to estimate the percentage ofdefective bulbs. There are100,000 of them, so testing all of them isn't feasible. Instead they go with a sample. They choose a random sample of1,000 bulbs. They are tested, and 20 of them arefound to be faulty. What does this mean at this level: This means that the sample defectrate was 2% (20/1000) From this sample data, they can extrapolate that 2 percent of the overall batch of 100,000 bulbs is probablydefective and that 2,000 bulbs are likely faulty. This conclusion is not the best,but rather a good approximation based on the sample.

2.7.2NavigatingtheSamplingLandscape:TypesofSamplingTechniques

Selecting a suitable sampling method is one of the factors that is critical in the research process since it affects the sample's representativeness and theresearch results' generalizability. Broadly, the two sampling techniques can be defined as Probability sampling: The method of sample selection gives each member of population known, non-zero chance of being chosen. This allowsforsamplerepresentationandenablesthepopulation'sstatisticalconclusions.

What you have is random sampling, where it is done randomly, thisrepresents something roughly along the lines of "with probability," so no bias should be around here. However, in non-probability sampling there is nopoint or indicator, and some bias is introduced into thesample.

ProbabilitySamplingTechniques:

- **SimpleRandomSampling:** This is the simplest form of probability sampling, wherein each individual in population has the same chance ofbeing chosen. It's kind of like drawing names from a hat. While thetechniqueisstraightforward,itisdifficulttoapplyat scale,particularlywhere populationsare geographically separated.
- **SystematicSampling**: It refers to selecting every nth member of population (here n is fixed sampling interval). For example, in case of a population size of 1,000 and sample you want to get of 100, your sampling interval will be: 1,000/ 100 = 10, every 10th member will be selected. While this is very efficient, it can introduce bias if there is some hiddenpattern in population.
- **StratifiedSampling:** Thistechniquesegmentsapopulationintostrataor subgroups according to specific characteristics (such as age, gender, or income).A basic random sample is subsequently extracted from each stratum in a manner that ensures the proportions of these traits in the sample mirror those seen in the population.This is especially beneficial when engaging with varied communities.
- **ClusterSampling:** Instratifiedsampling,thepopulationissegmentedinto clusters, such as geographical regions or educational institutions, from which random clusters are then chosen.All units inside the designated clusters are incorporated in the sample.
- **Multi-stageSampling**: This technique combines multiple sampling methods (eg, stratified, cluster), to createa sample that is both more efficient and representative. For instance, a researcher may want to first stratify the populationbyregionofthecountry,andthenrandomlyselectclustersfrom

- withineachregion,andthen take asimplerandomsamplefromclusters samples.

Non-ProbabilitySamplingTechniques:

- **ConvenienceSampling:** Where samples are selected within the reach of the researcher, and are easy to access. An example might be a researcher interviewing people walking by on a street corner. Cheap and easy to implement;however,methodhasbiasissues **Judgmentsampling**:Aprocess of collecting samples in an image while the researcher pulls from theirexpertise or skill of the material. In one, a marketing manager selects a sample of customers whom shebelieves accurately represents her target market. This is helpful when certain knowledge is required, but this leads to bias if the researcher's judgement was wrong (quantitative).
- **QuotaSampling**: In this method of sampling, a sample is selected according to a specific quota for certain types of characteristics such as sex or age group, education level, etc. That could be, for instance, a researcher who wants to interview an equal number of men and women. This is similar to stratified sampling, except that, you do not have to do therandom selection here.
- **SnowballSampling**: This sampling technique is applied in cases of some hard-to-access populations like drug users, or homeless individuals. Itis simplyidentifyingsmallgroupofpeopleinpopulationandaskingthemtorefer more. This method is useful for obtaining samples from hidden populations, however,could introduce bias in the outcome if the first group of individuals was not truly representative of population.

NumericalExample:

A university wants to understand how students feel about the services on campus. Sothey will perform stratified sampling. There are four strata in the student population: freshman, sophomore, junior, and senior. The university ensures that the sample is proportionally representative of each class. Alternatively, if the university's population consists of 25% each of the classes, Freshman,Sophomore,Junior,Senior,thenasampleof400wouldyield100

Freshman, 100 Sophomores, andso on. Doing so will ensure classes are not being misrepresented.

2.7.3 SizingUptheSample:DeterminingtheRightSampleSize

The size of the sample it generates in a sampling process is one of themajor components of sampling. If sample is small enough, it may misrepresent population, resulting in erroneous results. Or too large a sample size an unnecessary drain of time &money.

FactorsAffectingSampleSize:

- **PopulationSize:** Larger populations require larger samples to be representative. But it's not a straightline between the two. Once a population reaches a certain size, increasing the sample sizeprovides diminishing returns.
- **Precision**: The margin of error expresses precision, the range within which responses from the sample are presumed to reflect values in the population. Smaller margin of error requires alarger sample size.
- **VariabilityoftheCharacteristicsBeingInvestigated**: Largersample sizes are needed to detect substantial variation in the characteristics underscrutiny. Inanopinionneutralaboutanytopicanextremelylargesample size isneeded in order to identify difference.
- **Confidencelevel**: This is the degree of certainty that the sample outcome falls within the margin of error. A more confident levelneeds bigger sample size. Most commonconfidence levels are 95% and 99%.
- **SampleSizeFormulas**:

Depending on type of data being collected & desired level of precision, several different formulasmay be used to determine an appropriate sample size. The formula for sample size related to proportionis:

Theformulayou provided is:

$$n = \frac{Z^2 \cdot p \cdot (1-p)}{E^2}$$

Where:

- nissamplesize
- ZisZ-scorecorrespondingto desiredconfidence level
- pistheestimatedpopulation proportion
- Eisdesiredmarginoferror

To estimate number of voters supporting a specific candidate with 95% confidence level &a 3% margin of error, assuming a population proportion of 50%, the required sample size is:

n=(1.96^2 * 0.5 * 0.5) /0.03^2 =1067.11

Therefore,the researcher wouldneedasamplesizeofapproximately1,0

Sampling Distribution

A sampling distribution is the probability distribution of a sample statistic (such as mean, proportion, or variance) obtained from repeated random samples of the same size from a population.

In other words:

If we keep drawing samples from a population, compute a statistic (like the mean) for each sample, and then plot those values, the pattern we get is called a sampling distribution.

Key Features

1. Based on Samples, Not Population

It shows how a statistic (e.g., mean) varies from sample to sample.

2. Mean of Sampling Distribution

The average of the sample means equals the population mean (μ).

3. Standard Error (SE)

The standard deviation of the sampling distribution is called the standard error.

For sample mean:

$SE = \sigma / \sqrt{n}$

where σ = population standard deviation, n = sample size.

4. Effect of Sample Size

Larger sample size → smaller SE → more reliable estimate.

5. Central Limit Theorem (CLT)

For large n (usually $n \geq 30$), the sampling distribution of the mean tends to be normal, regardless of the population's distribution.

Example

Population: Average exam score of all students = 70, σ = 10.

Take repeated samples of size n = 25.

Compute mean score for each sample.

The distribution of these means forms the sampling distribution of the mean.

Its mean will still be 70, but its spread (SE) = $10 / \sqrt{25} = 2$.

SELFASSENMENTQUESTION

Multiple-ChoiceQuestions(MCQs)

1. Whatistheprobabilityofanimpossibleevent?

a. 1

b. 0.5

c. 0

d. 100%

2. Whichofthefollowingisatypeofprobabilitybasedonhistoricaldata?

a. Theoreticalprobability

b. Experimentalprobability

c. Subjectiveprobability

d. Axiomatic probability

3. Itisacharacteristicoftheadditivelawofprobabilitythatiftwoeventsaremutuallyexclusive,thentheprobabilityofeitherofthemoccurringisthesumoftheirprobabilities.Whatformulasignifiesthislaw?

a. P(A∩B) =P(A) +P(B)

b. P(A∪B)=P(A) +P(B)-P(A∩B)

c. P(A∪B) =P(A) +P(B)

d. P(A|B)=P(A)/ P(B)

4. Whatprobabilitydistributionisusedifanexperimentresultsinexactlytwopotentialoutcomes(successandfailure)?

a. Poissondistribution

b. Binomial distribution

c. Normal distribution

d. Exponentialdistribution

5. Whatpercentageofdataarewithinonestandarddeviationinanormaldistribution?

a. Fifty percent

b. Sixty-eight percent

c. Seventy-fivepercent

d. Ninety-fivepercent

6. Whatis1featureofthePoissondistribution?

a. Itisusedwhenthedatais continuous.

b. Itusedforoccasional eventsinaperiodoftime.

c. Itispossibleonlyincaseofnormaldistribution.

d. Itfollows abinomialdistribution.

7. GiventhatP(A)=0.6andP(B)=0.3,andthatoccurrencesAandB areindependent,whatisP(A∩B)?

a. 0.9
b. 0.18
c. 0.3
d. 0.6

8. Whichofthefollowingbestdefinesthedecisionruleinprobability?

a. Arulethathelpstochoosebetweentwo probabilities
b. Aruletodeterminewhethertorejector acceptanull hypothesis
c. Amethodtocalculate expectedvalues
d. Aformulaforbinomialprobability

9. Thesumofprobabilitiesofallpossibleoutcomesinasamplespacemustbe:

a. 1
b. 0
c. Between0and1
d. Greaterthan1

10. Whatistheprimaryassumptionofthecomposerofthebinomial?

a. Asmanyattemptsasyou like

b. Variablechanceofsuccess

c. Fixednumberoftrialsandindependentevents.

d. Probability Distribution

11. Thetheoremwhichdescribestheprobabilityofsomeotherincidentoccurringwhenanothereventhasalreadyoccurredisrepresentedby$P(A|B) = P(A \cap B) / P(B)$?

a. TotalLawof Probability

b. Bayes'sTheorem

c. Probabilitygiventhatsomethinghappens(botherhead2.3-7)

d. MultiplicationRule

12. Whatissignificanceofsamplinginprobability?

a. Itcomplicatesthestudy.

b. Italso assistsininvestigatinglargenumbersofpopulationsbysmall ones.

c. Itgivesthe resultswith100%accuracy.

d. Thatisitremovesall doubt.

13. Whichofthefollowingdistributionsiscontinuous?

a. Thebinomial distribution

b. Possion distribution

c. Gaussiandistribution

d. Hypergeometricdistribution

14. WhatisanapplicationofthePoissondistributioninreallife?

a. Passstudentinan examination

b. Numberofcallsreceivedin acallcentereveryhour

c. Heightsofthestudents inaclassaregivenby:

d. Monthlysalesof a product.

15. Inprobability,oneeventthathasnoimpactonanothereventis:

a. Dependent Event

b. Dependent eventifnotindependentevent

c. Conditional event.

d. Noneof theabove

ShortQuestions

1. Defineprobabilityanditssignificance
2. Explaintheadditiveandmultiplicativelawsofprobability.
3. Whatisthedecisionrule in probability?
4. Definebinomialdistributionandits properties.
5. Whatarethecharacteristicsofanormaldistribution?
6. ExplainthePoissondistributionandits applications.
7. Whatarethe basictheoremsof probability?
8. Whatistherole samplingin probability?
9. Whatroledoprobability distributionsplay indataanalysis?

LongQuestions

1. Describethevarioustypesofprobabilitywithexamples.
2. Explaintheadditionandmultiplicationlawsofprobabilitywithexamples.
3. Describethepropertiesofbinomial,Poisson,andnormal distributions.
4. Statethereal-lifeapplicabilityofthetheoremsofprobability.
5. Howisprobabilitybeingusedindecisionmakingin business?
6. Discusstheideaofsampling anditsdirnlicationto statistics.
7. Explainthedecisionruleinprobabilityanditssignificance.
8. Compareandcontrastbinomialandnormaldistributions.

Glossary for **Probability** and **Probability Distribution**:

Term	**Definition**
Probability	A measure of the likelihood that a particular event will occur, expressed as a number between 0 (impossibility) and 1 (certainty).
Random Experiment	A process or action that produces uncertain outcomes, such as rolling a die or drawing a card.
Sample Space	The set of all possible outcomes of a random experiment.
Event	A subset of the sample space; one or more outcomes for which we want to find the probability.
Independent Events	Two or more events where the occurrence of one does not affect the probability of the others.
Dependent Events	Events where the occurrence of one affects the probability of the others.
Conditional Probability	The probability of an event occurring given that another event has already occurred.
Probability Distribution	A mathematical function that describes how probabilities are distributed over the values of a random variable.
Discrete Probability Distribution	A distribution for a random variable that takes on countable values, with each possible value having an associated probability (e.g., Binomial, Poisson).
Continuous Probability Distribution	A distribution for a random variable that takes on infinitely many possible values, usually intervals of real numbers (e.g., Normal, Exponential distributions).
Random Variable	A numerical outcome of a random experiment, which can be discrete or continuous.
Expected Value (Mean)	The long-run average or weighted average of all possible values of a random variable, weighted by their probabilities.

Term	Definition
Variance	Ameasureofthespreadordispersionoftherandomvariablevalues around the expected value.
Standard Deviation	Thesquarerootofthevariance,representingtheaveragedistancefrom the mean.
Cumulative DistributionFunction (CDF)	Afunctionthatgivestheprobabilitythatarandomvariableislessthan or equal to a certain value.
ProbabilityMass Function (PMF)	Fordiscretevariables,thefunctionthatgivestheprobabilitythata random variable is exactly equal to a specific value.
ProbabilityDensity Function (PDF)	Forcontinuousvariables,thefunctionthatdescribestherelative likelihood of the variable taking on a particular value.

Summary:ProbabilityandProbabilityDistribution

1. Probability:

Probabilityisameasureofthelikelihoodorchancethataparticulareventwilloccur.Itranges from 0 (impossible) to 1 (certain).

ClassicalProbability:Basedonlogicalreasoningwhereoutcomesareequallylikely.

Example: Probability of getting a head in a coin toss = 1/2.

Empirical Probability: Based on observed data or experiments.

Example:Probabilityofastudentpassingbasedonpastresults.

SubjectiveProbability:Basedonpersonaljudgmentorexperienceratherthanexactcalculations.

Example: A doctor's estimate of recovery chances.

2. KeyTermsin Probability:

Experiment:A process that results in outcomes.

SampleSpace(S):Thesetofallpossibleoutcomes. Event:

A subset of the sample space.

MutuallyExclusiveEvents:Eventsthatcannothappentogether.

Independent Events: One event does not affect the other.

3. Probability Distribution:

Aprobabilitydistributiondescribeshowprobabilitiesaredistributedoverthevaluesofarandom variable.

RandomVariable:Avariablethattakesonvaluesbasedonoutcomesofarandomevent.Itcan be:

Discrete(countablevalues,likenumberof heads)

Continuous(infinitevalueswithinaninterval,likeheightorweight)

TypesofProbabilityDistributions:

1. DiscreteProbabilityDistribution:

Dealswithdiscrete randomvariables.

Example:BinomialDistribution,Poisson Distribution.

2. ContinuousProbabilityDistribution:

Dealswithcontinuousrandom variables.

Example:NormalDistribution,ExponentialDistribution.

Multiple-choicequestionsalongwiththeircorrectanswers:

1. What istheprobabilityofanimpossibleevent?

Answer c. 0

2. Whichofthefollowingisatypeofprobabilitybasedonhistoricaldata?

Answer b. Experimental probability

3. Theadditivelawofprobabilitystatesthatthelikelihoodoftwomutuallyexclusive events is the sum of their respective probabilities. What formula signifies this law?

Answerc.P(A∪B)= P(A)+ P(B)

4. Whichprobabilitydistributionisutilizedwhenanexperimentyieldsjusttwo possible outcomes (success or failure)?

Answer b. Binomial distribution

5. Inanormaldistribution,whatproportionofdatalieswithinonestandard deviation ofthe mean?

Answer b. Sixty-eight percent

6. What is a characteristic of a Poisson distribution?

Answer b. It is utilized for infrequent events inside a set interval.

7. Given that P(A) = 0.6 and P(B) = 0.3, and that occurrences A and B are independent,whatisP(A∩B)?

Answer b. 0.18

8. Which of the following best defines the decision rule in probability?

Answer b. A rule to determine whether to reject or accept a null hypothesis

9. Thesumofprobabilitiesofallpossibleoutcomesinasamplespacemustbe:

Answer a. 1

10. What is the key assumption of binomial distribution?

Answer c. Fixed number of trials with independent events

11. The theorem that articulates the likelihood of one event occurring given the occurrence of anothereventisexpressedasP(A|B)=P(A∩B)/P(B)?

Answer c. Conditional Probability

12. Why is sampling important in probability?

Answer b. It helps analyze large populations using smaller groups

13. Which of the following probability distributions is continuous?

c. Normal distribution

14. Which of the following is a real-life application of Poisson distribution?

b. Number of phone calls received at a call center per hour

15. Inprobability,aneventthatdoesnotaffecttheoutcomeofanothereventiscalled:

Answer b. Independent event

MODULE3CORRELATIONANDREGRESSIONA NALYSIS

Structure

UNIT3.1 Introduction to Correlation

UNIT3.2 PositiveandNegativeCorrelation

UNIT3.3 Karl Pearson'sCoefficientof Correlation

UNIT3.4 Spearman'sRankCorrelation

UNIT3.5 Introduction to Regression Analysis

UNIT3.6 LeastSquareFitofaLinearRegression

UNIT3.7 Two Lines of Regression

UNIT3.8 PropertiesofRegression Coefficients

OBJECTIVES

- Describethemeaningandimportanceof correlationinstatisticalanalysis.
- Determine&explainthedirection&strengthofrelationshipsamongvariables.
- CalculateandinterpretlinearcorrelationbythePearson'smethod.
- Compute and interpret the rank correlation coefficient of non-parametric data.
- Use linear regression and R-Square implementation with Least Square Method to fit a line to data and calculate your square of your fit another straight and another data group.
- Interpretations of the regression lines equation for the two variables Understand and interpret the equations of regression lines of two variables.
- Identify and discuss key properties and implications of regression coefficients.

UNIT3.1INTRODUCTIONTOCORRELATION

3.1IntroductionToCorrelation

3.1.1UnveilingtheRelationship:TheEssenceofCorrelation

Correlation is statistical concept that quantifies degree of association between two variables. It allows us to determine whether alterations in one variable are associated with modifications in another. The association does not imply causation, but shows correlation & dependencies that can be of great value in other areas.

- **DefiningCorrelation:**
  - Correlationanalysisinvestigatesthedegreeanddirectionofalinearrelationship between two quantitative variables.
  - Weuseittomakedecisions,suchas:"AsAincreases,doesBgoup, down, or stay the same."
- **TheSignificanceofCorrelation:**
  - Correlationisabedrockofdataanalysis,research,&decision making.
  - Inscience,itcanhelptoestablishpossiblelinksbetween observations.
  - It is useful in many business for understanding customers preferences and market orientations.
  - Infinance,itmeasuresthecorrelationbetweenassetprices.
- **Correlationvs.Causation:**
  - Itisessentialtonotethatcorrelationdoesnotimplycausality.The correlation betweentwo variables does not imply causation.
  - Theremightbeathird,unobservedvariableinfluencingboth,ortherelationship could be coincidental.
  - An investigation may reveal a correlation between ice cream sales &crime rates. Nonetheless, it seems more probable that elevated temperatures augment both ice cream sales & crime rates.

3.1.2 MeasuringtheStrengthandDirection:CorrelationCoefficients

Correlation coefficients yield a numerical value indicating degree & direction of linear association between two variables. Pearson's r is most often utilized coefficient.

Pearson'sCorrelationCoefficient(r):

- Pearson'srquantifieslinearcorrelationbetweentwovariables.
- Itrangesfrom-1to+1:

- +1signifiesan impeccablepositive association.
- -1signifiesan idealnegativecorrelation.
- 0indicatesnolinear correlation.

- **UnderstandingtheValues:**

- Valuesapproaching+1or-1signifyarobustassociation.
- Valuesapproaching0signifyaweakornonexistent association.

- Examplevalues.
- r =0.9: Strong positivecorrelation.
- r=-0.7: Strongnegativecorrelation.
- r=0.1: Weak positivecorrelation.
- r=-0.2: weak negativecorrelation.
- r =0: no correlation.

- **CalculatingPearson'sr:**
- Pearson'srformulaincorporatesthecovarianceofthetwovariables along with their standard deviations.

Formula:

- $r=[\Sigma(x -\bar{x})(y-\bar{y})]/[\sqrt{(\Sigma(x -\bar{x})^2)} * \sqrt{(\Sigma(y -\bar{y})^2)}]$
- Where:
- xand y arethevariable values.
- $\bar{x}$and $\bar{y}$arethemeansof x and y.
- Σdenotesthesum.

UNIT3.2POSITIVEANDNEGATIVECORRELATION

3.2 PositiveAndNegativeCorrelation

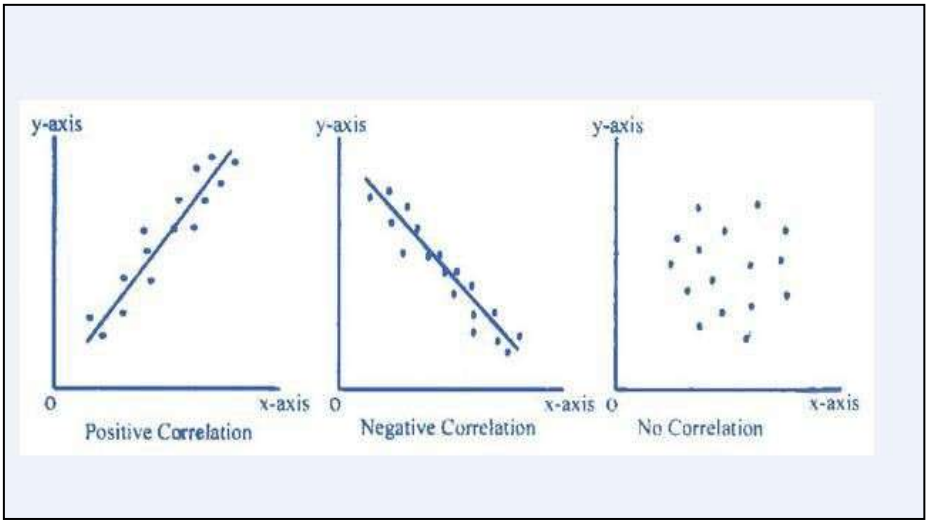

Figure3.2:PositiveandNegativeCorrelation.

1. UnderstandingCorrelation:TheFoundationofRelationships

- Correlation is statistical metric that quantifies degree to which two variables fluctuate in relation to one another. This is a key notion in data analysis that enables the identification of patterns and correlations within datasets.
- It is essential to recognize that correlation does not signify causality. A relationship between two variables does not mean a cause-and-effect-relationship between them. There may be other factors at play that Nino is influencing.
- We will delve into how correlation is calculated, how it is read, and what cannot be told from correlation.
- Correlationsarescored from-1to+1.
- Avalueof+1correspondstoperfectpositivecorrelation.
- A-1valuerepresentsperfectnegative correlation.

- Avalue of0 indicatesnocorrelation.

VisualizingCorrelation:ScatterPlots:

Scatter plots are essential tools for depicting the relationship between two variables.Each point on the graph represents a pair of values, with onevariable shown on the x-axis and the other on the y-axis.

- Byexaminingtheconfigurationofthepoints,wemayascertainthe intensity and direction of the link.
- Atrendofpointsascendingfromlefttorightsignifiesafavorable association.
- Adecreasingtrendofpointsfromlefttorightsignifiesanegative association.
- Randomlyspreadpointsindicate minimalorno association.

TheCorrelationCoefficient:

The correlation coefficient, represented as "r," measures the degree and direction of the linear relationship between two variables.

Pearson's correlation coefficient is the primary type of correlationcoefficient, evaluating the linear relationship between two continuous variables.Comprehending the magnitude of association.

- Valuesapproaching+1 or-1signifyarobustassociation.
- Valuesapproaching0signifyaweakornonexistentassociation.

For instance:

- r=0.9:Indicatingarobustpositive association
- r=-0.7:Indicating astrongnegativeconnection
- r=0.1:indicatesa weak positiveconnection.

NumericalexampleofcalculatingCorrelation:

- Toshowabasicexample,wewilluseasmall dataset.
- Letssay wehavethefollowing data ofstudy hoursand exam scores.
- StudyHours(x): 1,2, 3,4, 5.
- ExamScores(y):50, 60,65,80, 90.
- We can then calculate Pearson correlation coefficient. This involves finding mean of x &y, standard deviation of x &y, & covariance of x &y.

- After the calculations, we would find a high positive correlation. This means that as study hours increase, exam scores also increase.
- Explaining the formula of Pearsons correlation is very technical, therefore it is more important to explain the meaning of the resulting number.

3.2.1 PositiveCorrelation:WhenVariablesMoveTogether

- **DefinitionandCharacteristics:**

  - A positive correlation transpires when two variables simultaneously grow or decrease. In other words, an increase in one variable correlates with an increase in other variable, whereas a reduction in one variable correlates with decrease in other variable.

  - Thisrelationshipisrepresentedbyapositivecorrelationcoefficient.
  - Examplesofpositivecorrelationareabundantinvarious fields.
- **Real-WorldExamples:**
  - **HeightandWeight:**Generally,tallerpeopletendtoweighmore, demonstrating a positive correlation.
  - **StudyTimeandExamScores:**Asstudydurationgrows,examination scores often enhance.
  - **AdvertisingSpendingandSales:**Increasedadvertisingspendingoften leads to increased sales.
  - **TemperatureandIceCreamSales:** Asthetemperaturerises,thesalesof ice cream tend to increase.
  - **ExerciseandCalorieExpenditure:**Themoresomeoneexercisesthe more calories they will burn.

- **NumericalExample:**

  Letusexaminethecorrelationbetweenweeklyexercisedurationand caloric expenditure.

  Data:

  - Hoursof Exercise(x): 1, 2,3, 4, 5
  - CaloriesBurned(y):200,400,600,800, 1000
  - In this example, as number of hours spent exercising increases, number of calories burned also increases proportionally. This is a clear illustration of positive correlation.
  - If we were to plot this data on a scatter plot, the points would form an upward sloping line.
  - IfwecalculatedthePearsonsCorrelationcoefficient,theresultwouldbea number very close to 1.

UNIT3.3KARLPEARSON'SCOEFFICIENTOFC ORRELATION

3.3 KarlPearson'sCoefficientOfCorrelation

Karl Pearson's correlation coefficient 'r' is a statistic that quantifies linear correlation between two continuous variables. It quantitatively assesses extent to which a linear equation can represent the relationship between those variables. The coefficient resides within the interval of -1 to +1, where:

- +1 signifies perfect positive linear correlation, indicating that when one variable risesby 2,other alsoincreasesproportionallyby2,with allpoints aligning precisely on a straight line with positive slope.
- -Acorrelation of -1 indicates perfect negative linear relationship, wherein an increase in one variable corresponds to a drop in other, with all data points aligning precisely along a straight line with negative slope.
- 0 means no linear correlation, so no straight-line relationship between variables.Thisdoesn'tnecessarilymeanthereisnorelationship,itmaybe non-linear relationship**.**
- A value between -1 & +1 signifies varying degrees of linear correlation. The value between +1 and-1 quantifies linear relationship strength. The closer the value is to0, weaker linear relationship is.

This is determined by ratio of covariance of two variables to the product of their standard deviations. Covariance measures the degree to which two random variables co-vary, whereas standard deviation quantifies extent to which values of each variable diverge from the mean. Karl Pearsons Coefficient of CorrelationFormula:

$$r=Cov(X,Y)/(\sigma X*\sigma Y)$$

Where:

- risPearsoncorrelationcoefficient.
- Cov(X,Y)iscovariancebetweenvariablesX &Y.

- σXisstandarddeviationofvariable X.
- σYisstandarddeviationofvariable Y.

Alternatively, using raw scores, the formula can be expressed as:

$$r=[n(\sum XY)-(\sum X)(\sum Y)]/\sqrt{\{[n(\sum X^2)-(\sum X)^2][n(\sum Y^2)-(\sum Y)^2]\}}$$

Where:

- nisnumberofdata pairs.
- ∑XYissumofproductsofpairedscores.
- ∑Xis sumof Xscores.
- ∑Yisthesumof Y scores.
- $\sum X^2$issumofsquaredXscores.
- $\sum Y^2$issumofsquaredY scores.

NumericalExample:

Now let us consider a numerical example, calculating Karl Pearson's correlation coefficient. Let us say we have the following dataset for the Study hours (X) &Test scores (Y) of 6 students:

Student	**StudyHours (X)**	**TestScores (Y)**
1	2	50
2	3	60.0
3	4	65
4	5	75
5	6	80
6	7	90

Tocalculate'r', weneed tocomputeFollowing:

1. **Calculate∑X,∑Y,∑XY,$\sum X^2$,and $\sum Y^2$:**
   - ∑X=2 +3+4 +5+6 +7 =27

- $\sum Y = 50 + 60 + 65 + 75 + 80 + 90 = 410$
- $\sum XY = (2*50) + (3*60) + (4*65) + (5*75) + (6*80) + (7*90) = 1940$
- $\sum X^2 = 2^2 + 3^2 + 4^2 + 5^2 + 6^2 + 7^2 = 159$
- $\sum Y^2 = 50^2 + 60^2 + 65^2 + 75^2 + 80^2 + 90^2 = 28850$

2. **Plugthevaluesintotheformula:**

$r = [6(1940) - (27)(410)] / \sqrt{\{6(159) - (27)^2 - (410)^2\}}$ r =

[11640 - 11070] / √{954 - 729}

$r = 570 / \sqrt{\{(225)(5000)\}}$ r

= 570 / √1125000

r =570 / 1060.66

r ≈ 0.537

Therefore, the Karl Pearson's coefficient of correlation between study hours and test scores is approximately 0.537. It can be observed that this is apositive linear relationship. Test scores rise as time spent studying rises, but the connection is slightly less than perfectly linear.

InterpretationandSignificance

Correlation coefficientshouldbeinterpretedtakingintoaccount itsmagnitude and sign.

- **Magnitude:**Theabsolutevalueof'r'signifiesintensityoflinearcorrelation.
  - |r|≥ 0.8: Strong correlation
  - 0.5≤ |r|<0.8:Moderatecorrelation
  - 0.2≤ |r|<0.5:Weak correlation
  - |r|<0.2:Veryweakor no correlation
- **Direction:**Thesignof'r' indicatesdirectionoflinearrelationship.
  - Positive'r':Positivelinearcorrelation(variablesincrease together).
  - Negative'r':Negativelinearcorrelation(variablesmoveinopposite directions).

Itisessentialtorecognizethatcorrelationdoesnotimplycausality.Themore they are positively correlated does not mean that if it happens A B it nicrosoftm means that it isAB. There could be other variables affecting both, or this relation might be spurious.

UNIT3.4 SPEARMAN'SRANKCORRELATION

3.4 Spearman'sRankCorrelation

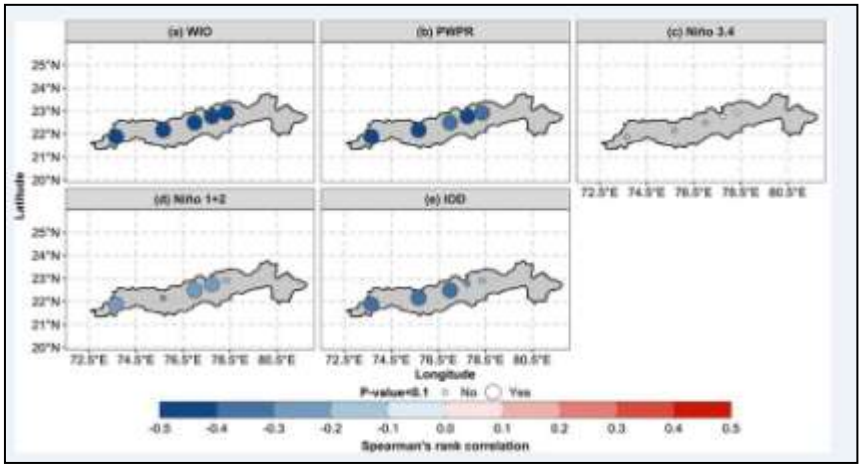

Figure3.3:Spearman'sRankCorrelationCoefficient.

3.4.1 UnderstandingNon-ParametricCorrelation

Spearman's Rank Correlation (ρ) serves as non-parametric alternative to Pearson's correlation coefficient. Pearson's correlation is confined to linear associations among continuous variables, Spearman's correlation analyzes monotonic relationships between ranked data, where outliers and non-normally distributed data will not affect results significantly. Basically, it describes how well the relationship between two variables can be explained through monotonic functions: If one variable goes up, the other one will also go up (or down) but that does not have to be on a constant rate. Hence, Spearman rank correlation is especially valuable when dealing with ordinal data, such as survey Likert-scale responses, or when data is continuous but violates the assumptions of normality that are necessary for a valid Pearson's correlation. To be even more specific, heart of Spearman's correlation is converting the raw data to ranks and then finding a correlation coefficient on these ranks. This method works because it removes the influence of extreme values andconsiders the relative ranks of the data points we have, so we can get a true measure of the association regardless of the skewness in the distributionoroutliers.Sinceyouareconcernedonlywithranksinsteadof

data points, Spearman's correlation focuses on thetrend of how two variables vary with respect to each other, regardless of the exact numerical distances between them. Due to its applicability to diverse datasets, it serves as a potent instrument in disciplines such as social sciences, psychology, and market research, where data seldom adhere to normal distribution. The coefficient, ρ, which varies from -1 to +1, indicates the presence of a statistical relationship between the data, whether positive or negative. +1 signifies a perfect positive monotonic relationship, -1 denotes a perfect negative monotonic relationship, &0 represents the absence of a monotonic relationship. The intensity of the association is shown by the size of the coefficient, while its direction is denoted by the sign.

CalculatingandInterpretingSpearman'sRankCorrelation:AStep-by-StepGuidewithNumericalExamples

collected from five students their scores on that exam. Data were with an example to understand the process. For instance, consider examining theimpact of the number of hours students dedicate to preparing for an impending

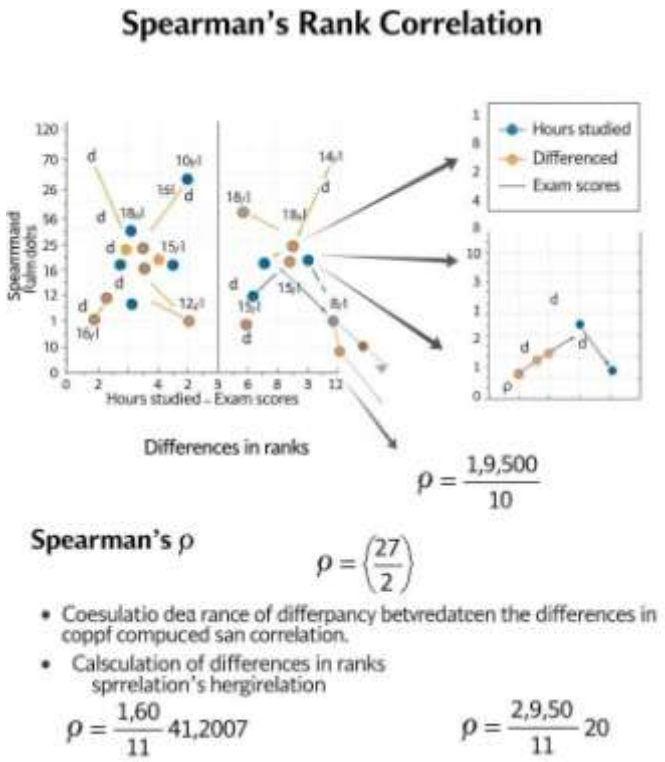

Figure3.4:Spearman'sRankCorrelation.

examination on correlation, which is computed by: Now let us go through thestepsSpearman's Rank:

Student	HoursStudied (X)	ExamScore(Y)
A	10	20
B	15	25
C	8	18
D	20	35
E	12	22

Step1:RanktheData

First, we rank the values of X & Y separately in ascending order. If there areties, we assign the average rank to the tied values.

Student	Hours Studied (X)	Rank ofX (Rx)	Exam Score (Y)	Rank ofY (Ry)
A	10	2	20	2
B	15	4	25	4
C	8	1	18	1
D	20	5	35	5
E	12	3	22	3

Step2:CalculatetheDifferencesinRanks(d)

Next,wecalculatethedifference(d)betweenranksofeachpairof observations (Rx - Ry).

Student	Rx	Ry	d(Rx -Ry)
A	2	2	0
B	4	4	0
C	1	1	0
D	5	5	0
E	3	3	0

Step3:SquaretheDifferences(d^2)

Wethensquarethedifferences(d^2) toeliminatenegativevalues.

Student	d	d^2
A	.00	0
B	.00	0
C	.00	0
D	.00	0
E	.00	0

Step4:SumSquaredDifferences(Σd^2)

Wesumthesquareddifferences(Σd^2).Inourexample,$\Sigma d^2=0+0+0+0+0$ $=0$.

Step5:ApplySpearman'sRankCorrelationFormula

TheformulaforSpearman'sRankCorrelationis:

$\rho=1-(6\Sigma d^2)/(n(n^2-1))$ Where:

- ρisSpearman'sRankCorrelationcoefficient.
- Σd^2issumofsquareddifferencesinranks.
- nisnumberofdata pairs.

Inourexample,n=5,and$\Sigma d^2=0$.Pluggingthesevaluesintoformula: $\rho = 1 -$

$(6 * 0) / (5(5^2 - 1)) \rho = 1 - 0 / (5 * 24) \rho = 1 - 0 \rho = 1$

This result indicates a perfect positive monotonicrelationship between number of hours studied & exam scores.

AMoreComplexExamplewithTies

Let'sconsideranotherexamplewithtiesinthedata:

Student	StudyTime(X)	Exam Performance(Y)
F	12	75
G	15	80
H	10	70
I	15	80
J	18	90

Step1:RanktheDatawithTies

ForX:10,12,15,15,18.Theranksare1,2,3.5,3.5,5(15istied,sowetake theaverageof3 and 4).ForY: 70, 75, 80, 80, 90.Theranks are1, 2, 3.5, 3.5, 5 (80 is tied, so we take the average of 3 and 4).

Student	X	Rx	Y	Ry
F	12	2	75	2
G	15	3.5	80	3.5
H	10	1	70	1
I	15	3.5	80	3.5
J	18	5	90	5

Step2:CalculateDifferences(d)

Student	Rx	Ry	d
F	2	2	0
G	3.5	3.5	0
H	1	1	0
I	3.5	3.5	0
J	5	5	0

Step3:SquaretheDifferences(d^2)

Student	d	d^2
F	0.0	0
G	0.0	0
H	0.0	0
I	0.0	0
J	0	0

Step4:SumtheSquaredDifferences(Σd^2)

$\Sigma d^2 = 0$

Step5:ApplytheFormula

$\rho = 1 - (6 * 0) / (5(5^2-1))$ $\rho = 1$

Again,weget perfect positive correlation.

Let'sconsideradifferentset ofdatathatcreatesaresult thatisnot 1.

Student	StudyTime (X)	ExamPerformance (Y)
K	10	90
L	12	80
M	15	75
N	18	70
O	20	60

Step1:RanktheData

Student	X	Rx	Y	Ry
K	10	1	90	5
L	12	2	80	4
M	15	3	75	3
N	18	4	70	2
O	20	5	60	1

Step2:CalculateDifferences(d)

Student	Rx	Ry	d
K	1	5	-4

Step3:SquaretheDifferences(d^2)

Student	d	d^2
K	-4	16
L	-2	4
M	0	0
N	2	4
O	4	16

Step4:SumtheSquaredDifferences(Σd^2)

$\Sigma d^2 = 16 + 4 + 0 + 4 + 16 = 40$

Step5:ApplytheFormula

$\rho = 1-(6*40) / (5(5^2-1)) \rho = 1-(240)/(5*24) \rho = 1-240/120 \rho = 1-2 \rho$

$= -1$

Inthiscase,wehave aperfectnegative correlation.

Now,let'sconsiderascenariowithlessperfect correlation.

Student	StudyTime(X)	ExamPerformance(Y)
P	10	75
Q	12	80
R	15	70
S	18	85
T	20	65

Step1:RanktheData

Student	X	Rx	Y	Ry
P	10	1	75	3
Q	12	2	80	4
R	15	3	70	2
S	18	4	85	5
T	20	5	65	1

Step2:CalculateDifferences(d)

Student	Rx	Ry	d
P	1	3	-2
Q	2	4	-2
R	3	2	1
S	4	5	-1
T	5	1	4

Step3:SquaretheDifferences(d^2)

Student	d	d^2
P	-2	4
Q	-2	4
R	1	1
S	-1	1
T	4	16

Step4:SumtheSquaredDifferences (Σd^2)

$\Sigma d^2 = 4 + 4 + 1 + 1 + 16 = 26$

Step5:ApplytheFormula

$\rho = 1-(6*26)/(5(5^2-1))$ $\rho = 1-(156)/(5*24)$ $\rho = 1-156/120$ $\rho = 1-1.3$
$\rho = -0.3$

Inthiscase,wehave amoderatenegative correlation.

InterpretingtheResults

- **ρ = +1:** Ideal positive monotonic correlation. As one variableescalates, the other concomitantly escalates consistently.
- **ρ = -1:** Ideal negative monotonic correlation. As one variableescalates, the other invariably diminishes.
- **ρ = 0:** No monotonic correlation. The variables are not related in a consistent increasing or decreasing manner.
- **Valuesbetween-1and+1:** Indicate varying degrees of correlation. Theproximityofvalueto+1or -1indicatesahigherassociation.Acorrelation closer to 0 indicates a weaker relationship.

UNIT3.5 INTRODUCTIONTOREGRESSIONANALYSIS

3.5 IntroductionToRegressionAnalysis

Itcan either be simple or multiple depending upon the number of which they have to relate. The primary aim is to understand the correlation between changes in independent factors and changes in dependent variable. Insummary, regression seeks to establish a line or curve that accuratelyrepresents relationship between variables, enabling use of independent variable

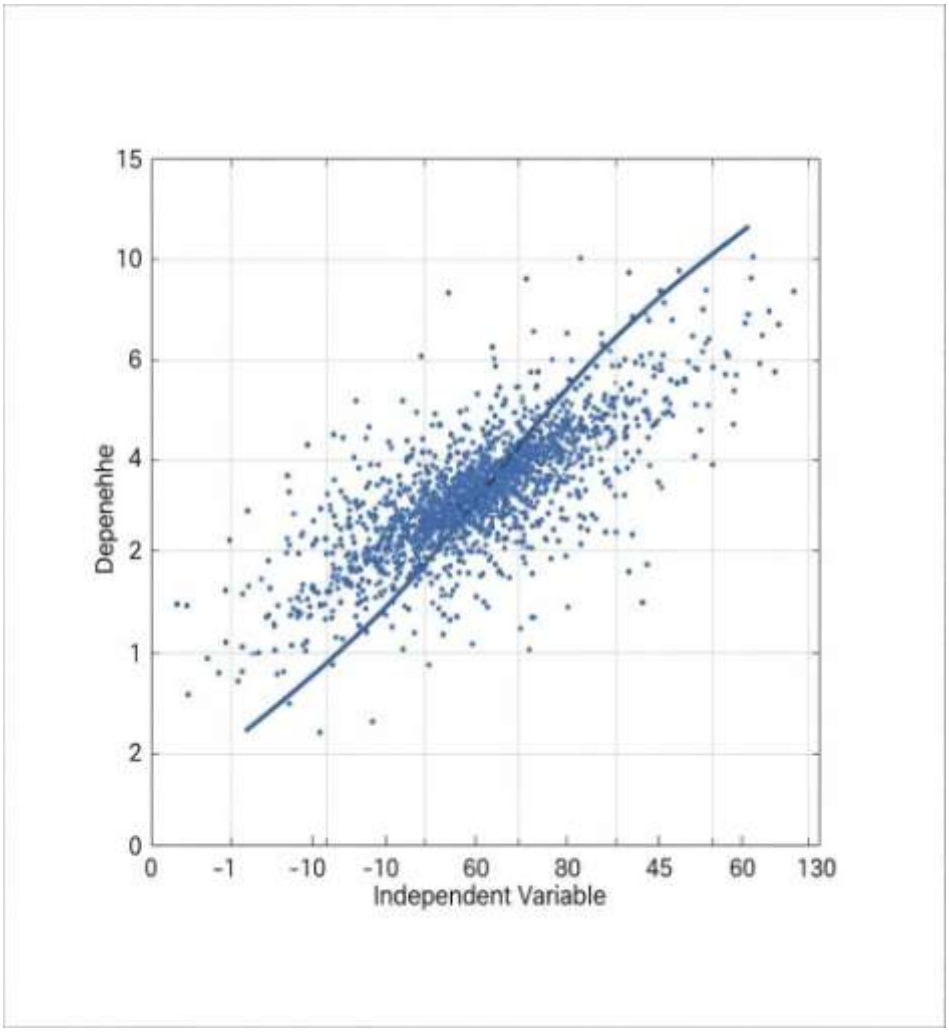

Figure3.5:foundationalconceptsandpurposeofregressionanalysis.

values to forecast the dependent variable's value. Regression's capacity to forecast future outcomes from historical data renders it one of the most essential statistical models now employed, with applications across diverse domains such as economics, finance, social sciences, and engineering. This will allow researchers to detect and study these interactions, quantify their strengthanddirection,andsopredictandgeneralizeresults.Linearregression is the fundamental form of regression that assumes a linear relationship between variables, although polynomial regression and multiple regression can accommodate non-linear correlations and numerous predictors. Regression provides methods to evaluate model's goodness of fit, indicating its explanatory power about the data, and to analyze the statisticalsignificance of the predictors.; and flag potential outliersor influential data points. Well, it is essential since regression analysis offers a mechanismthat helps understand and qualify relationships, including how variables influence each other.

BuildingandInterpretingaLinearRegressionModel:AStep-by-StepNumericalExample

We will use a numerical example todemonstrate how to build and interpret a simple linear regression model. Let's say we wish to study the effect of number of hours students' study for an exam (independent variable, X) on their score in exam (dependent variable, Y). Data we collected from six students:

Student	HoursStudied(X)	ExamScore(Y)
A	2.0	55
B	3.0	60
C	4.0	68
D	5.0	72
E	6.0	78
F	7	85

Step1:CalculateMeanofXandY

First,wecalculate meanofX(denoted as$\bar{X}$)&meanof Y(denoted as$\bar{Y}$).

$\bar{X}$=(2+3+4+5+6+7)/6=27/ 6=4.5$\bar{Y}$=(55+ 60+68+72+78+85) / 6 = 418 / 6 =69.67

Step2:CalculateDeviationsfromMean

Next,wecalculatedeviationsofeachXvaluefrom$\bar{X}$(x=X-$\bar{X}$)&deviations ofeachYvaluefrom$\bar{Y}$(y =Y-$\bar{Y}$).

Student	X	Y	x(X-$\bar{X}$)	y(Y-$\bar{Y}$)
A	2	55	-2.5	-14.67
B	3	60	-1.5	-9.67
C	4	68	-0.5	-1.67
D	5	72	0.5	2.33
E	6	78	1.5	8.33
F	7	85	2.5	15.33

Step3:CalculatetheProductsofDeviations(xy)andSquaredDeviations(x^2)

We then calculate the product of deviations (xy) and squared deviations of X(x^2).

Student	x	y	xy(x*y)	x^2(x*x)
A	-2.5	-14.67	36.675	6.25
B	-1.5	-9.67	14.505	2.25
C	-0.5	-1.67	0.835	0.25
D	0.5	2.33	1.165	0.25
E	1.5	8.33	12.495	2.25
F	2.5	15.33	38.325	6.25

Step4:CalculateSumsofxyandx^2

Wecalculatesumsofxy (Σxy)andx^2(Σx^2).

Σxy=36.675+14.505+0.835+1.165+12.495+38.325=104Σx²=6.25+ 2.25 +0.25 +0.25 +2.25+6.25 =17.5

Step5:CalculateSlope(b)andIntercept(a)

Theslope (b)of regressionlineiscalculated as:

b =Σxy / Σx² =104 / 17.5 =5.94(approximately)

Theintercept(a)iscalculated as:

a=Ȳ -bX̄= 69.67 -(5.94 * 4.5)= 69.67 -26.73=42.94(approximately)

Step6:WriteRegressionEquation

Theregressionequationis:

Ŷ=a+bX Where:

- Ŷispredictedvalueof Y.
- aisintercept.
- bis slope.
- Xisindependentvariable.

Inourexample,regressionequationis:

Ŷ=42.94 +5.94X

Step7:InterprettheResults

- **Slope(b):** The slope of 5.94 signifies that for each additional hour studied, exam score is anticipated to rise by an average of 5.94 points.
- **Intercept(a):** The intercept (42.94) is the estimated exam score when the number of hours studied is zero. However, in this case, this may not be meaningful as one does not read for zero hours.

- **RegressionEquation:** The equation $\hat{Y} = 42.94 + 5.94X$ can be used to predict exam scores for different study times. For example, if a student studies for 8 hours, the predicted exam score would be: $\hat{Y} = 42.94 + (5.94 * 8) = 42.94 + 47.52 = 90.46$.

Step8:AssesstheGoodnessofFit(R-squared): R-squared (R^2) quantifies proportion of variance in dependent variable that can be anticipated from independent variable. It varies from 0 to 1, with 1 signifying an ideal fit.

To calculate R-squared, we need to find sum of squares regression (SSR) and total sum of squares (SST).

$SSR=\Sigma(\hat{Y} - \bar{Y})^2 SST = \Sigma(Y-\bar{Y})^2$

Then, $R^2 = SSR / SST$

Usingstatisticalsoftwareorcalculators,wecandeterminetheR-squaredvalue for this example. A high R-squared value indicates that model fits the data well.

Step9:TesttheSignificanceoftheRegressionCoefficients

Then, we can conduct hypothesistests to check if the slope and the intercept are statistically significant. Therefore, computing t-statistics and p-values. Reject null hypothesis if p-values are below significance level (e.g., 0.05), indicating that coefficients are significant.

Step10:AnalyzeResiduals: These residuals are the differences of actual Y and predicted $\hat{Y}$. Residuals analysis also assists in detecting outliers, non-linearities, and assumption violations. To check for patterns we can plot residuals against predicted values or independent variables.

MultipleRegression: With more than one independent variable involved, we conductmultiple regression.Theprocessissimilar, butthemathgets trickier. Multiple regression analysis is typically undertaken using statistical software.

UNIT3.6 LEASTSQUAREFITOFLINEARREGRESSION

3.6LeastSquareFitOfLinearRegression

Linear regression is arguably most elementary statistical technique for modeling relationship between two variables: an independent variable (predictor) & dependent variable (target). We are doing linear regression to identify the line that optimally fits this data in terms of least squares. The predominant approach for doing this is "least squares fit" method. It aims to minimize squared sum of the discrepancies between the observed values of the dependent variable and the values predicted by linear function. These discrepancies,termed residuals, represent the errorsbetween themodel and the actual data points. This would reduce the total error: the aggregate of all squared projected errors throughout the dataset to identify the line that most accurately represents the linear connection, offering a valuable framework for analyzing or predicting trends. This foundational technique is employed across various fields, including economics, finance, engineering, & social sciences, enabling analysis & prediction of linear relationships. The derived linear equation, typically expressed as $y = mx + b$ (where m represents slope & b denotes the y-intercept), offers a straightforward and efficient method for analyzing relationships and generating data predictions. The slope (m)indicates variation in the dependent variable for each unit change in the independent variable, whereas y-intercept (b) denotes value of dependent variable when independent variable is zero. We choose the least squares method because it isan optimal and unique solution and makes sure that the resulting line is the best linearization of the data. It is also mathematically tractable, familiar formulas for slope and intercept can be derived, making it feasible to do the math’s manually and not just place the formula on the computational side.

CalculatingtheLeastSquaresLine:AStep-by-StepNumericalExample

Now, we will demonstrate how to find the leastsquares fit using an example with numbers. Let's say we're trying tofigure out the relationship betweenhow many hours students' study (x) & their exam scores (y): We collect following data:

HoursStudied(x)	ExamScore(y)
1	2
2	4
3	5
4	4
5	7

Step1:CalculatetheSums

Wefirst calculate thesumsof x,y, x^2, andxy:

- $\Sigma x = 1 + 2 + 3 + 4 + 5 = 15$
- $\Sigma y = 2 + 4 + 5 + 4 + 7 = 22$
- $\Sigma x^2 = 1^2 + 2^2 + 3^2 + 4^2 + 5^2 = 1 + 4 + 9 + 16 + 25 = 55$
- $\Sigma xy = (1*2) + (2*4) + (3*5) + (4*4) + (5*7) = 2 + 8 + 15 + 16 + 35 = 76$

Step2:CalculateNumberofDataPoints(n)

Inthiscase,n=5.

Step3:CalculateSlope(m)

Theformulafor theslope(m)is:$m = (n\Sigma xy - \Sigma x\Sigma y) / (n\Sigma x^2 - (\Sigma x)^2)$

Pluggininthevalues:$m = (5*76 - 15*22) / (5*55 - 15^2) m = (380 - 330) / (275 - 225) m = 50 / 50 m = 1$

Step4:CalculatetheY-Intercept(b)

The formula for the y-intercept (b) is:

$b = (\Sigma y - m\Sigma x) / n$

Plugging in thevalues:

b =(22-1 * 15)/ 5 b=(22 -15) / 5 b =7 / 5 b =1.4

Step5:WritetheLinearEquation

Theequation ofleast squareslineis:y =mx +b y=1x +1.4 y =x+1.4

Interpretation: Our slope (where m = 1) means that for every extra hour studied, exam score is 1 point more. The y-intercept (b=1.4): this is the predicted amount of exam score when the student spends0 hours studying

AssessingtheFit: The coefficient of determination (R^2) can be computed to assess adequacy of line's fit to the data. R^2 multiplied by 100 yields the percentage of variance in y that is accounted for by x. NOTE: A higher R^2 means the regression fits data better.

CalculatingR^2

1. Calculatemean of y ($\bar{y}$):$\bar{y} = \Sigma y / n = 22 / 5 = 4.4$
2. Calculatetotalsumofsquares(SST):$SST = \Sigma(y - \bar{y})^2$
3. Calculate the regression sum of squares (SSR): $SSR = \Sigma(\hat{y} - \bar{y})^2$ (where $\hat{y}$ is the predicted y)
4. $R^2 = SSR / SST$

By computing these sums and applying the formula, we can determine the R^2 value and assess the goodness of fit of the linear regression model.

ApplicationsandImportance: Least squares linear regression, used throughout many areas. In economics, it can model the correlation between GDP and unemployment. In finance, it can forecast stock prices from the market indicators. In engineering, itcan study correlation between input and output variables in a system. In the world ofsocial sciences, it can concisely describe the relationship between educational attainment and income.

UNIT3.7 TWOLINESOFREGRESSION

3.7UnderstandingRegressionanditsDualNature

Regression analysis is statistical technique used to model and examine relationship between two or more variables. For two variables, it aims to determine a line that optimally fits data points on a scatter plot, enabling prediction of one variable's value based on other variable's value. The concept of "best fit" can be understood in two distinct manners, resulting in two regression lines: the Y on X regression line (Y = a + bX) and X on Yregression line (X = c + dY). The regression line of Y on X is utilized to forecast the values of Y based on value of X, with X being the independent variable (predictor) and Y dependent variable (response). The regression lineof X on Y is utilized to forecast the values of X based on the values of Y, with Y designated as the independent variable and X as the dependent variable. These two lines illustrate differing viewpoints of the same relationship, withthe slope and intercept defining the nature and degree of that association. The mean for both variables is the intersection point of these two lines. Having an understanding of the context of the data and where you want to predict is

Figure3.6IndexNumberRealValues

important to identify which regression lineto use. Overlap of data on those linessuggests the accuracy level of prediction.

CalculatingandInterpretingTwoLinesofRegression:APracticalApproachwith NumericalExamples

Now I want to give you a numerical example todemonstrate the computation and meaning of two lines of regression. Let us assume we want to study relationshipbetweennumberofhoursstudents'study(X),&theirexamscores (Y).Wegatherdatafrom 5 students.:

Student	HoursStudied(X)	ExamScore(Y)
A	2	50
B	4	60
C	6	70
D	8	80
E	10	90

1. CalculateMeansofX&Y:

- Meanof X ($\bar{X}$) = (2 +4+6 +8 +10) / 5 =30 / 5=6
- Meanof Y ($\bar{Y}$) = (50 +60 +70 +80 +90) / 5 =350 / 5 =70

2. CalculatetheSumofSquaresandCross-Products:

- $\Sigma(X-\bar{X})^2=(2-6)^2+(4-6)^2+(6-6)^2+(8-6)^2+(10-6)^2=16+4+0+4+16 =40$
- $\Sigma(Y - \bar{Y})^2 = (50-70)^2 + (60-70)^2 + (70-70)^2 + (80-70)^2 + (90-70)^2 = 400 +100 + 0 + 100 + 400 = 1000$
- $\Sigma(X - \bar{X})(Y - \bar{Y}) = (2-6)(50-70) + (4-6)(60-70) + (6-6)(70-70) + (8-6)(80-70) + (10-6)(90-70) = 80 + 20 + 0 + 20 + 80 = 200$

3. CalculatetheRegressionCoefficients:

- **RegressionCoefficientofYonX(b):** $b = \Sigma(X - \bar{X})(Y - \bar{Y}) / \Sigma(X - \bar{X})^2 = 200 / 40 = 5$
- **RegressionCoefficientofXonY(d):** $d=\Sigma(X-\bar{X})(Y-\bar{Y})/\Sigma(Y-\bar{Y})^2= 200 / 1000 = 0.2$

4. CalculatetheIntercepts:

- **InterceptofYonX(a):**a=$\bar{Y}$-b$\bar{X}$=70 -(5* 6)=70-30 =40
- **InterceptofXonY(c):**c=$\bar{X}$-d$\bar{Y}$=6 -(0.2 *70)=6 -14 =-8

5. WritetheRegressionEquations:

- **RegressionLineofYonX:** Y= a+bX =40+5X
- **RegressionLineofXonY:**X=c+dY =-8+0.2Y

Interpretation:

- **YonX(Y=40+5X):** For every one-hour increase in study time (X), exam score (Y) is predicted to increase by 5 points. The intercept, 40, represents predicted exam score when no hours are studied, though this may not be practically meaningful.
- **XonY(X=-8+0.2Y):** For every one-point increase in exam score (Y), the study time (X) is predicted to increase by 0.2 hours. The intercept, -8, represents the predicted study time when the exam score is zero, which is also not practically meaningful.

UsingtheEquationsforPrediction:

- Ifastudentstudiesfor7hours(X=7),thepredictedexamscore(Y)is:Y =40 +(5 *7)=40 +35=75.
- If a student scores 85 on the exam (Y = 85), the predicted study time (X) is: X = -8 + (0.2 * 85) = -8 + 17 = 9 hours.

ImportantNotes:

- The regression lines should intersect at the mean values ($\bar{X}$, $\bar{Y}$), which in our exampleis (6, 70).
- The coefficients (b and d) represent the extent of change in dependent variablecorrespondingtoaunitchangeinindependentvariable.

UNIT3.8 PROPERTIESOFREGRESSIONOEFFICIENTS

3.8.1 UnderstandingtheFoundationofRegressionCoefficients

Regression analysis, a prevalent activity in statistical modeling, seeks to ascertain the response of a dependent variable (Y) to variations in one ormore independent variables (X).This approach centers on regression coefficients, which indicate the amount and direction of the influence of each independent variable on the dependent variable.In a fundamental linear regression model ($Y = \beta_0 + \beta_1 X + \varepsilon$), the coefficients denote the Y-intercept (β_0, the value of Y when X equals zero) and the slope (β_1, the variation in Y for each unit increment in X).In these cases, the least squares method is utilized to ascertain the coefficient values that minimize the sum of squared residuals between the observed Y and the predicted values Y hat.The characteristics of these coefficient values, such as unbiasedness, consistency, and efficiency, are essential for the reliability and validity of the regression model.Understanding these qualities enables researchers to make informed decisions about model selection, interpretation, and inferential implications. The coefficients are random variables which are calculated from

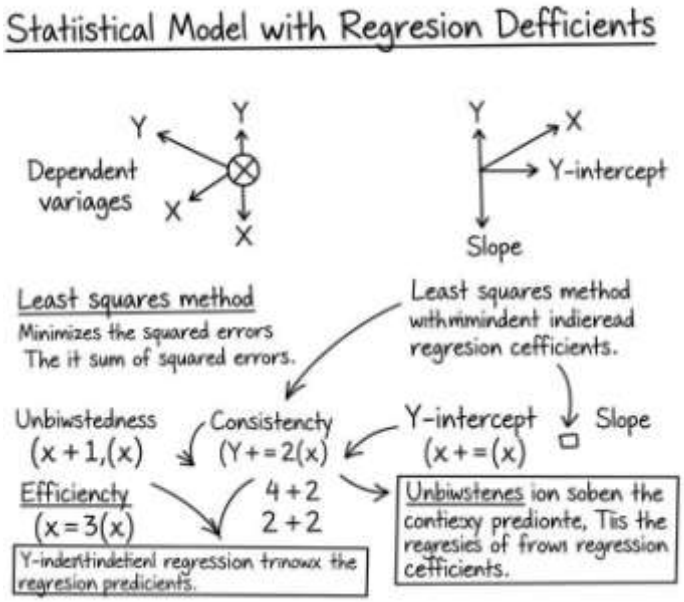

Figure3.7StatisticalModelWithRegressionDefficients

sample data, and their distributions are necessary for hypothesis testing and confidence interval construction. They are subject to the assumptions of the linear regression model (e.g., linearity, independence, homoscedasticity, normality of errors). If these assumptions are violated, the estimates may become biased or inefficient, which can affect the accuracy andgeneralizability of the regression outcomes.

KeyPropertiesandNumericalIllustration:DeconstructingtheBehaviorof β_0 and β_1

Regression coefficients have several important properties that make them reliableandusefulinstatisticalinference. Theordinaryleastsquares estimators of regression coefficients are unbiased when the classical linear regression model (CLRM) conditions hold. This indicates that, on average, the predicted coefficients will correspond to the genuine population coefficients. Secondly, they exhibit consistency, indicating that as sample size rises, calculated coefficients converge to true population values. Third, they are efficient, i.e. OLS estimators have minimum variance among every linear unbiased estimator. Fourth, OLS estimators follow a normal distribution which aids in hypothesis testingand creating confidence intervals. The covariance between the estimated coefficientsreveals the degree of interdependence among them. Now, let usproceed with anumerical example to put together these properties. Example: In correlation analysis, we may want to study the relation between no. of hours of study (X) and the unsigned exam scores (Y) for a group of students. We collect following data:

Student	HoursStudied(X)	ExamScore(Y)
A	2.0	60
B	3.0	70
C	4.0	80
D	5.0	90
E	6.0	100

Wewanttoestimatesimplelinearregressionmodel:$Y=\beta_0+\beta_1X+\varepsilon$.

1. CalculatingRegressionCoefficients:

Wecancalculatetheregressioncoefficientsusingthefollowingformulas:

$\beta_1=\Sigma[(Xi - \bar{X})(Yi-\bar{Y})]/\Sigma(Xi-\bar{X})^2\beta_0=\bar{Y}-\beta_1\bar{X}$

Where:

- $\bar{X}$ismeanof X.
- $\bar{Y}$ismeanof Y.

$\bar{X}=(2 +3+4 +5+6) /5 =4\bar{Y}=(60+70 +80+90 +100)/ 5 = 80$

Now,wecalculatethenecessarysums:

$\Sigma[(Xi-\bar{X})(Yi-\bar{Y})]=(-2)(-20)+(-1)(-10)+(0)(0)+(1)(10)+(2)(20)=40+$
$10+0+ 10+40=100\Sigma(Xi-\bar{X})^2= (-2)^2+(-1)^2 +(0)^2+ (1)^2+(2)^2=4+ 1+0$
$+1 +4 =10$

$\beta_1 =100 / 10 = 10\ \beta_0 =80-10 * 4 =80 -40 = 40$

Therefore,theestimatedregressionequationis: Y=40+10X.

2. Unbiasedness: In repeated sampling, the mean of the predicted β_1 values would converge to the true population β_1. If we were to replicate sampling and estimatingproceduremultiple times,averageofthe$\beta_1$valueswouldbecloseto 10.

3. Consistency: As the sample size increases, the estimated β_1 and β_0 values become closer to the true population values. If we collected data from a larger group of students, the estimated coefficients would be more accurate.

4. Efficiency: Among all linear unbiased estimators, the Ordinary Least Squares (OLS) estimators exhibit the minimal variation. This indicates that the predicted coefficients are the most accurate.

5. Normality: Under the CLRM assumptions, the estimated coefficients are normally distributed. This allows us to perform hypothesis tests and construct

confidenceintervals.Forinstance,wecantestnullhypothesisthat$\beta_1=0$(no relationshipbetweenhoursstudied&examscores)usingat-test.

6. Covariance: The covariance between β_0 & β_1 indicates how they vary together. A negative covariance suggests that as β_1 increases, β_0 tends to decrease, and vice versa. This is often observed in regression models.

7. VarianceoftheCoefficients: The variances of regression coefficients are crucial for assessing reliability of theestimates. They are calculated asfollows:

$Var(\beta_1) = \sigma^2 / \Sigma(Xi- \bar{X})^2 Var(\beta_0) = \sigma^2 [1/n+\bar{X}^2 /\Sigma(Xi-\bar{X})^2]$

Where σ^2 is variance of error terms. The standard errors of coefficients are square roots of these variances.

8. R-squaredandAdjustedR-squared: Understanding R-squared and Adjusted R-squared in Statistical Modeling

R-squared (R^2) is one of the most widely used metrics for evaluating the goodness-of-fit of statistical models, particularly in regression analysis. At its core, R^2 represents the proportion of variance in the dependent variable that is explained by the independent variable(s) in the model. This metric provides analystswitha straightforward interpretation:an R^2valueof 0.75indicatesthat approximately 75% of the variability in the outcome can be explained by the predictor variables included in the model.

However, R^2 has a fundamental limitation that necessitates caution in its application and interpretation. By mathematical construction, the R^2 value will always increase or, at minimum, remain unchanged when additional independent variables are introduced to the model, regardless of whether these new variables genuinely contribute meaningful explanatory power. This property creates a problematic incentive in model building, as it can lead analysts to artificially inflate their models with superfluous variables merely to achieve a higher R^2 value, potentially resulting in overfitting and reducedmodel generalizability.

This inherent limitation of R^2 led to the development of adjusted R-squared, which incorporates a penalty for each additional predictor variable added tothe model. Unlike standard R^2, adjusted R-squared increases only if the new variableimproves themodel morethanwould be expectedby chancealone. In some cases, adjusted R-squared can decrease when irrelevant variables are added, providing a more reliable indicator of model quality and a safeguard against unnecessarily complex models. When applying these concepts to practical data analysis, calculating both R^2 and adjusted R-squared offers valuable insights about model performance. The R^2 value provides a straightforward indication of how well the model captures the variance in the dependent variable, while adjusted R-squared serves as a check against overfitting by balancing explanatory power against model complexity. Together, these metrics form an essential part of the model evaluation toolkit, although they should be interpreted alongside other diagnostic measures such as residual analysis, hypothesis tests, and information criteria for a comprehensive assessment of model adequacy.

3.8.1HypothesisTesting:

Wecanconductt-teststoascertainstatisticalsignificanceofregression coefficients.Forinstance,wecanassessif$\beta_1$isstatisticallydistinct from zero.

T-tests in hypothesis testing are essential in regression research, offering a rigorous statistical framework to ascertain whether the patterns identified inour data likely represent true linkages in the larger population or are simplydue to sampling variability.In regression analysis, we derive coefficient estimates (such as β_1) that quantify the associations between independent variables and the dependent variable.Nevertheless, these estimates are proneto sampling error, necessitating a methodical approach to assess their trustworthiness. The t-test for regression coefficients fulfills this requirement by enabling us to evaluatewhether acoefficient significantly differsfromzero. A non-zero coefficient indicates that the associated independent variable significantly influences the dependent variable, while a coefficient indistinguishable from zero signals that the variable may lack substantial explanatory power in the model.

Theprocedurecommenceswiththeformulationofnullandalternative hypotheses.Thenullhypothesis(H_0)positsthatthecoefficientiszero(H_0:$\beta_1 = 0$), indicating an absence of correlation between the independent variableand the dependent variable.The alternative hypothesis (H_1) posits that the coefficient is not equal to zero (H_1: $\beta_1 \neq 0$), signifying the presence of a significant link. To conduct the test, we compute a t-statistic by dividing the estimated coefficient by its standard error : $t = \beta_1/SE(\beta_1)$.The t-statistic quantifies the number of standard errors the calculated coefficient deviates from zero.The greater the absolute value of the t-statistic, the more compelling the evidence against the null hypothesis.

We then compare this t-statistic to critical values from the t-distribution with theappropriatedegreesoffreedom(typicallyn-k-1,wheren isthesample size and k is the number of independent variables). Alternatively, we can calculate the p-value, which represents the probability of observing a t-statistic as extreme as ours if the null hypothesis were true. A small p-value (typically below 0.05)suggests that it's unlikely to observeourresults by chance aloneif no relationship exists, leading us to reject the null hypothesis. In business applications, thesetests help determine which variables significantly influence outcomes of interest. For example, a marketing team might analyze whether advertising expenditure significantly affects sales, or a financial analyst might assess whether certain economic indicators reliably predict stock returns. By applying hypothesis testing to regression coefficients, business professionals can make data-driven decisions with quantifiable levels of confidence, distinguishing between meaningful factors and statistical noise. While hypothesis testing provides valuable insights, it's important to interpret results in context, considering practical significance alongside statistical significance, particularly when working with large sample sizes where even small effects may appear statistically significant. Additionally, multiple hypothesis testing requires appropriate adjustments to control error rates across the entire set of tests.

3.8.2**ConfidenceIntervals:**

Confidence intervals provide range of plausible values for regression coefficients. They are calculated as:

$\beta_1 \pm t(\alpha/2,n-2) * SE(\beta_1) \beta_0 \pm t(\alpha/2,n-2) * SE(\beta_0)$

Where $t(\alpha/2, n-2)$ is critical value from t-distribution with n-2 degrees of freedom.

In this post, we will cover some essential properties of regression coefficients and whatthey can tell you about the relationships between variables in your data. These properties are crucial to the validity and utility of regression analysis in various disciplines.

Confidence Interval (CI)

1. Meaning

A Confidence Interval is a range of values, derived from sample data, that is likely to contain the true population parameter (like mean, proportion, or variance) with a certain level of confidence.

It gives both an estimate and the precision of that estimate.

Example:

If the sample mean height = 165 cm and the 95% CI is (162, 168), it means we are 95% confident that the true population mean lies between 162 cm and 168 cm.

2. Key Terms

Point Estimate: A single value estimate of a parameter (e.g., sample mean).

Interval Estimate: A range around the point estimate (CI).

Confidence Level: The probability that the CI includes the true parameter (commonly 90%, 95%, 99%).

Margin of Error (E): The amount added and subtracted from the point estimate to form the interval.

3. General Formula

For a population mean:

$CI = X +- Z \times \sigma / \sqrt{n}$

Where:

X = Sample mean

s = Population standard deviation (if known)

n = Sample size

Z = Z-score corresponding to the confidence level

If σ is unknown, use sample standard deviation s and the t-distribution:

$CI = X +- t \times s / \sqrt{n}$

4. Confidence Interval for Proportion

For population proportion p:

CI = p+- Z x√ p (1-p) / n

Where:

p = sample proportion

n = sample size

5. Interpretation

A 95% confidence level does not mean there is a 95% chance that the true value is in the interval.

It means that if we take many random samples, 95% of the confidence intervals constructed will contain the true parameter.

6. Factors Affecting CI

1. Sample size (n): Larger n → narrower CI → more precision.

2. Confidence level: Higher confidence (99% vs. 95%) → wider CI.

3. Variability in data (σ): Higher variability → wider CI.

7. Example

Suppose a sample of 100 students has a mean score = 70 and standard deviation = 10. Find the 95% CI for population mean.

CI = X +- Z x s / √n

= 70 +- 1.96 x 10 / √{100

= 70 +- 1.96

CI = (68.04, 71.96)

Interpretation: We are 95% confident that the population mean lies between 68.04 and 71.96.

8. Common Confidence Levels

90% CI → Z = 1.645

95% CI → Z = 1.96

99% CI → Z = 2.576

In short:

Confidence Interval gives a range estimate instead of just one value.

It combines point estimate + margin of error.

Higher confidence = wider interval, larger sample = narrower interval.

SELFASSENMENTQUESTION

Multiple-ChoiceQuestions(MCQs)

1. Whatdoescorrelationmeasure?

a. Thedifferencebetween two variables
b. Thestrengthanddirectionoftherelationshipbetweentwovariables
c. Thecausationbetweentwo variables
d. Theaveragevalue oftwovariables

2. Whichofthefollowingcorrelationvaluesindicatesthestrongestrelationship?

a. -0.85
b. 0.65
c. 0.25
d. -0.20

3. Whatdoesapositivecorrelationindicate?

a. Onevariableincreasewhiletheother decreases
b. Bothvariablesincreaseordecrease together
c. Thereisnorelationshipbetween variables
d. Onevariableremainsconstantwhiletheother increases

4. Whichmethodiscommonlyusedtomeasurecorrelation?

a. Standard deviation
b. KarlPearson'sCoefficientof Correlation
c. Movingaverage method
d. Chi-square test

5. WhatistherangeofKarlPearson'scorrelationcoefficient?

a. -2to 2
b. 0 to 1
c. -1to 1
d. -∞to ∞

6. WhichtypeofcorrelationdoesSpearman'sRankCorrelationmeasure?

a. Linear correlation
b. Non-linearcorrelation
c. Rank-basedcorrelation
d. Noneof the above

7. Whichofthefollowingisakeydifferencebetweencorrelationandregression ?

a. Correlation measures dependence, while regression measures association
b. Correlationdoesnotimplycausation,whereasregressiondoes
c. Correlationonlydescribestherelationship,whileregressionpredicts one variable based on another
d. Correlationrequiresmoredatapointsthanregression

8. WhatdoestheregressionequationY=a+bXrepresent?

a. Acorrelationequation
b. Therelationshipbetween independentanddependent variables
c. Thecalculationofmeanand median
d. Aprobabilitydistributionfunction

9. Whatarethetwolinesofregressioncalled?

a. RegressionofXon Y and RegressionofYon X
b. SimpleregressionandMultipleregression
c. KarlPearson'sregressionandSpearman'sregression
d. Linearregressionand Non-linearregression

10. WhatdoestheLeastSquaresMethodinregressiondo?

a. Itfindsthemedianofthe dataset

b. Itminimizesthesumofsquareddifferencesbetweenobservedand predicted values

c. Itmaximizesthecorrelationcoefficient

d. Iteliminatesallerrorsindata

11. Whichofthefollowingisapropertyofregressioncoefficients?

a. Theyarealwaysgreaterthan1

b. Theyareindependentof measurementunits

c. Theyremainconstantfor alldatasets

d. They indicate the change in the dependent variableforaunit changein the independent variable

12. WhichofthefollowingisNOTanapplicationofregressionanalysis?

a. Predictingstockprices

b. Findingrelationshipsbetweeneconomicindicators

c. Calculatingthemeanof a dataset

d. Forecastingbusinesstrends

13. Whatisthemainadvantageofusingregressionanalysis?

a. Ithelpsinestablishingcauseandeffect relationships

b. Itcalculatesaveragesquickly

c. Iteliminateserrorsinstatisticaldata

d. Itensuresthatcorrelationisalwaysequaltoone

14. Whichtypeofregressionisusedwhentherearemultipleindependentvariables?

a. Simplelinearregression

b. Multiple regression

c. Rankregression

d. Exponentialregression

15. Infinancialforecasting,regressionanalysisisusedtopredict:

a. Historicalstockprices
b. Futuretrendsbasedonpastdata
c. Fixedvaluesof assets
d. Theprobability ofan event occurring

ShortQuestions:

1. Definecorrelationand explainitsimportance.
2. Whatisthe differencebetweenpositiveandnegativecorrelation?
3. ExplainKarlPearson'sCoefficientofCorrelation.
4. WhatisSpearman'sRankCorrelation?
5. Defineregressionandits significance.
6. Whatarethetwolinesof regression?
7. Explaintheleastsquaremethodin regression.
8. Whatarethepropertiesofregression coefficients?
9. Howdoescorrelationdifferfromregression?
10. Whataretheapplicationsofregressionanalysisinbusiness?

LongQuestions:

1. Explaincorrelationanalysisandits significance.
2. DiscussthedifferencebetweenPearson andSpearmancorrelation.
3. Explaintheregressionanalysiswith examples.
4. Describetheleastsquare method anditsapplicationin regression.
5. Whatarethepropertiesofregression coefficients?
6. Explainhowcorrelationandregressionareusedinreal-worldscenarios.
7. CompareKarlPearson's andSpearman'scorrelationmethods.
8. Whataretheadvantagesandlimitationsofregression analysis?
9. Howdoescorrelationhelpinpredictive analytics?
10. Discusstheroleofregressioninfinancial forecasting.

Term	Definition
Correlation	The statistical measure of the relationship or association between two or more variables.Itindicateshowstronglyvariablesmovetogether,butdoesnotimply causation
Correlation Coefficient	Anumericalvalue(typicallybetween-1and1)thatmeasuresthestrengthand direction of a linear relationship between variables. Values close to 1 or -1 indicate strong positive or negative relationships, respectively; a value near 0 indicates no correlation
Positive Correlation	Arelationshipwhereincreasesinonevariableareassociatedwithincreasesin another
Negative Correlation	Arelationshipwhereincreasesinonevariableareassociatedwithdecreasesin another variable
Pearson Correlation	Acommonmethodtomeasurethelinearrelationshipbetweenvariables, sensitive only to that linearity
Spearman Correlation	Amethodtoassessmonotonic(notnecessarilylinear)relationshipsbetween variables, using ranked values
Independence	Aconditionwheretwovariableshavenodependence;theircorrelation coefficient is zero

Summary:CorrelationandRegressionAnalysis

Correlationandregressionanalysisarestatisticaltoolsusedtostudytherelationshipbetween two or more variables.

1. CorrelationAnalysis:

Correlationmeasuresthestrengthanddirectionofthelinearrelationshipbetweentwovariables. It does not imply causation, only association.

TypesofCorrelation:

Positive Correlation: Both variables increase or decrease together.

NegativeCorrelation:Onevariableincreasesastheotherdecreases. Zero

Correlation: No identifiable relationship between variables.

MeasuresofCorrelation:

Pearson'sCorrelationCoefficient(r):Measureslinearcorrelation;values rangefrom-1to+1.

+1:Perfectpositivecorrelation

-1:Perfectnegativecorrelation 0:

No correlation

Spearman'sRankCorrelation:Anon-parametric measureforrankeddata orordinalvariables.

2. RegressionAnalysis:

Regressionisusedtopredictthevalueofonevariable(dependentvariable)basedonthe value(s) of another variable (independent variable).

Itprovidesan equationthatbest fitsthe data.

TypesofRegression:

Simple Linear Regression: One independent variable and one dependent variable.

MultipleRegression:Twoormoreindependentvariablespredictingadependentvariable.

RegressionEquation(Simple Linear):

$Y=a+bX$

Y = Dependent variable

X=Independentvariable

a= Intercept(valueofYwhenX=0)

b=Slope(rate of changein Yfor aunitchangein X)

KeyTerms:

CoefficientofDetermination(R^2):Representstheproportionofvarianceinthedependent variable explained by the independent variable.

Residual:Thedifference betweenobservedandpredictedvalues.

Applications:

Predictingtrends(sales,performance,growth) Business

forecasting

Social science research

Healthandbiologicalstudies

MultipleChoiceQuestionswithAnswers:

1. Whatdoescorrelation measure?

Answerb.Thestrengthanddirectionoftherelationshipbetweentwo variables

2. Whichofthefollowingcorrelationvaluesindicatesthestrongest relationship?

Answera.-0.85

3. Whatdoesapositivecorrelationindicate?

Answerb.Both variablesincreaseordecreasetogether

4. Whichmethodiscommonlyusedtomeasurecorrelation?

Answerb.KarlPearson'sCoefficientofCorrelation

5. WhatistherangeofKarlPearson'scorrelationcoefficient?

Answer c.-1to1

6. Which typeofcorrelationdoesSpearman'sRankCorrelationmeasure?

Answerc.Rank-basedcorrelation

7. Which of the following is a key difference between correlation and regression?

Answerc.Correlationonlydescribestherelationship,whileregressionpredictsone variable based on another

8. What does the regression equation Y = a + bX represent?

Answer b. The relationship between independent and dependent variables

9. What are the two lines of regression called?

Answer a. Regression of X on Y and Regression of Y on X

10. WhatdoestheLeastSquaresMethodinregressiondo?

Answerb.Itminimizesthesumofsquareddifferencesbetweenobservedandpredicted values

11. Whichofthe followingis aproperty of regression coefficients?

Answerd.Theyindicatethechangeinthedependentvariableforaunitchangeinthe independentvariable

12. WhichofthefollowingisNOTanapplicationofregressionanalysis?

Answer c. Calculating the mean of a dataset

13. Whatisthemain advantageof usingregression analysis?

Answera.Ithelpsinestablishingcauseandeffect relationships

14. Whichtypeofregressionisusedwhentherearemultipleindependentvariables?
Simple linear regression

Answer b. Multiple regression

15. Infinancialforecasting,regressionanalysisisusedtopredict:

b.Futuretrendsbasedonpast data

MODULE4TIMESERIESANALYSIS

Structure

OBJECTIVES

- Explaintheconcept,significance,andapplicationsoftimeseries analysis.
- Recognizeanddescribethedifferentcomponentsoftimeseries.
- Explain and compare additive, multiplicative, & mixed models of time series.
- Understandconceptoftrendanalysisanditsimportancein forecasting.
- Explainandimplementfreehandcurve,semi-averages,movingaverages, and least square methods for trend estimation.

UNIT4.1 INTRODUCTIONTOTIMESERIESANALYSIS

4.1INTRODUCTIONTOTIMESERIESANALYSIS

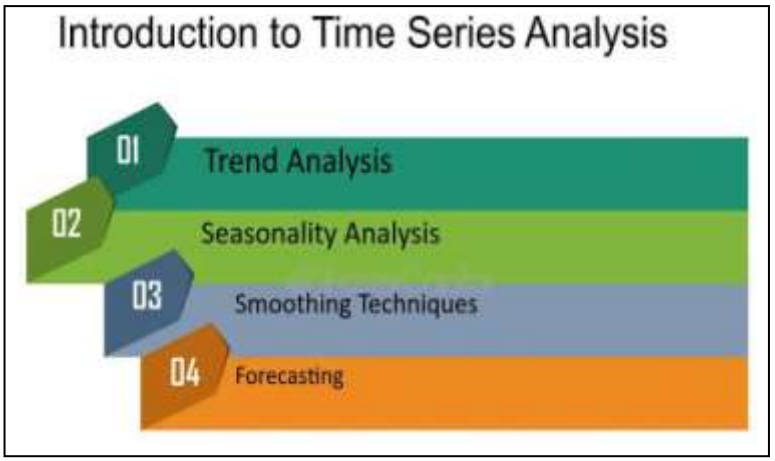

Figure4.1:IntroductiontoTimeSeriesAnalysis.

Time series analysis is study of data points collected, or recorded, at specific time intervals and allows you to analyze the data point readings over time to better understand what happens in the future based onpreviously determined values. In contrast to cross-sectional data, which reflects a snapshot of observations at a given point in time, time-series data exposed trends, seasonality, and cyclical behavior that are endemic to temporal sequences. Such analysis is vital inmany fields ranging from economics (predicting stock prices or inflation) to environmental science (weather andclimate patterns) to even signal processing (understanding the variation in audio waves). A time series is a type of dependent data; for any point in time, the value will usually depend on the previous value. For a better analysis of time series, we usually decompose it into a few components: a trend (long-term movement), seasonality (repeated patterns with a fixed time interval), cyclical component (long-term variance), and random or irregular components (unpredictable noise). By comprehending these factors, we can simulate the fundamental mechanisms and generate educated forecasts. For example: retail sales may show ayearly trend of increase, seasonal peaks around holidays, and outlier drop/ups due to unexpected occurrences.

NumericalExample:AnalyzingMonthlySalesData

Let's illustrate time series analysis with a simple numerical example. Suppose we have monthly sales data for a small bookstore over a year:

Month	**Sales (Units)**
Jan	120
Feb	130
Mar	150
Apr	160
May	170
Jun	180
Jul	190
Aug	200
Sep	180
Oct	160
Nov	220
Dec	250

1. VisualizingtheTimeSeries:

The first task is to plot data, specifically time series with months forx-axisand sales for the y axis. This image shows a positive line, indicating sales are better throughout the year. You also see a peak of sales in November and December, which suggests some seasonality due to holiday shopping.

2. IdentifyingTrend:

To identify the trend, we can use a moving average. A 3-month movingaveragesmoothsoutshort-termfluctuations&highlightsthelonger-termtrend. For example, the moving average for March is (120 + 130 + 150) / 3 = 133.33.

Month	Sales(Units)	3-MonthMoving Average
Jan	120	-
Feb	130	-
Mar	150	133.33
Apr	160	146.67
May	170	160
Jun	180	170
Jul	190	183.33
Aug	200	190
Sep	180	193.33
Oct	160	180
Nov	220	200
Dec	250	210

3. DetectingSeasonality:

Seasonal indices can be calculated in order to detectseasonality. Forease of calculation, let's examine the December spike. We will take the average sales across all months and compare the sales for December to this average. Monthly Average Sales:

(120+130+150+160+170+180+190+200+180+160+220+250)/12=184.17
Decemberseasonalityindex=250/184.17=1.36Thismeansthatsalesin December is about 36% morethan monthly average sales.

4. SimpleForecasting:

Wecancomputeanaive forecastusingtrendandseasonality.Using seasonaladjustment,extrapolateuptoJanuaryofthefollowingyearassuming those trends hold. But for convenience, we may also take theaverage of the last few months moving average, and consider slight uptrend.

FurtherAnalysis: Applicationsfor more broad-spectrum techniques such as ARIMA models, exponential smoothing, decomposition methods, can also be used for more clarified forecasting here. These are adjusted forautocorrelation, the correlation of values at different time points. This is a simple example on how time seriesanalysis works. Analyzing loaddata, we traintimeseriesmodelstomakepredictionsinproductionsystems.

UNIT4.2 COMPONENTSOFTIMESERIES

UnravelingDynamicsofTime-DependentData

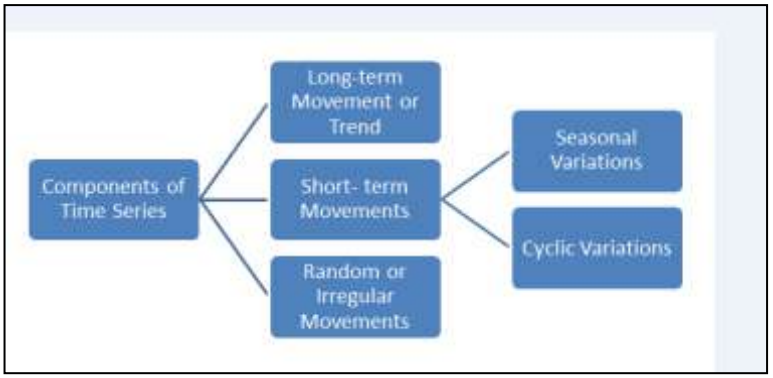

Figure4.2:ComponentsConcerningTimeSeries.

FeatureEngineeringforTimeseriesdataTimeseriesdataiskindofdatathatis used in time series analysis which is an important analytical method that used to analyze time series data to extract interesting statistics and other characteristics data. Seemingly,this data sets are collected over time, and are coming in at regular intervals, and such data usually has complex patterns about them which can be broken down into several components.Understanding and separation of these elements are necessary for proper prognostication & rationalization of the business decision. Trend, seasonality, cyclicalvariations, & irregular fluctuations that are four main components of any time series. The trend refers tolong-term movement of data, whether upor down, over several months or years. Seasonality is the repetitive patternsthat happen on a shorter time span, like daily, weekly, monthly or yearly. Cyclical variations are long-run oscillations of indefinite frequency associated with business cycles or economic conditions. Finally, uneven oscillations (or random noise) are variations that cannot be attributed to any of the other components; they are unfurling in a random manner. Extracting these components from a time series provides us with useful information aboutthe main mechanisms that drive the time series, helps generate better predictions, and helps develop a clearer picture of the underlying process that generates the observed results.

NumericalExample:DecomposingSalesData

Consider a company's quarterly sales data for three years (12 quarters). Let's illustrate how these components might manifest and how we can conceptualize their impact.

Quarter	Year1	Year2	Year3
Q1	110	130	155
Q2	120	145	170
Q3	105	125	150
Q4	135	160	190

1. Trend: Note that totalsales figures are increasing over the three years. This also means that thetrend isapositiveone. Thus, ifweplot thequarterlysales, we can see the general upward slope. From week to week, it can look like a mountain rangeso using a simple moving average to smooth the bumps outand show the general trend helps. For example,a four-quarter moving average would smooth sales over four successive quarters, uncovering the underlying upward trend.

2. **Seasonality:** Notethat Q4 always has the highest sales, while Q3 has the lowest. “Such seasonal patterns may be driven by holiday shopping-related eventsinQ4.Wecandiscussseasonalindicestoquantifythisseasonality.We can compute the average sales for that quarter across years andthen divide it by the overall average sales. This measures the amount that seasonal effects cause an individual quarter tovary from the overall mean.

- AverageQ1: (110+130+155)/3 = 131.67
- AverageQ2: (120+145+170)/3 =145
- AverageQ3: (105+125+150)/3 = 126.67
- AverageQ4: (135+160+190)/3 = 161.67
- Overall Average:

(110+120+105+135+130+145+125+160+155+170+150+190)/12=143.33

- Seasonalindexfor Q1:131.67/143.33= 0.92

- Seasonalindexfor Q2:145/143.33= 1.01
- Seasonalindexfor Q3:126.67/143.33= 0.88
- Seasonalindexfor Q4:161.67/143.33= 1.13

These indices show Q4 sales are about 13% higher than average due to seasonality, and Q3 sales about 12% lower.

3. CyclicalVariations: Werethis company to exist in a cyclical industry, we mightwitnesslonger-termswingsbeyondseasonaltrends.Salesmightdropoff over a few years and then recover behind a broader economic downturn,for instance. Spotting cyclical fluctuationstypically needs longer time series data and advanced statistical methods.

3. **IrregularFluctuation**: After removing trend, seasonality, and cyclical variations from thedata, therewill bestill berandom variations. Thesemay be because somethingunexpected happened, like a shift in consumer behavior,the unexpected success of a marketing campaign, or asupply chain problem. These variationsare non-deterministic and are usually described as a random noise.

By identifying and separating these components we are able to create more accurate forecasting models. We can time-shift the data by dividing the actual sales by the seasonal indices to separate out what underlying trend isactually there. It can capture the longer-term trend as well as the repeating seasonal patterns fora better prediction of future sales.

UNIT4.3 MODELOFTIMESERIES

4.3 MODELOFTIMESERIES

Timeseries data which is asequenceofobservations recorded overa period of timeusually showcomplex patterns that can hide underlying trends orseasonal fluctuations. In short,we can usedifferent techniquesto decomposetimeseries into itselements to then analyze and forecast it. These elements often consistof a trend component (long-term trend), a seasonal component (repeatable fluctuations), a cyclic component (long-term disturbances), and a residual or irregular component (random noise in general). Additive, multiplicative, and mixed models are among the common decomposition models that help determine the models as per how the components interact. The selection of model is depending on data as well as the different relationships among its constituent components. All components are assumed to be independent and additively contribute to the final outcome in the additive model. A multiplicative model multiplies the components together with dependent effects. A mixed model is a combination of both approaches, which providesa better representation for more complicated time series. This analysis offers crucial insights into the underlying dynamics, allowing businesses and researcherstobeequippedwithdata-drivendecisionsandpredictionsbasedon past behavior and trends these become apparent.

4.3.1 AdditiveandMultiplicativeModels:ContrastingApproaches

This algebraic equation of additive time series model for Yt which is the value/time series is the sum oraddition of Trend (Tt), Seasonal (St), Cyclical (Ct), and Irregular (It). This is ideal for seasonality when the absolute size of the seasonal variations are similar, over time, independent of the trend level. For examples, suppose monthly ice cream sales, increase or decrease by a fixedamounteveryyearregardlessofthetotalsalestrend.Thiswouldindicate that the additivemodel would be appropriate.

Multiplicative Model: This model assumes that time series is result of componentsmultiplytogethertogivethetimeseries Yt=Trend(Tt)*

Seasonal (St) * Cyclical (Ct) * Irregular (It). This model is suitable when amplitudeof the seasonal variation's changes in proportion with trend level. Forinstance, multiplicativemodel would bemore suitableifthe monthly sales of a luxury product go through a more pronounced seasonal variability when sales are high and amore moderate seasonal variability when sales are low.

NumericalExample:ComparingAdditiveandMultiplicativeModels

Let's illustrate these models with a numerical example. Suppose we have quarterly sales data for a product over two years:

Quarter	Year1Sales	Year2Sales
Q1	110	121
Q2	120	132
Q3	130	143
Q4	140	154

1. TrendComponent:

First,wecalculatethetrendusingamovingaverage.Forsimplicity,we'll usea 4-quarter moving average.

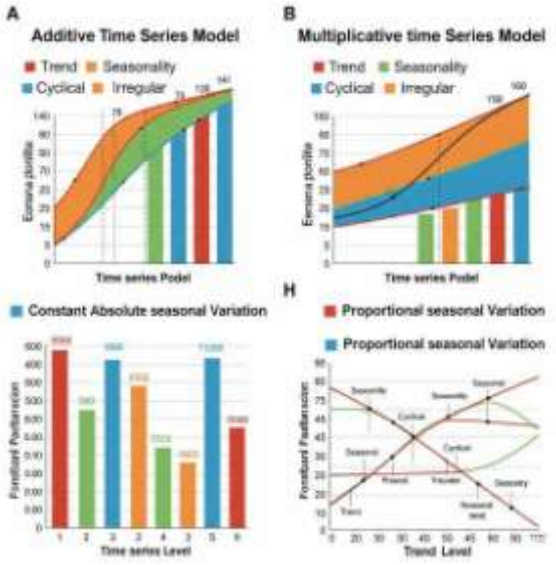

Figure4.3:AdditiveandMultiplicativeModels:

- (110+120+130+140)/4=125Year2:
- (121+132+143+154)/4=137.5

2. SeasonalComponent(AdditiveModel):

To estimate the seasonal component for additive model, we calculate average deviation from the trend for each quarter.

- Q1:(110-125)+(121-137.5)/2=-15.5
- Q2:(120-125)+(132-137.5)/2=-5.5
- Q3:(130-125)+ (143-137.5)/2 =5.5
- Q4:(140-125)+ (154-137.5)/2 =15.5

3. SeasonalComponent(MultiplicativeModel):

Forthemultiplicativemodel,wecalculateaveragerationofactualsalesto trend for each quarter.

- Q1: (110/125) +(121/137.5)/2 =0.88 +0.88/2 =0.88
- Q2: (120/125) +(132/137.5)/2 =0.96 +0.96/2 =0.96
- Q3:(130/125)+(143/137.5)/2 =1.04 +1.04/2 =1.04
- Q4: (140/125) +(154/137.5)/2 =1.12 +1.12/2 =1.12

4. DecomposedValues:

- **AdditiveModel:**
  - Year1Q1:125 -15.5 =109.5
  - Year1Q2:125 -5.5 =119.5
  - Year1Q3:125 +5.5 =130.5
  - Year1Q4:125 +15.5 =140.5
  - Year2Q1:137.5 -15.5=122
  - Year2Q2:137.5 -5.5 =132
  - Year2Q3:137.5 +5.5 =143
  - Year2Q4:137.5 +15.5=153

- **MultiplicativeModel:**
  - Year1Q1:125 *0.88 =110
  - Year1Q2:125 *0.96 =120
  - Year1Q3:125 *1.04 =130
  - Year1Q4:125 *1.12 =140
  - Year2Q1:137.5 *0.88 =121
  - Year2Q2:137.5 *0.96 =132
  - Year2Q3:137.5 *1.04 =143
  - Year2Q4:137.5 *1.12 =154

In this simplified example, the multiplicative model exactly reproduces the original data, suggesting it is a better fit. However, real-world data is rarelythis perfect.

4.3.2 MixedModelandModelSelection

The mixed model is a combination of both the additive model and multiplicative model, and implementations of this model can be more complex than both components. For instance, it could assume that trend and cyclical components are additive, but seasonal and irregular ones are multiplicative. A log additive model is beneficial in cases where the data hasboth additive and multiplicativecomponents.Amixedmodelcanbearticulatedinseveralforms' contingent upon its intended application. For example, Yt = Tt + St It.This involves examining features of the time series to identify trending behavior or seasonal patterns within it. An initial impression can be obtained throughvisual inspection of the time series plot. Seasonal fluctuations can be constant or can be proportional to the trendstatistical tests like the F-test for homogeneity of variance can be performed in order to decide. Also, the analysisofthenextresiduals(thedifferencebetweentherealanddecomposed values) can inform us about the model chosen. If the residuals form a random pattern then model is said to bea good fit. Looking at the residuals should all be random and independent of the fitted values, if they are systematic, including being auto correlated or het exorcistic, weneed to adjust the models.

UNIT4.4 TRENDANALYSIS

I still consider myself a newbie in thisdomain, but I like to know about Trend Analysis which is a statistical analysis made over time series data to identify patterns and direction. So it looks at data that gets collected regardless, at regularintervals,likedailyfiguresonsales,monthlyreportsonwebvisitors,or annual statistics on economic metrics, so that it can analyze the trends they form and project the likely values they will have at future points. While descriptive statistics provide a summary of data at a specific moment in time, trend analysis looks at change in data over timeto identify long-term trends, seasonal variations and cyclical shifts. Accurate forecasting is necessary for decision-making in many domains, ranging from business forecasts and financialplanningto scientificresearchandsocialpolicyformulation. Through data analysis and the identification of trends, organizations canforesee challenges and opportunities on the horizon, optimize resource allocation, and implement proactive measures. A retailer, for instance, may use trend analysis to anticipate seasonal demand for goods, a financial analystcould use it to project stock prices, or a public health official may use it to monitor the spread of a disease. Time series analysis is essentially about breaking down the time-series data and separating the trend, seasonality,cycles, and noise. This allows us to decompose the time series into various components as wealready see, where one often cares about the trend, which is the long-term movement in the data after removing the effects of other component. The trend (meaning up, down, or flat) tells you whether we are growing, declining, orstable. Different techniques likemoving averages, linear regression, and exponential smoothing are used to model and forecast none of which have a monopoly onstrengths or weaknesses.

4.4.1 *MethodsandNumericalExample:LinearTrendAnalysis*

Linear trend analysis is one of the easiest and popular methods for trend analysis where its assumption is the datais following a linear pattern in time. Linear Regression:This method involves fitting straight line to time series by linear regression, utilizing time as independent variable & observed values as dependent variable. The equation of line is expressed as y = a + bx, where a represents y-intercept &b denotes slope. The 'b' represents the slope of the linear trend, indicating rate of change, whereas 'a' (the intercept) denotes the initial value. To have further insight, let us do a numerical example. Let us examinethe subsequentsalesstatisticsofthecompany overa five-year period.:

Year(X)	**Sales(Y)(inthousands)**
1	10
2	12
3	15
4	18
5	20

To perform linear trend analysis, we first need to assign numerical values tothe years. We can simply use the year number (1, 2, 3, 4, 5) as the independent variable. Next, we calculate the necessary sums:

- $\Sigma X=1+2+3+4+5=15$
- $\Sigma Y=10+12+15+18+20=75$
- $\Sigma X^2=1^2+2^2+3^2+4^2+5^2=55$
- $\Sigma XY=(1*10)+(2*12)+(3*15)+(4*18)+(5*20)=249$
- n=5 (numberof data points)

Now, we can calculate slope 'b' and the intercept 'a' using the following formulas:

- $b=(n\Sigma XY-\Sigma X\Sigma Y)/(n\Sigma X^2-(\Sigma X)^2)$
- $a=(\Sigma Y-b\Sigma X)/n$

Plugging in the values:

- b=(5 *249 -15 *75) / (5 * 55-15²)=(1245 -1125) / (275 -225)=120/ 50 =2.4
- a=(75-2.4 * 15)/ 5 =(75-36) /5 =39 / 5 =7.8

Therefore, the linear trend equation is y = 7.8 + 2.4x. This equation indicates that the company's sales are increasing by 2.4 thousand units per year, with a starting point of 7.8 thousand units. To forecast sales for the next year (Year 6), we can plug in x = 6:

- y =7.8 +2.4 * 6 =7.8+14.4 =22.2

Thus, the forecasted sales for Year 6 are 22.2 thousand units. This method provides a simple and effective way to estimate and project linear trends, but it's important to note that it assumes a constant rate of change, which may not always hold true in real-world scenarios.

4.4.2 BeyondLinearity:AdvancedTrendAnalysisTechniques

linear trend is a great fit for simple datasets, most time series in the real world exhibit more complextrends. These complexitiesrequire advanced techniques to capture them. For example, moving averages smooth out short-term fluctuations by averaging data points over specified period. By averaging, we mitigate random noiseand may spot hidden trends. Where exponential smoothing applies exponentially decreasing weights to past observations, focusing more on recent observations. This method is especially effective at predicting time series that has trends andseasonality. Statistical Methods for Logistic Regression Seasonal Decomposition Seasonal decomposition is an effective techniqueemployed to disaggregate time series into its constituent components: trend, seasonal, & residual elements. This allows analysts to examine each individual segment without deciphering concealed meanings in the data.As an example, a retailer can use seasonal decomposition to analyze sales data and determine the seasonal peaks and troughs. techniques such as spectral analysis and wavelet analysis can also be applied to cyclical fluctuationsthatareessentiallylong-termvariationsofthetrend.Such

techniques enable the classification of periodic patterns and project future cycles. Apart from these classical methods, various machine learning techniques like ARIMA (Autoregressive Integrated Moving Average) and neural networks are also being used for trend analysis. Such ARIMA models tend to capture the autocorrelation and moving average components while neural networks are able to learn complex non-linearities. These advanced methodsoffermorepreciseforecastsandinsights,particularlyforintricate and fluctuating time series. They do, however, also need more computational resources and expertise. Assessing trend analysis accuracy is key to making accurate predictions. Different metrics, like mean absolute error (MAE), root mean square error (RMSE), and mean absolute percentage error (MAPE), are commonly used to measure 1 the discrepancy between forecasted and actual values. Introduction In order to decide what trend analysis method to use, analysts can compare the performance of various trend analysis methods. The data characteristics, accuracy requirements and resource availability will determine the appropriate trend analysis method. Time-series analysis is a powerful tool that can be used to extract insights from a wide range of data sources, and by understanding different methods available, analystscan better leveragethesetechniquestoinformtheirdecision-makingprocess.

UNIT4.5 METHODSOFTRENDANALYSIS

4.5.1 TheSignificanceofTrendAnalysis

Trend analysis is an important statistical approach that is used to analyses the patternanddirectionoftimeseries data.Analyzingtrendsinvolves discerning patterns and trends in values recorded over time, usually during regular intervals. This is critical across different fields, including economics and finance, environmental science, andmarketing. By identifying long-term movements, cyclical variations and seasonal fluctuations, businesses can forecast sales, governments canplan infrastructure and researchers can gainan understanding ofchanging phenomena. Trend analysis allowsus to identify thesignal from the noise thebasictrend that adataset is following and predict where it might head in the future. This data however is crucial for the comprehension of the past, present and possible future of datasets, it is inevitable. There are multiple approaches to accomplish this, which vary in benefits and constraints,and are more or less suitable for various data types and analytical requirements.

1. FreeHandCurve:AVisualApproachtoTrendIdentification

Logically, the easiestand subjective method of trend analysis is the freehand curvemethod. Theseinvolveplotting time-series dataand drawing agraph by hand, a smooth curve which best fits the general trend. This quick and simple method requires no complex calculations,suitable for a preliminary overview orwithsmalldatasets.Butitssystemissubjective,sodifferentanalysts might draw different curves and thus get different results. As an example, take the yearly sales figures of asmall book shop for 5 years: [20, 25, 30, 35, 40]. Ifwe plot these pointsand fit a line that tends to follow the upward direction,we can get a rough idea of the trend in sales. Although it is useful for a preliminary overview, it is not precise and objective as more sophisticated ways. Itismostusefulforarapidfirstpassatthedata,mostspecificallywhen a back-of-the envelope sense of the trend is all that is required.

2. Semi-Averages Method: Simplifying Trend Calculation

Thesemi-averagesmethodtriestoaddmoreobjectivityintotrendanalysis,for each half, you need to calculate theaverage value immediately. Averages are computed and then plotted at the midpoint of theirrespective time periods, with a straight line drawn between them. This line showsthe trajectory. For example,youmayhavetenyears'worthofsalesdata:[10,12,15,18,20,22, 25,28,30,32].Splittingitlikethisleadsusto[10,12,15,18,20]and[22, 25, 28, 30, 32]. They're averaging 15 and 27.4, respectively. Plotting these averages at the midpoints of their halves and drawing a connecting line givesa trend line. This method is easy and straightforward and also lesssubjective in comparison with a custom freehand curve. Yet, it assumes a linear behavior and it may not eventually reflectmore complex behavior. It is handy whenyou need a fast, less subjective approximation ofa linear trend.

3. MovingAveragesMethod:SmoothingOutFluctuations

The moving averages method is another highly popular method, which allows smoothing out the noise/volatility in the data and highlight the generaldirection in a long-term. The employed technique is moving average, which computes average value of specified number of successive data points. That average is then displayed at the halfway point of the period that theaverage covers. The more numberof data points you take for the average, smootherwillbethetrendline.Forinstance,forthesalesdata[10,12,15,18,20,22, 25, 28, 30, 32], we compute three-year moving averages like (10+12+15)/3 = 12.33,(12+15+18)/3 =15,etc. Plottingtheseaveragesshows a smoother trend linethantherawdata.Moving averages method is themost common technique used to smooth the dataas it effectively smooths with time ahead and helps to identify the long-term trend by reducing the impact of random variation. However, it may lag actual data especially during periods of rapid change and does not correspond to trends for the beginning or end of the time series. Choosing the moving average period is important and should be based on characteristicsofthedata&desiredamountofsmoothing.

4. LeastSquareMethod:PreciseTrendLineFitting

The least squares method is statistical technique that determines the optimal straight line by reducing total of squared deviations between observed data points & line. Its accuracy based solely onmath's, unlike always subjective based judgments. Trend-related equations are typically expressed as: y = a + bx, where y represents predicted value, x denotes time period, a signifies the y-intercept, and b indicates the slope. Let us examine data set [5, 8, 10, 12, 15] as an example. The slope & intercept of optimal line can be determined using the least squares approach. The slope signifies the pace of variation. While the intercept refers to the starting value. A method often used for forecasting, trend analysis, particularly when it is assumed that there is a lineartrend;themethodisquiteaccurate.Becauseitisoftencomputationally expensive and may not perform well withnonlinear trends. If accuracy and objectivity are paramount, as is the case with most statisticalapplications, use the least squares method that produces a trend line with the strongest statistical characteristics. The least squares method is a widely usedstatistical technique for determining the optimal straight line that best fits a given set of data points. It is primarily employed in regression analysis and trend forecasting to establish a mathematical relationship between dependent and independent variables. By minimizing the sum of the squared deviations between observed data points and the fitted line, the least squares method ensures an optimal representation of the data trend.

Unlike subjective judgment-based methods, which may introduce bias or inconsistency, the least squares method relies purely on mathematical principles. This makes it a preferred approach for analysts and researchers who seek objective and statistically robust models for decision-making.

Trend Analysis – Problem with Solutions

Common Methods

1. Freehand Curve Method
2. Semi-Average Method
3. Moving Average Method
4. Least Squares Method

Problem 1: Freehand Curve Method

The sales (₹000) of a company for 6 years are:

Year	2015	2016	2017	2018	2019	2020
Sales	120	135	150	165	170	185

Solution

Plot years on the X-axis and sales on Y-axis.

Draw a smooth curve passing through or near the points.

The curve shows the trend (upward in this case).

Sales are rising steadily over time.

Problem 2: Semi-Average Method

The production (in tons) of a factory for 8 years is:

Year	2012	2013	2014	2015	2016	2017	2018	2019
Output	80	90	92	100	110	115	120	130

Solution

Step 1: Divide the data into two equal parts.

First half (2012–2015): 80, 90, 92, 100 → Average = (80+90+92+100)/4 = 90.5

Second half (2016–2019): 110, 115, 120, 130 → Average = (110+115+120+130)/4 = 118.75

Step 2: Find mid-points of years.

First half midpoint = 2013.5, average = 90.5

Second half midpoint = 2017.5, average = 118.75

Step 3: Equation of trend line (Y = a + bX).

Slope b =118.75 - 90.5 / 2017.5 - 2013.5 = 28.25 / 4 = 7.0625$

Using point (2013.5, 90.5):

90.5 = a + 7.0625(2013.5)

Solve for a. After simplification, the trend equation:

Y = -14156.44 + 7.0625X

This equation can be used to forecast future production.

Problem 3: Moving Average Method

The sales (₹000) for 9 years are:

Year	2011	2012	2013	2014	2015	2016	2017	2018	2019
Sales	22	24	26	30	28	32	34	36	38

Solution

Take a 3-year moving average.

(22+24+26)/3 = 24

(24+26+30)/3 = 26.67

(26+30+28)/3 = 28

(30+28+32)/3 = 30

(28+32+34)/3 = 31.33

(32+34+36)/3 = 34

(34+36+38)/3 = 36

Moving averages: 24, 26.67, 28, 30, 31.33, 34, 36

Trend shows steady upward sales growth.

Problem 4: Least Squares Method

The income (₹000) of a firm for 5 years is:

Year	2016	2017	2018	2019	2020
Income	10	12	13	16	18

Solution

Equation of trend line:

$Y = a + bX$

Let 2018 = 0, then X values:

2016 = -2, 2017 = -1, 2018 = 0, 2019 = 1, 2020 = 2

X	-2	-1	0	1	2
Y	10	12	13	16	18
XY	-20	-12	0	16	36
X^2	4	1	0	1	4

Step 1: Calculate totals

sum Y = 69, sum X = 0, sum XY = 20, sum X^2 = 10

Step 2: Find constants

a = sum Y/ n = 69 / 5 = 13.8

b = sum XY / sumX^2 =20 / 10 = 2

Step 3: Trend equation

$Y = 13.8 + 2X$

Forecast for 2021 (X = 3):

$Y = 13.8 + 2(3) = 19.8$

SELFASSENMENTQUESTION

MultipleChoiceQuestions(MCQs)

1. WhatisTimeSeriesAnalysis?

A) Thestudyofhistoricaldatatoidentifypatterns over time

B) Theprocessofcalculatingaveragesofunrelated data

C) Amethodused onlyforfinancial forecasting

D) Atechniquetocollectsurveydata randomly

2. WhichofthefollowingisNOTacomponentoftimeseries?

A) Trend

B) Seasonality

C) RandomVariations

D) HypothesisTesting

3. Inanadditivetimeseriesmodel,howarethecomponentscombined?

A) Multiplication

B) Subtraction

C) Addition

D) Division

4. Whichofthefollowingisanexampleofamultiplicativetimeseriesmodel?

A) Y=T+S+C+RY=T +S+C+RY=T+S+C+R

B) Y=T×S×C×RY=T\timesS\timesC\times RY=T×S×C×R

C) Y=(T+S)×CY=(T + S)\times CY=(T+S)×C

D) Y=T−S−C−RY=T -S -C-RY=T−S−C−R

5. WhatdoestheFree-HandCurvemethodhelpinidentifying?

A) Cyclicalvariations

B) Trendcomponent

C) Seasonalvariations

D) Residualerror

6. WhatistheSemi-Averagesmethodusedfor?

A) Tocalculatemoving averages

B) Tosplitdataintotwoequalpartsandfindtrends

C) Toanalyzecyclicalvariations

D) Tomeasureseasonaleffects

7. IntheMovingAveragemethod,whathappenswhenthewindowsizeincreases?

A) Thetrend linebecomessmoother

B) Thefluctuationsincrease

C) Theseasonalvariationsbecomemore prominent

D) Theanalysisbecomesless reliable

8. TheLeastSquaresMethodisprimarilyusedto:

A) Findtherelationshipbetweentwoindependent variables

B) Fitatrendlinetohistoricaldata

C) Removeseasonal fluctuations

D) Analyzerandomvariations

9. Whichofthefollowingisamajorapplicationoftimeseriesanalysis?

A) Medicalresearch

B) Forecastingfuturesales

C) Analyzingsurvey responses

D) Predictingelectionresults

10. WhyisTimeSeriesAnalysisimportantinforecasting?

A) Itidentifiestrendsandpatternsinhistoricaldata

B) Iteliminatesallfluctuationsindata

C) Itremovesrandomnessfromfinancialmarkets

D) Itguaranteesaccuratefuturepredictions

11. WhatistheprimaryobjectiveofTrendAnalysisintimeseries?

A) Identifyinglong-termmovementindata

B) Removingseasonal fluctuations

C) Adjustingcyclical variations

D) Predictingshort-termrandomchanges

12. WhichofthefollowingisNOTatrendanalysistechnique?

A) Free-HandCurve Method

B) Semi-AveragesMethod

C) RegressionAnalysis

D) MonteCarlo Simulation

13. InwhichsectorisTimeSeriesAnalysiswidelyused?

A) Financial markets

B) Meteorology

C) Salesforecasting

D) All ofthe above

14. HowdoesTimeSeriesAnalysishelpinstockmarketpredictions?

A) Byensuring futurestockprices

B) Byidentifyinghistoricalpatternsandtrends

C) Byeliminatingmarketrisks

D) Byremovingexternaleconomicfactors

15. WhatisacommonchallengeinTimeSeriesForecasting?

E) Dataisalways accurate

F) Markettrendsremain constant

G) Presenceofrandomvariationsandexternalfactors

H) Lackofstatisticalmodels

ShortQuestions:

1. Whatistimeseries analysis?
2. Explainthedifferentcomponentsofatimeseries.
3. Whatisthedifferencebetweenadditiveandmultiplicative models?
4. Describethefree-hand curvemethod for trend analysis.
5. Whataresemi-averagesintimeseries analysis?
6. Howistheleastsquaremethodusedintrend analysis?
7. Whataretheapplicationsoftimeseries analysis?
8. Howdoestimeseriesanalysishelpin forecasting?
9. Whatistheimportanceoftrend analysis?

LongQuestions:

1. Explaintimeseriesanalysisandits significance.
2. Describethedifferentmodelsusedintimeseriesanalysis.
3. Discussthevariousmethodsoftrendanalysiswithexamples.
4. Explaintheleastsquaremethodanditsapplicationintimeseries.
5. Whataretheadvantagesofusingmovingaveragesintrend analysis?

6. Howdoestimeseries analysishelpinbusinessforecasting?
7. Comparethedifferenttrendanalysistechniques.
8. Discusstheimpactoftimeseriesanalysisonfinancial decision-making.
9. Explaintheroleoftrendanalysisin stockmarketpredictions.
10. Whatarethechallengesintimeseries forecasting?

Term	Definition
Time Series	Asequenceofdatapointscollectedorrecordedatregulartime intervals.
Stationarity	Apropertyofatimeserieswherestatisticalfeaturessuchasmean, variance, and autocorrelation are constant over time.
Trend	Thelong-termmovementordirectioninatimeseriesdataset.
Seasonality	Patternsthatrepeatatregulartimeintervals,suchasdaily,weekly, or yearly cycles.
Autocorrelation	Thecorrelationofatimeserieswithitsownpastandfuture values.
Lag	Theoffsetbetweencurrentandpreviousobservationsinatime series, often measured in time units.
Moving Average	Amethodtosmoothoutshort-termfluctuationsandhighlight longer-term trends by averaging data points over a specific window.
ARIMA (AutoRegressive IntegratedMovingAverage)	Apopularstatisticalmodelusedforanalyzingandforecasting time series data by combining autoregression, differencing (integration), and moving average components.
Forecasting	Theprocessofpredictingfuturevaluesinatimeseriesbasedon previously observed data.
SeasonallyAdjusted	Datathathavebeenmodifiedtoremovetheeffectsofseasonal patterns.
Differencing	Atransformationappliedtoatimeseriestomakeitstationaryby calculating the differences between consecutive observations.
White Noise	Atimeseriesofrandomvalueswithconstantmeanandvariance, and no autocorrelation over time.
ExponentialSmoothing	Aforecastingtechniquethatappliesexponentiallydecreasing weights to past observations.

Term	Definition
Residual	Thedifferencebetweenobservedandpredictedvaluesintime series modeling.
Decomposition	Theprocessofseparatingatimeseriesintotrend,seasonal,and residual components.

SummaryonTimeSeriesAnalysis

TimeSeriesAnalysisisastatisticaltechniqueusedtoanalyzedatapointscollectedorrecorded at specific time intervals. It helps in understanding underlying patterns, trends, seasonal variations, and forecasting future values.

KeyFeaturesofTime Series:

1. TimeOrder:Dataisarrangedinchronologicalorder.
2. ContinuousorDiscreteTime:Datamayberecordedhourly,daily,monthly,yearly,etc.
3. DependentNature:Currentvaluesdependonpreviousvaluesinthe series.

ComponentsofTimeSeries:

1. Trend(T):Long-termmovementordirectioninthedata.
2. Seasonality(S):Regularpatternthatrepeatsoveraspecificperiod(e.g.,yearly,quarterly).
3. CyclicVariations(C):Long-termfluctuationsnotofafixedperiod(oftenrelatedtoeconomic cycles).
4. IrregularVariations(I):Randomorunpredictableinfluences(e.g.,naturaldisasters).

TypesofTimeSeries Models:

1. AdditiveModel:

$Y_t=T_t +S_t + C_t +I_t$

2. MultiplicativeModel:

$Y_t=T_t\timesS_t\timesC_t \timesI_t$

ApplicationsofTimeSeriesAnalysis:

Forecastingsales,demand,andstockprices.

Weather prediction.

Economic and financial data analysis.

Monitoringandcontrollingindustrialprocesses.

MethodsUsedinTimeSeriesAnalysis:

Moving Averages

ExponentialSmoothing

ARIMA(Auto-RegressiveIntegratedMovingAverage) Decomposition

Techniques

BenefitsofTimeSeriesAnalysis:

Identifies patterns and

trends.Enhancesdata-drivendecision-

making. Enables accurate forecasting.

Challenges:

Requiresalargeamountofhistoricaldata.

Irregularitiescanaffect accuracy.

Complexmodelsmaybedifficulttointerpret.

MultipleChoiceQuestions:

1. WhatisTimeSeries Analysis?

Answer:A)Thestudyofhistoricaldatatoidentifypatternsovertime

2. WhichofthefollowingisNOTacomponentoftimeseries?

Answer: D) Hypothesis Testing

3. Inanadditivetimeseriesmodel,howarethecomponentscombined?

Answer: C) Addition

4. Whichofthefollowingisanexampleofamultiplicativetimeseries model?

Answer A) Y= T + S+ C + R

5. WhatdoestheFree-Hand Curve methodhelp in identifying?

Answer: B) Trend component

6. WhatistheSemi-Averagesmethodusedfor?

Answer:B) Tosplit data intotwo equal partsand find trends

7. IntheMovingAveragemethod,whathappenswhenthewindowsize increases?

Answer:A)Thetrendlinebecomessmoother

8. TheLeastSquaresMethodisprimarilyusedto:

Answer: B) Fit a trend line to historical data

9. Whichofthe followingisamajor applicationof timeseriesanalysis?

Answer: B) Forecasting future sales

A)

B) $Y=T \times S \times C \times R$

C) $Y=(T+S)\times C$

D) $Y=T-S-C-R$

5.

6.

7.

10. Why is Time Series Analysis important in forecasting?

Answer:A)Itidentifiestrendsandpatternsinhistoricaldata

11. WhatistheprimaryobjectiveofTrendAnalysisintimeseries?

Answer: A) Identifying long-term movement in data

12. WhichofthefollowingisNOTatrendanalysistechnique?

Answer:D) Monte Carlo Simulation

13. InwhichsectorisTimeSeriesAnalysiswidelyused?

Answer: D) All of the above

14. HowdoesTimeSeriesAnalysishelpinstockmarketpredictions?

Answer:B)Byidentifyinghistoricalpatternsandtrends

15. Whatisacommon challengeinTimeSeriesForecasting?

Answer: B) Handling missing data

MODULE5 DECISIONTHEORY

Structure

OBJECTIVES

- Explainthe concept,significance,andapplicationsof decisiontheoryin problem-solving.
- Understand andapplydecision-makingprinciples insituationswithknown outcomes.
- Develop decision trees to visualize and evaluate different decision-making scenarios.

UNIT5.1 INTRODUCTIONTODECISIONTHEORY

5.1DefiningDecisionTheoryandItsRelevance

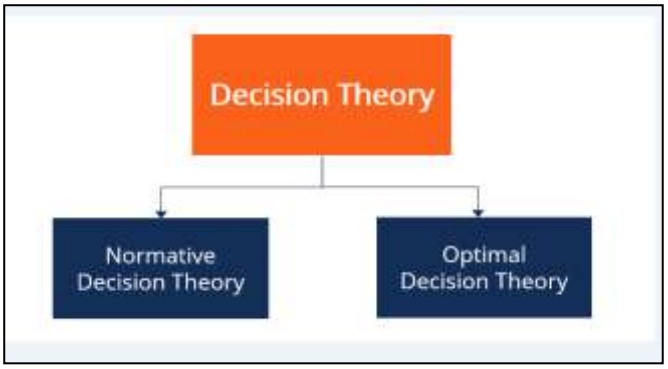

Figure5.1:DecisionTheory

At a basic level decision theory is the study of howhumans and organizations make choices. It's an interdisciplinary field, pulling from economics, psychology, statistics,philosophy, computer science and others. It attempts to understand the processes that underlie decision-making, which can includeboth descriptive (how people actually decide) and normative (how people

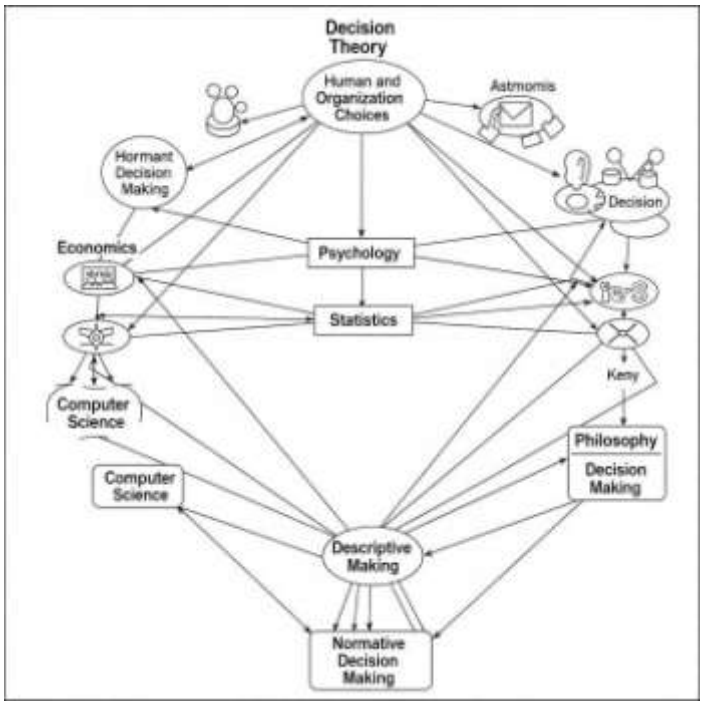

Figure5.2:RelevanceofDecisionTheory

should decide). We start with the innate complexity of choice. We are constantly faced with decisions in life, from the mundane and quotidian (what to eat for breakfast) to the profound and life-changing (career choices, investments, etc.). This article is suggestedby Decision theory. The basic idea is that decisions are made in face of uncertainty. We seldom know enough about the consequences of ourdecisions. You may not know everything you need to know to make predictions, or events may defy predictions, or other beings may choose actions that create uncertainty in the future, even with optimal knowledge. Decision theory uses elements likeprobabilities, utilities, andrisktoavoidbecomingmiredinuncertainty.Utilitiesareusedtoreflectthe expected value or satisfaction coming from particular scenarios, while probabilities show how likelythose scenarios are. Risk, in turn, represents the possibility ofdownside.

We need to distinguish descriptive from normative decision theory.Influencing behavioral decision making, the descriptive decision theory analyses how people do make decisions and commonly describe biases and irrationalities. For instance, the field of behavioral economics has revealed psychological phenomena, such as loss aversion, which is the fact that we feel more pain from losses than we derive pleasure from equivalent gains. In contrast, normative decision theory sets out how one should decide to achieve the most preferred outcomes, generally in a rational manner. This method is based on principles such as expected utility maximization, which takes into account thepotential results of each decision and balances them in accordance to their probabilities and utilities. Decision theory is not just an ivory tower exercise; it is more than alot of theorems stated without proof; it has strong real life implications. In business, it guides strategic planning, investment decisions and risk management. In medicine, it informs treatmentdecisionsand public health policies. In A.I. it forms the foundation forthe creation of intelligent agents capable of makes autonomous decisions. It can guide us to better decision-making in day-to-day scenarios.

KeyConceptstoIntroduceandElaborate:

- **Decision-MakingProcess:** The steps taken in decision-making process such as recognizing issues, gathering information, developing options, evaluating options, and making a decision and reviewing it.
- **Rationality**:Theideaofmakingeconomicdecisionsthatarealigned with your preferences and values. Embrace the imperfection ofrationality and understand bounded rationality.
- **UncertaintyandRisk:** Understanding the difference between uncertainty (when the outcomes are not known) and risk (when the probabilities of outcomes are known). How could we collaborate to identify types of risk (financial, operational, etc.)
- **Chance:**Theprobabilityofaneventuates happening.Introduce subjective probability andobjective probability.
- **Utility:** The subjective valueor satisfaction associated with an outcome. Paraphrase
- **ExpectedValueandExpectedUtility**: Teaching how to compute expected value (average outcome) and expected utility (average satisfaction).
- **DecisionTrees**: A visual decision-making process used to examine possible outcomes.
- **RealLifeExamples**: Give examples of how decision theory is used in business, finance,medicine, public policy, etc.
- **CognitiveBiases:** You can explain cognitive biases and how they impact yourdecision making. For example availability heuristic, anchoring bias, confirmation bias.

II. NavigatingtheUnknown:ToolsandFrameworksinDecisionTheory

Now that we have established foundational knowledge, we can go into some of the core tools and frameworks possessed by the field of decision theory that we can leverageto analyze and improve decision process. This is where youapply your theoretical learnings in practice. Perhaps, one of the simplest foundational tools is a decision matrix, where you line up potential choices, their possible outcomes, and the relative utilities or payoffs.Itfacilitatesanstructuredcomparisonofoptions'Acompany

deciding whether or not to launch a new product could, for instance, build a decision matrix that lays out the potential outcomes (success, moderatesuccess, failure) againstthe profits or losses for each scenario.

Bayesian decision theory updating and the test outcome. Sequential decision-making, where decisions are made based on an evolving body of information,is a key application for Bayesian beneficial especially when not all the information is available or well-defined. Examples would include like how a physician diagnosing a patient would use Bayesian reasoning to revise probability of a disease based on the symptoms that the patient presents with evidence to update probabilities. This method is a complementary powerful framework, incorporating previous beliefs and new and prisoner's dilemmahelp in understanding how individuals and organizations behave in strategic situations.interactions(e.g.,inauctions,negotiations,orcompetitivemarkets).
Concepts from game theory such as the Nash equilibriumcircumstances with multiple decision-makers that might have conflicted or aligned goals. It studies strategicJust as decision theory studies choice under uncertainty, game theory generalizes it to MCDA methods allow for prioritization and weighting ofthese objectives.involves the location of a factory, where you decide based on cost, environment, and nearness to customers, etc. Tools such as conflicting objectives. An instance Simultaneously, multi-criteria decision analysis (MCDA) addresses decision-making involving many criteria random sampling analysis studies the effect of varying inputs on outputs, whereas scenario planning investigates possible future scenarios and their consequences. Monte Carlo simulation modelsthe probability of different outcomes within business refers to the variability of future outcomes, and methods while quantifying and managing risk include sensitivity analysis, scenario planning, and Monte Carlo simulation, etc. For Looking Back — Sensitivity Analysis and Scenario Planning: Sensitivity in decision theory. Riskanalysis is a fundamental discipline.

KeyConceptstoIntroduceandElaborate:

- **DecisionMatrices:**Constructingandinterpretingdecisionmatrices

- **BayesianDecisionTheory:**Bayes'theorem,priorandposterior probabilities,belief updating.
- **Game Theory**: Nash equilibrium, prisoner's dilemma, strategic interactions
- **Multi-CriteriaDecisionAnalysis(MCDA):**Weightingofcriteria, scoring of alternatives, ranking approaches.
- **RiskAnalysis**:Sensitivity,scenario,Monte Carlo.
- **ValueofInformation**-Whatisthecostofobtainingfurther information?
- **InformationSystems**:Roleoftechnologyindecision support.
- **RealWorldExamples**:Instancesofaccuratetechniquesinrespective fields.

III. The Human Element: Behavioral Insights and EthicalConsiderations

and psychology that people frequently diverge from rationality, often as a result of cognitive biases, emotions and social influences.bases decisions on cold calculations and rational choices, but we must remember the humanity behind it all. It has been shown by behavioral economics Normative decision theory to combat them. on the first informationgiven), and loss aversion (the tendency to prefer avoiding losses over acquiring equivalent gains). By being aware of these biases, we can make better choices and design interventions The field of behavioral decision theory delves into the nature of these deviations, examining conceptual occurrences such as framing effects (the impact of how a decision is framed on the decision), anchoring bias (the tendency for an individual to rely too heavily.

Emotions drive many of the decisions we make. These feelings of fear, regret, excitement, can affect our choices; sometimes in even irrational ways. Decision theory asset us to understand and navigate these emotional ensnarement's. Social bonds also affect our choices. Meaning, we are affected by what other people think of us and do,as well as what others sayisrightorwrong.Decisiontheorycanhelpusmakesenseofhow

these social influences impact our decisions. Ethical considerations are paramount in decision-making. Any decision we make has the potential to affect either others or society greatly, and as such, we need to also therefore be wary of the ethics of ourdecisions. For example, the principlesand values that should dictate the choices we make can be framed using decision theory.

Also Important are Long-term vs. Short-term decisions. Most decisions are made on the basis of immediate gratification; however, the best decision may be the one that'll give the best outcome in the long run. Consideration of decisiontheoryallowsustonarrowdownapreferredlongtermaction.

UNIT5.2 DECISIONMAKINGUNDERCERTAINTY

5.2DecisionMakingUnderCertainty

Figure5.3:Decision-MakingunderConditionsofCertainty

Decision theory, a cornerstone of rational choice, provides a framework for understanding and analyzing how individuals and organizations make choices in the face of uncertainty. It is a deep dive into the ways that we assesschoices, consider the potential consequences, and finally make a decision that is consistent with our objectives. Basically, decision theory is the systematic study of decision-making, making choices that maximize the expected payoff and minimize the expected loss. It is a trans-disciplinary field that spans economics, psychology, statistics, philosophy, artificial intelligence, management, etc. The written word is the most efficient route for conveying a structured framework down to addressing complex matters, whether components of everyday living, enhancing strategic objectives or critical planning decisions. Both decision theory and HJB theory are not based on the ideathat decision making is arandom occurrence, but that we aredeliberatein our choices given our beliefs, preferences, and available data. It can help codify these influences so that we can construct models to predict and prescribe the best choices. Decision theory starts with some basics: Alternatives, outcomes, probabilities and utilities. Alternatives are the actions or decisions that the decision-maker can take or make, each with different outcomes. Outcomes are the results of these events and can be known outcomesorunknownoutcomes.Probabilitiesmeasurehowlikelyeach

outcome is to happen, capturing the decision-maker's beliefs about howthe world works.

Utilities are, instead, the subjective value or desirability of each outcome and therefore embody the preferences of the decision-maker. Decision-makingcan be roughly defined as the process of selecting the alternative that maximizes expected benefit, influenced by many factors. This entails calculating the weighted average of the utilities of all potential outcomes, with the weights corresponding to the probability of those possibilities. Decision theory distinguishes between decisions made under certainty, risk, and uncertainty. Decision-Making under Conditions of Certainty Decision-making under certainty pertains to scenarios where the outcomes of all alternatives are unequivocally known. While this is a rather basic situation, it serves as a foundation for more complex cases. Decision-making under risk refers to circumstances where the outcome is uncertain, but the probabilities of outcomes are known or can be estimated. This is the most basic situation covered in decisiontheory, where on the basis of expected utility a concrete conclusion is drawn. How to makesdecision under uncertain -- the situations where the results are not guaranteed, and the redundancies of these results are nothing but guess orestimation that mayormay not work. This becomes quite a task since expected utility calculations cannot be applied normally. Anumber of different approaches have been devised for this, including subjective probabilities, robust decision-making, andambiguity aversion. The first examines deductive normative approaches,while the second explores a variety of both normative and descriptive approaches. Normative decision theoryisanattempttotellrationalpeoplehowto makedecisionsaccordingto rules of logic and axioms. It sets up in ideal standard of decision making, thereby giving a yardstick to measure reality against. In contrast, descriptive decision theory fares an attempt to characterize the waypeople really make decisions, often admitting that human behavior is irrational. It integrates psychological elements, including cognitive biases and heuristics, to understand where such deviations arise. We are all taught the great key concepts of decision theory the principle of dominance, which is whenrationaldecision-makerswillalwayschoosetheoptionthatisbestinallstates

of the world, and so on. They can be used to describe very different preferencesofdecision-making:thetransitivityaxiomstates thatifa decision-maker prefers alternative A over alternative B and B over alternative C, then A must be preferred over C too, whereas the independence axiom states that preference between A and B must not change if a third alternative, not relevant to the choice, is included. These principles underlie rationalchoice theory, which posits that rationalbeings make consistent and coherent choices.

Decision trees and influence diagrams are two important tools used to help people understand decision problems and to analyze complex scenarios in decision theory. They can help us understand decision trees, which are graphical representations of the decision situation, explaining the sequence of decisions, chance events, and the resulting outcomes. They are especially useful for sequential decision problems where the outcome of one decision impacts future decisions. Other than Influence diagrams highlight the relationships among the variables, decisions, and outcomes, showing the dependencies and the flow of information, They are useful for the study of complex systems with multiple causes interacting. Game theory, a closely related field, generalizes decision theory to cases with multiple decision-makers with conflicting or aligned interests. It studies strategic interactions, where the payoff of one decision maker's action depends on the actions of others. Game theory explains competitive and cooperative behavior, with applications in fields from economics and political science to evolutionary biology.Behavioraldecisiontheorytakesinsightsfrompsychologytoexplain how cognitive. It recognizes that human decision making may not always be rational in the sense of expected utilitytheory. Such biases include framing effects -- when the way a problem is presented makes a difference to the choices made; anchoring effects -- as when the first piece of information received biases subsequent judgments; and availability heuristics, when information that comes to mind easily is overweighted.

These perceptual and cognitive biases can introduce or exacerbate systematic errorsinhowwemakeimportantdecisions;andso,theyareindangerof

being misunderstood or misapplied, highlighting the need for a thorough understandingof the sources and influences of these sugars. Decision theory also investigates the phenomena of risk aversion, where individuals prefer known risks over unknown risks, given the same expected value. Individual preferences, cultural factors, and situational context affect people's risk attitudes. Another area of focus is makingdecisions under ambiguity, where probabilities areunknown oruncertain. Ambiguous Aversion: Likely toavoid from options with unknown probabilities even when the expected utility is likely the same as options that have known probabilities. Robust decision making is concerned with making decisions under deep uncertainty; where the probabilities of the outcomes are poorly understood. It means creating strategies that will prove robust to a broad range of potential futures, insteadof aiming for accurate predictions. This obviously include new advancements and ideas from various fields. It offers an empowering platform for understanding and improving decision-making across a diverse scope of frameworks. These are the key to better decisions leading to improved outcomes, whether they be individual or organizational. Except that an instructional process that is prescriptive (top-down rules) does not allow for any abductive reasoning about shared context between multiple disciplines. We live in a time of uncertainty and complexity, and in such an environment, decision theory can serve as an important guide for how we approach the challenges, challenges will face us, and opportunities ahead, that we need rational and effective decision-making.

UNIT5.3 CONSTRUCTIONOFDECISIONTREES

5.3ConstructionofDecisionTrees

CorePrinciples: One powerful tool within business analytics is decisiontrees, a visual representation of the decision-making process, including potential outcomes, probabilities, and costs associatedwith each choice. They are constructed based on a recursive partitioning scheme, where the data is divided according to values of attributes that maximize information gain or minimize some measure of impurity. This begins from a root node that contains the entire data set and divides into internal nodes, which represent decision points around a specific attribute. The leaf nodes, which are the terminal points, represent the final outcomes, classified according to their respective categories or numerical values.

Forthisreason,theprimaryobjectiveisanaccuratemodeltopredictoutcomes in addition with interpretability so that the business could comprehend why decisions are made. One of the common algorithms for this construction is called decision trees, which utilize the chosen splitting criteria, such as Gini impurity or entropy for categorical variables and variance reduction for numerical variables, to choose what attributes at each node provides the most information. The basic idea behind pruning is an application of techniques, such as cost-complexity pruning, to reducethe complexity of the model and help ensure the model does not overfit to the train dataset, and does well on previously unseen data. The structure of the tree is constructed in an iterative manner, where all possible splits are evaluated, and the one that separatesthe outcomes best is selected, and this is done until some stopping condition is reached, such as minimum number of samples in a leaf node or maximum depth of the tree. This yields what we call the decision tree: a clear, hierarchical decision space allowing the organization to see the risk versus reward of each of the decisions.

DataProcessingandPreprocessing: Before any decision tree is made, it is essential to start from quality input data. Until the construction, the datashould be cleaned and preprocessed carefully. This includes dealing with issingvaluesthroughimputationorremoval,addressingoutlierswhichcan

skew the model and transforming variables when required. Featureengineering is key, where you can create new features based on the existing ones to improve predictive power. Data cleaning is a form of organizationin its purest form, ensuring consistency and accuracy while eliminatingduplicates and errors. Dependingon the data preprocessing that one applies, categorical variables are transformed into numeric values (like one-hot encoding or label encoding) to make it easier to work with them. Dimensionality reduction methods, such as feature selection, can help you focuson most relevant features to improvemodel's performance. Thedatasets can be divided into training and testing datasets, training datasets are used to construct the tree while testing datasets are used fortesting the constructed tree statistically. This split allows to coveron model generalization to unseen data and prevent overfitting. Youassess the distribution of classes within the dataset, and you may employ techniques like oversampling or under sampling to balance imbalanced datasets, ensuring that all classes are adequately represented in the model. If there are a number of different numerical features that are on very different scales, data normalization or scaling may be necessary since some of the splitting criteria can be sensitive to feature magnitude. The Preprocessing phase is an iterative one and might need tobe adjusted as you try to fit and test your model.

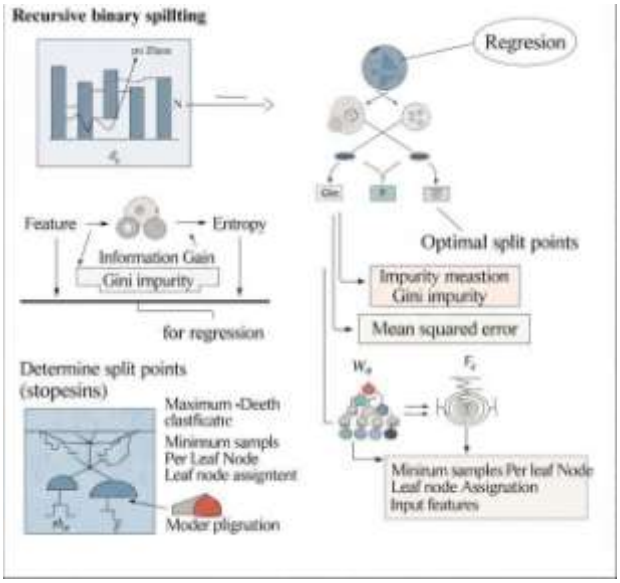

Figure5.4:DecisionSplitting

SelectionofSplittingCriteria: This is one of most important aspectsof a decision tree. For categorical type target variables; Gini impurity and entropy are widelyadopted. Gini impurity estimates the likelihood of mislabeling a randomly chosen item if it is randomly labeled according to the distributionof labels in the subset. Gini impurity: a lower value means a more homogeneous subset. Entropic, on the other hand, measures the unruliness or randomness in afraction. This change in entropy (less entropy value) when we split ona specific attribute is termed as information gain; informationgain is derived from entropy. The maximuminformation gain is chosen asthe splitting criterion. Variance reduction is commonly used for numerical target variables. This is based on the variance reduction when dividing the node according to a certain attribute.

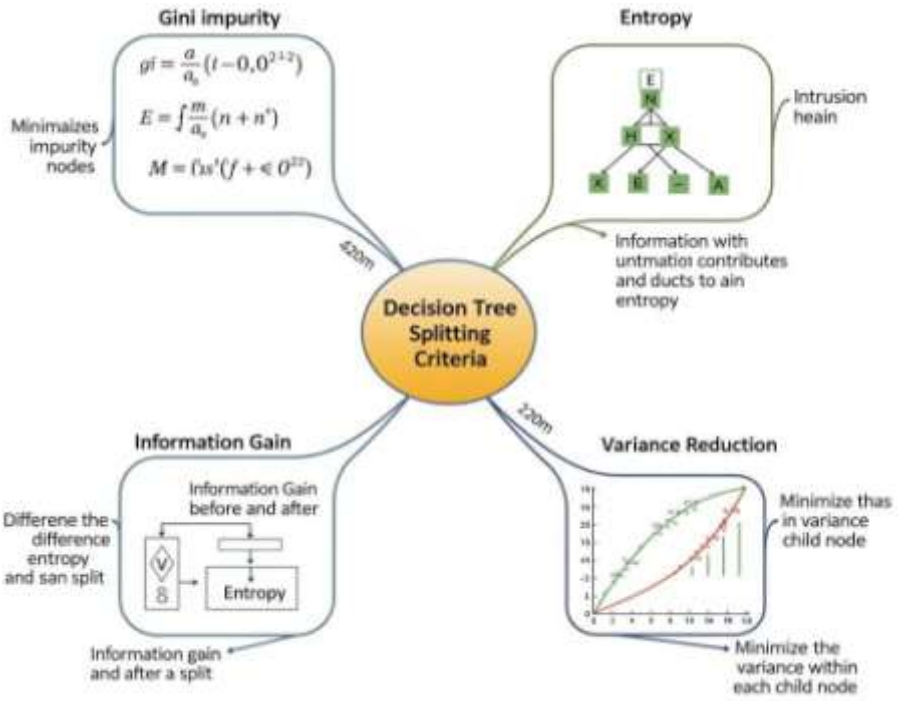

Figure5.5:DecisionTreeSplittingCriteria

Asindecisiontrees,theattributeresultinginmaximumreductionofvariance is chosen. Split can also be assessed using other criteria, for example chi- square test. The decision

between splitting criteria depends on the nature of dataset & particular aspectof analysis that one is interested in. Gini impurity, for example, is computationallyfasterthanentropy,andthereforewellsuitedforhandlingvery large datasets. Choosing the splitting criterion is a canonical step in the construction process that significantly affects the capability of the tree to accuratelyclassifyorpredictoutcomes.Foreachpotentialsplit,thecriteriaare calculated and the split that creates the maximum of the selected criterion is used.

TreeGrowthandPruning:the growth of the tree is similar to building the database recursive partitioning the data, and stops when a criterion is met. Some common criteria include minimum leaf node sample, maximum tree depth, maximum number of leaf nodes. Because decision trees are prone to overfitting the training data when pruning is not applied, this often results in weak model performance in terms of generalizing to unseen data. Techniques used to prune trees to avoid overfitting. One popular approach for pruning Decision Trees is cost-complexity pruning, also referred to as weakest link pruning. It introduces a complexity parameter - alpha - that governs thebalance between accuracy and size of the tree. The algorithmpruning beginsby cutting off the weakest link, that is, the node that provides the least amount of error reduction, and continues until the desired pruning level. The value of alpha is typically optimized through cross-validation to strike a balance betweenbias and variance. There are also other pruningmethods like Lower Error Pruning and Pessimistic Error Pruning which help trim the tree by removing those nodes that do not yield significant improvement. The second tree is simpler and even more interpretable than the first tree, thus it will be easier to understand and keepin mind while applying it to business decisions.

EvaluationandInterpretation: To provide the optimal information for the system, proper data running strategies should be in place. Evaluation Metrics are based on types of target variable. Common metrics for categoricalvariables include: accuracy,Precision, recall, F1-score, and area underreceiver operating characteristic curve (AUC). Accuracy quantifies the proportionofcorrectlyclassifiedcases.Precisionisratioofaccurately

anticipated positive instances to the total expected positive instances. Recall: The proportion of True Positives to Total Positives. Precision and recall are derived from F1-score, which represents harmonic mean of both metrics. AUC represents a comprehensive measure of performance across all potential classification criteria. When predicting numerical target variables, metrics such as mean squared error (MSE), root mean squared error (RMSE), or mean absolute error (MAE) can be employed to assess the model's predictive capability. This can aid in comprehending the outcomes by tracing the paths from the root node to each leaf node, which describes decision rules and distribution of outcomes across leaf nodes. You can evaluate featureimportance by checking how often a feature is used to split a node and how much impurity or variance is reduced due to afeature. Graphviz, for example, can be a straightforward way to visualize a tree, ascan the plot tree function from the scikit-learn. The resulting decision tree offers visual representation of the data, highlighting the factors that contribute to different outcomes. Train data until the decision tree is retrieving better results Evaluation & Interpretation: This stage confirms if the decision tree is accurate and usable,so thatit can provide useful insights about business decisions.

ApplicationsinBusiness: In marketing, they are oftenapplied for purposes like customer segmentation, target audience identification, and forecasting customer turnover. In the field of finance, they can be used for credit risk assessment,frauddetection,andportfolio management.Inbusiness operations, they can be employed for streamlining the supply chain, tracking inventory levels, and maintaining quality control. In HR, they can apply to employee performance evaluations, hiring, and training. They are also used in decision support systems where thealgorithm recommends a best decision foracomplicated decision-making scenario involving multiple attributes. In health care, they are used for diagnosis fordiseasediagnosis, treatment, and assessment of patient risk. Decision trees are interpretable whichmakes them very useful especially when you need to understand how decisions are made.

Decision Tree – Problems with Solutions

Problem 1: New Product Launch

A company is considering launching a new product.

Investment cost = ₹50,000

If successful → Profit = ₹1,20,000

If unsuccessful → Loss = ₹30,000

Probability of success = 0.6, failure = 0.4

Should the company launch the product?

Solution

1. Expected Monetary Value (EMV):

EMV = (Success x Profit) + (Failure x Loss)

EMV = (1,20,000 x 0.6) + (-30,000 x 0.4)

= 72,000 - 12,000 = 60,000

2. Since EMV > 0, launching gives a positive expected return.

Decision: The company should launch the product.

Problem 2: Market Research Decision

A company is unsure whether to invest in a new factory.

Investment = ₹5,00,000

If market is good → Profit = ₹10,00,000

If market is bad → Loss = ₹2,00,000

Probability (Good market) = 0.5, (Bad market) = 0.5

The company can first pay ₹50,000 for market research:

If research predicts "Good," probability of success increases to 0.8

If research predicts "Bad," probability of success decreases to 0.3

Should the company spend on market research?

Solution

1. Without Research:

EMV = (10,00,000 x 0.5) + (-2,00,000 x 0.5) = 5,00,000 - 1,00,000 = 4,00,000

2. With Research:

If prediction = Good (prob 0.5):

EMV = (10,00,000 x 0.8) + (-2,00,000 x 0.2) = 8,00,000 - 40,000 = 7,60,000

If prediction = Bad (prob 0.5):

EMV = (10,00,000 x 0.3) + (-2,00,000 x 0.7) = 3,00,000 - 1,40,000 = 1,60,000

Weighted EMV = (0.5 × 7,60,000) + (0.5 × 1,60,000) = 3,80,000 + 80,000 = 4,60,000

Minus research cost (50,000): 4,60,000 – 50,000 = 4,10,000

Decision: Since 4,10,000 > 4,00,000, the company should spend on market research.

Problem 3: Oil Drilling Decision

An oil company considers drilling a site.

Drilling cost = ₹1,00,000

If oil is found → Profit = ₹8,00,000

If dry → Loss = ₹1,00,000

Probability of finding oil = 0.3

The company may first conduct a geological survey at a cost of ₹20,000.

If survey positive → Probability of oil = 0.6

If survey negative → Probability of oil = 0.1

Should the company conduct the survey?

Solution

1. Without Survey:

EMV = (8,00,000 x 0.3) + (-1,00,000 x 0.7)

= 2,40,000 - 70,000 = 1,70,000

2. With Survey:

Survey Positive (prob = 0.3):

EMV = (8,00,000 x 0.6) + (-1,00,000 x 0.4) = 4,80,000 - 40,000 = 4,40,000

Survey Negative (prob = 0.7):

EMV = (8,00,000 x 0.1) + (-1,00,000 x 0.9) = 80,000 - 90,000 = -10,000

Weighted EMV = (0.3 × 4,40,000) + (0.7 × -10,000) = 1,32,000 - 7,000 = 1,25,000

Minus survey cost (20,000): 1,25,000 – 20,000 = 1,05,000

Decision: Since 1,70,000 > 1,05,000, the company should drill directly without survey.

SELFASSENMENTQUESTION

Multiple-ChoiceQuestions(MCQs)

1. WhatisDecisionTheoryprimarilyconcernedwith?

a. Probabilitycalculations
b. Makingoptimalchoicesunderuncertainty
c. Financial accounting
d. Manufacturingprocesses

2. Which of the following is NOT a type of decision-makingenvironment?

a. Decision-makingundercertainty
b. Decision-makingunder uncertainty
c. Decision-makingunder dictatorship
d. Decision-makingunderrisk

3. Whichdecision-makingconditioninvolvescompleteknowledgeofoutcomes?

a. Uncertainty
b. Risk
c. Certainty
d. Probability-baseddecision-making

4. Adecisiontreeismainlyusedfor:

a. Predictingfinancial losses
b. Evaluatingdecisionalternatives systematically
c. Conductingexperiments
d. Measuringeconomic growth

5. WhichcomponentisNOTpartofadecisiontree?

a. Decisionnodes
b. Probability nodes
c. Regressionequations
d. Outcomenodes

6. Whichofthefollowingrepresentsadecision-makingtechniquethatevaluatesmultiplepossibleoutcomes?

a. Decisiontree
b. Pie chart
c. Histogram
d. Timeseries analysis

7. Whatdoes"Maximin"strategyimplyindecision-making?

a. Choosingthealternative with thebestworst-casescenario
b. Maximizingprofitsatanycost
c. Ignoringuncertainties
d. Selectingrandom alternatives

8. Indecision-makingunderrisk,probabilitiesofoutcomesare:

a. Unknown
b. Known
c. Assumedto beequal
d. Ignored

9. WhatisthepurposeofExpectedMonetaryValue(EMV)indecision-making?

a. Todeterminethe worstpossible outcome
b. Tocalculatethemost likely profitor loss
c. Toeliminateuncertainty
d. To ignore risks

10. WhichofthefollowingisNOTacomponentofdecisiontheory?

a. Alternatives
b. Outcomes
c. psychologicalfactors
d. Payoffs

11. Whatisakeyadvantageofusingdecisiontrees?

a. Theyeliminaterisk
b. Theyprovideastructured andvisualrepresentationofchoices
c. Theyguaranteemaximum profit
d. Theyareonlyusefulforlarge businesses

12. Bayesiandecisiontheoryisbasedon:

a. Subjectiveopinions
b. Probabilityandstatistics
c. Randomselection
d. Maximizinglosses

13. TheHurwiczcriterionisusedwhendecision-makers:

a. Arehighlyrisk-averse
b. Areoptimisticor pessimisticabout outcomes
c. Havecomplete certainty
d. Usedecision treesonly

14. Whichtooliscommonlyusedfordecision-makingunderuncertainty?

a. Probability distributions
b. Regressionanalysis
c. SWOT analysis
d. Demand forecasting

15. Whichofthefollowingbestdescribesa"PayoffMatrix"?

a. Amathematicaltoolshowingpossibleoutcomesforeachdecision alternative
b. Agraphicalrepresentationoffinancial trends
c. Atypeofaccounting statement
d. Atime-seriesmodel

ShortQuestions

1. Whatisdecisiontheory?
2. Explaindecision-makingundercertainty.
3. Whataredecision treesin statistics?
4. Howdoesdecisiontheoryimpactbusinessdecisions?
5. Whataretheadvantagesofdecisiontrees?

LongQuestions:

1. Explaintheprocessof decision-makingin uncertainty.
2. Discusstheimportanceofdecisiontreesinbusiness strategy.

Glossary-D**ecisiontheory**:

Term	**Definition**
DecisionTheory	Thestudyofhowagentsmakechoicesamongalternatives,takinginto account uncertainty and values. It integrates probability, economics, philosophy, and statistics to guide rational decision-making
Agent	Theentity(e.g.,individual,group,orsystem)makingadecisionbasedon preferences and available information
Preference	Therankingororderinganagentassignstodifferentpossibleoutcomesor prospects, reflecting desirability or value
Utility	Anumericalmeasureofsatisfactionorvaluethatanagentassociateswith a specific outcome; central to quantifying preferences
ExpectedUtility	The weighted average of utilities across all possible outcomes, where weightsaretheirprobabilities;usedtodetermineoptimalchoicesunder uncertainty
Belief	Thesubjectiveprobabilityanagentassignstoaparticularoutcomeor event occurring
Rationality	Thepropertyofmakingchoicesthatmaximizeexpectedutility, given beliefsandpreferences;consideredoptimalor“best”withinthe theory
Normative DecisionTheory	The branch concerned with prescribing how agents *ought* to make decisionsrationallyandoptimally,oftenunderidealizedconditions
Descriptive DecisionTheory	Thestudyofhowagentsactuallymakedecisions,includingpsychological, social, and cognitive factors that impact real-world choices
Prescriptive DecisionTheory	Focusesonpracticalstrategies,tools,ormethodstohelpagentsimproveor guide decision-making in real settings
ChoiceUnder Uncertainty	Makingdecisionswhenoutcomesareunknownandmustbeconsideredin terms of probabilities
Bayes’Theorem	Amathematicalruleforupdatingprobabilitiesbasedonnewevidence, widely used in Bayesian decision theory
Axiom	Afundamentalassumptionorpremisewithinadecision-makingmodelthat is accepted without proof
Heuristic	Arule-of-thumborshortcut usedtosimplifycomplexdecisions, often

Term	Definition
	sacrificingthoroughnessforspeedandefficiency1.
Bounded Rationality	Theconceptthatagents'capacityforrationaldecision-makingislimitedby information, cognitive abilities, and time constraints1.

SUMMARY:DecisionTheory

Definition:

Decision theory is a framework for making logical choices in the face of uncertainty. It involvesidentifyingpossiblealternatives,evaluatingtheoutcomes,andselectingthebest option based on certain criteria.

Key Concepts:

1. Decision-MakingEnvironment:

Certainty:Outcomesofallalternativesareknown. Risk:

Probabilities of outcomes are known.

Uncertainty:Probabilitiesofoutcomesarenotknown.

2. ElementsofDecisionTheory:

Decisionmaker:Theindividualorgroupmakingthechoice.

Alternatives: Different courses of action available.

Statesofnature:Eventsnotunderthecontrol of thedecision maker.

Payoffs:Outcomesresultingfromeachcombinationofdecisionandstateofnature. Probabilities:

Likelihood of each state of nature (in case of risk).

TypesofDecision-MakingCriteria:

Under Certainty:

Choosethebestoption directlybasedonknown outcomes.

UnderRisk(withknown probabilities):

ExpectedMonetaryValue(EMV):Weightedaverageofoutcomes.

ExpectedOpportunityLoss(EOL):Costofnotchoosingthebestoptionforeachstate. Decision TreeAnalysis: Graphical method to evaluate decisions and outcomes.

UnderUncertainty(withoutknownprobabilities):

Maximin:Choosetheoptionwiththebestoftheworstoutcomes.

Maximax: Choose the option with the best possible outcome.

MinimaxRegret:Minimizethemaximum regret.

LaplaceCriterion:Assumesallstatesareequallylikely;chooseoptionwithbestaverage payoff.

Applications:

Business strategy

Economics

Operationsresearch

Risk management

Public policy

Multiple-ChoiceQuestions(MCQs)–DecisionTheory

1.What is DecisionTheory primarily concerned with?

Answerb.Makingoptimalchoicesunderuncertainty

2.WhichofthefollowingisNOTatypeofdecision-making environment?

Answerc. Decision-makingunder dictatorship

3.Whichdecision-makingconditioninvolvescompleteknowledgeofoutcomes?

Answer c. Certainty

4. Adecision treeismainly used for:

Answerb.Evaluatingdecision alternatives systematically

5.WhichcomponentisNOTpartofadecisiontree?

Answer c. Regression equations

6. Whichofthefollowingrepresentsadecision-makingtechniquethat evaluates multiple possible outcomes?

Answera. Decision tree

7. Whatdoes"Maximin"strategyimplyindecision-making?

Answera.Choosing thealternative withthe bestworst-case scenario

8. Indecision-makingunderrisk,probabilitiesofoutcomesare:

Answer b Known

9. Whatisthe purpose of ExpectedMonetary Value (EMV) in decision-making?

Answer b To calculate the most likely profit or loss

10. Whichof the following isNOT a component of decision theory?

Answerc. Psychological factors

11.Whatisa key advantage of using decision trees?

Answerb. They provide a structured and visual representation of choices✓

12. Bayesian decision theory is based on:

Answer bProbability and statistics

13. TheHurwiczcriterionisusedwhendecision-makers:

Answer b. Are optimistic or pessimistic about outcomes

14. Which tool is commonly used for decision-making under uncertainty?

Answer a. Probability distributions

15. Which of the following best describes a "Payoff Matrix"?

Answera.Amathematicaltoolshowingpossibleoutcomesforeach decision alternative

Reference

MODULEI:IntroductiontoStatistics

1. Hogg,R.V.,McKean,J.W.,&Craig,A.T.(2018).IntroductiontoMathematicalStatistics.Pearson Education.
2. Levin,R.I.,&Rubin,D.S.(2019).StatisticsforManagement.PearsonEducation.
3. Bowerman,B.L.,O'Connell,R.T.,&Murphree,E.S.(2016).BusinessStatisticsinPractice. McGraw-Hill Education.
4. Lind,D.A.,Marchal,W.G.,&Wathen,S.A.(2017).StatisticalTechniquesinBusinessand Economics. McGraw-Hill Education.
5. Black,K.(2019).BusinessStatistics:ForContemporaryDecisionMaking.Wiley.

MODULEII:ProbabilityandProbabilityDistribution

1. Walpole,R.E.,Myers,R.H.,Myers,S.L.,&Ye,K.(2016).ProbabilityandStatisticsforEngineers and Scientists. Pearson Education.
2. DeGroot,M.H.,&Schervish,M.J.(2014).ProbabilityandStatistics.Addison-Wesley.
3. Evans,M.J.,&Rosenthal,J.S.(2019).ProbabilityandStatistics:TheScienceofUncertainty.W.H. Freeman.
4. Hogg,R.V.,&Tanis,E.A.(2019).ProbabilityandStatisticalInference.PearsonEducation.
5. Mendenhall,W.,Beaver,R.J.,&Beaver,B.M.(2018).IntroductiontoProbabilityandStatistics. Cengage Learning.

MODULEIII:CorrelationAnalysis

1. Hair,J.F.,Black,W.C.,Babin,B.J.,&Anderson,R.E.(2018).MultivariateDataAnalysis.Cengage Learning.
2. Montgomery,D.C.,Peck,E.A.,&Vining,G.G.(2021).IntroductiontoLinearRegressionAnalysis. Wiley.
3. Kutner,M.H.,Nachtsheim,C.J.,Neter,J.,&Li,W.(2019).AppliedLinearStatisticalModels. McGraw-Hill Education.
4. Chatterjee,S.,&Hadi,A.S.(2015).RegressionAnalysisbyExample.Wiley.
5. Draper,N.R.,&Smith,H.(2014).AppliedRegressionAnalysis.Wiley.

MATS Center for Distance and Online Education, MATS University

MODULEIV:TimeSeries Analysis

1. Box,G.E.P.,Jenkins,G.M.,Reinsel,G.C.,&Ljung,G.M.(2015).TimeSeriesAnalysis:Forecasting and Control. Wiley.
2. Chatfield,C.,&Xing,H.(2019).TheAnalysisofTimeSeries:AnIntroductionwithR.Chapman and Hall/CRC.
3. Brockwell,P.J.,&Davis,R.A.(2016).IntroductiontoTimeSeriesandForecasting.Springer.
4. Wei,W.W.S.(2019).TimeSeriesAnalysis:UnivariateandMultivariateMethods.Addison-Wesley.
5. Hyndman,R.J.,&Athanasopoulos,G.(2018).Forecasting:PrinciplesandPractice.OTexts.

MODULEV:DecisionTheory

1. Peterson,M.(2017).AnIntroductiontoDecisionTheory.CambridgeUniversityPress.
2. Raiffa,H.,&Schlaifer,R.(2000).AppliedStatisticalDecisionTheory.Wiley.
3. Berger,J.O.(1985).StatisticalDecisionTheoryandBayesianAnalysis.Springer.
4. Clemen,R.T.,&Reilly,T.(2013).MakingHardDecisionswithDecisionTools.CengageLearning.
5. French,S.,Maule,J.,&Papamichail,N.(2009).DecisionBehaviour,AnalysisandSupport. Cambridge University Press.

MATS UNIVERSITY

MATS CENTRE FOR DISTANCE AND ONLINE EDUCATION

UNIVERSITY CAMPUS: Aarang Kharora Highway, Aarang, Raipur, CG, 493 441

RAIPUR CAMPUS: MATS Tower, Pandri, Raipur, CG, 492 002

T : 0771 4078994, 95, 96, 98 Toll Free ODL MODE : 81520 79999, 81520 29999

Website: www.matsodl.com