Chapter 3

Data Analysis Using R Programming Language

☞ Learning Objectives and Outcomes

In this unit student will be able to learn essentials of :

- ☞ Data analysis
- ☞ Terminology and concepts of data Analysis
- ☞ Introduction to statistics,
- ☞ Central tendencies and distributions,
- ☞ Defining Variance, Distribution properties and arithmetic,
- ☞ Samples and Central Limit Theorem (CLT),
- ☞ Basic machine learning algorithms,
- ☞ Linear regression,
- ☞ Support Vector Machine (SVM),
- ☞ Naive Bayes.

3.1 INTRODUCTION TO APPLIED STATISTICAL TECHNIQUES

Statistics is a branch of mathematics dealing with the collection, analysis, interpretation, presentation, and organization of data. In applying statistics to, e.g., **a scientific, industrial,** or **social problem,** it is conventional to begin with a statistical population or a **statistical model** process to be studied. Populations can be diverse topics such as "all people living in a country" or "every atom composing a crystal." Statistics deals with all aspects of **data including the planning of data collection in terms of the design of surveys and experiments.**

Applied statistics is a branch which covers natural processes and phenomena and provides us the knowledge for decision making as detailed in sketch shown in Figure 3.1.

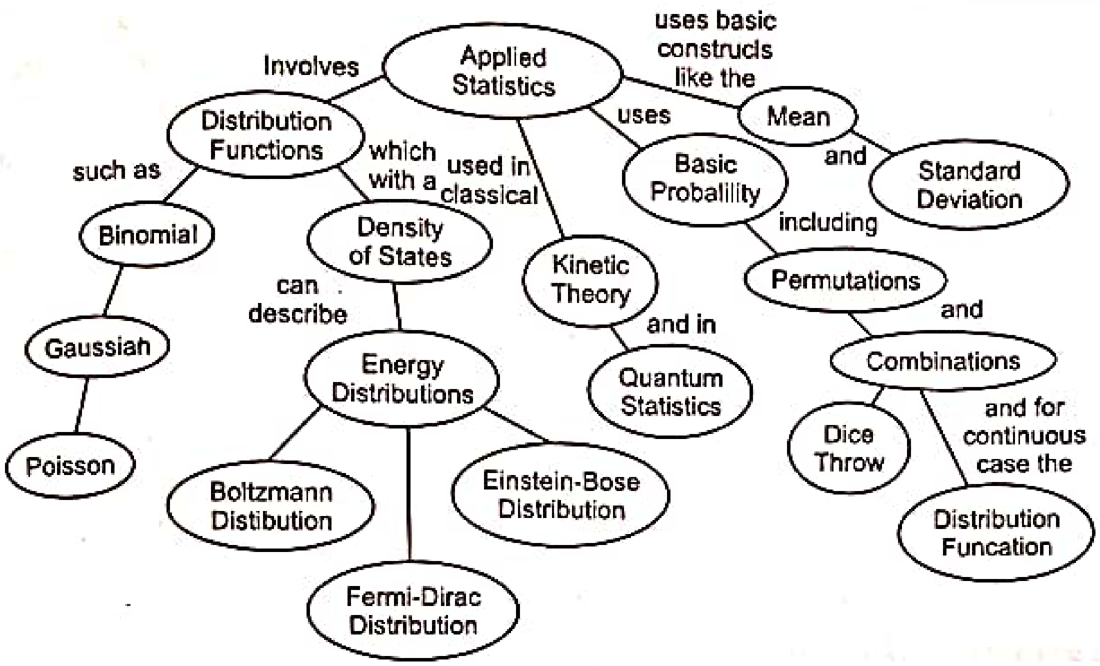


Fig. 3.1: Coverage of Applied Statistics.

3.1.1 Data Analysis

Analysis is the process of breaking a complex topic or substance into smaller parts in order to gain a better understanding of it. As such it refers to breaking a whole into its separate components for individual examination. Data analysis is a process for obtaining raw data and converting it into information useful for decision-making by users. Data is collected and analyzed to answer questions, test hypotheses or disprove theories. The term data analysis is sometimes used as a synonym for data modelling.

3.1.2 Data Analytics

Analytics is the discovery, interpretation, and communication of meaningful patterns in data. Especially valuable in areas rich with recorded information, analytics relies on the simultaneous application of statistics, computer programming and operations research to quantify performance. Data analytics uses analysed data to draw conclusions from it. It is the technique of getting automated insights into a given dataset. Data analytics can be described as a part of data science and it does find its applications in analyzing big data. The main focus of data analytics is inferencing some conclusion from the given data.

According to Merriam-Webster dictionary, *analysis* is the separation of a whole into its component parts, and *analytics* is the method of logical analysis. Marketers leverage both to drive all types of decisions, and each specific application supports the unique insight challenges inherent in dissecting customer behaviors.

Sole Difference

Analysis looks backwards over time, providing marketers with a historical view of what has happened. Typically, analytics look forward to model the future or predict a result.

Data analytics is a broader term and includes data analysis as necessary subcomponent. Analytics defines the science behind the analysis. The science means understanding the cognitive processes an analyst uses to understand problems and explore data in meaningful ways. Analytics also include data extract, transform, and load; specific tools, techniques, and methods; and how to successfully communicate results.

The growing maturity of the concept more starkly delineates the difference between big data and Business Intelligence:

1. Business Intelligence uses descriptive statistics with data with high information density to measure things, detect trends, etc..
2. Big data uses inductive statistics and concepts from nonlinear system identification to infer laws (regressions, nonlinear relationships, and causal effects) from large sets of data with low information density to reveal relationships and dependencies, or to perform predictions of outcomes and behaviors.

3.1.3 Statistical Data Analysis

The goal of statistical data analysis is to understand a complex, real-world phenomenon from partial and uncertain observations. The uncertainty in the data results in uncertainty in the knowledge we get about the phenomenon. A major goal of the theory is to quantify this uncertainty.

It is important to make the distinction between the mathematical theory underlying statistical data analysis, and the decisions made after conducting an analysis. The former is perfectly rigorous; perhaps surprisingly, mathematicians were able to build an exact mathematical framework to deal with uncertainty. Nevertheless, there is a subjective part in the way statistical analysis yields actual human decisions. Understanding the risk and the uncertainty behind statistical results is critical in the decision-making process.

In this unit, we will see the basic notions, principles, and theories behind statistical data analysis, covering in particular how to make decisions with a quantified risk. Of course, we will always show how to implement these methods with Python.

Exploratory methods allow us to get a preliminary look at a dataset through basic statistical aggregates and interactive visualization.

Statistical inference consists of getting information about an unknown process through partial and uncertain observations. In particular, estimation entails obtaining approximate quantities for the mathematical variables describing this process.

Decision theory allows us to make decisions about an unknown process from random observations, with a controlled risk.

Prediction consists of learning from data, that is, predicting the outcomes of a random process based on a limited number of observations.

Univariate and multivariate methods

In most cases, you can consider two dimensions in your data:

1. Observations (or samples, for machine learning people)
2. Variables (or features)

Typically, observations are independent realizations of the same random process. Each observation is made of one or several variables. Most of the time, variables are either numbers, or elements belonging to a finite set (that is, taking a finite number of values). The first step in an analysis is to understand what your observations and variables are.

Your problem is univariate if you have one variable. It is bivariate if you have two variables and multivariate if you have at least two variables. Univariate methods are typically simpler. That being said, univariate methods may be used on multivariate data, using one dimension at a time. Although interactions between variables cannot be explored in that case, it is often an interesting first approach.

When working with statistics, it's important to recognize the different types of data:

3.2 TYPES OF STATISTICAL DATA

Data are the actual pieces of information that you collect through your study.

When working with statistics, it's important to recognize the different types of data:

1. Numerical *(discrete and continuous),*
2. Categorical, and
3. Ordinal.

Discrete data are called *digital or Binary* data and Continuous data types are called *Analog data* in information technology

3.2.1 Numerical Data

These data have meaning as a measurement, such as a person's height, weight, IQ, or blood pressure; or they're a count, such as the number of stock shares a person owns, how many teeth a dog has, or how many pages you can read of your favourite book before you fall asleep. (Statisticians also call numerical data ***quantitative** data.*)

Numerical data can be further broken into two types: **discrete and continuous.**

Discrete data represent items that can be counted; they take on possible values that can be listed out. The list of possible values may be fixed (also called *finite*); or it may go from 0, 1, 2, on to infinity (making it *countably infinite*). For example, the number of heads in 100 coin flips takes on values from 0 through 100 (finite case), but the number of flips needed to get 100 heads takes on values from 100 (the fastest scenario) on up to infinity (if you never get to that 100th heads). Its possible values are listed as 100, 101, 102, 103, . . . (representing the countably infinite case).

Continuous data represent measurements; their possible values cannot be counted and can only be described using intervals on the real number line. For example, the exact amount of gas purchased at the pump for cars with 20-gallon tanks would be continuous data from 0 gallons to 20 gallons, represented by the interval [0, 20], inclusive. You might pump 8.40 litres, or 8.41, or 8.414863 litress, or any possible number from 0 to 20. In this way, continuous data can be thought of as being uncountably infinite. For ease of recordkeeping, statisticians usually pick some point in the number to round off. Another example would be that the lifetime of a C battery can be anywhere from 0 hours to an infinite number of hours (if it lasts forever), technically, with all possible values in between. Granted, you don't expect a battery to last more than a few hundred hours, but no one can put a cap on how long it can put on service.

3.2.2 Categorical Data or Qualitative Data

Categorical data represent characteristics such as a person's gender, marital status, hometown, or the types of movies they like. Categorical data can take on numerical values (such as "1" indicating male and "2" indicating female), but those numbers don't have mathematical meaning. You couldn't add them together, for example. (Other names for categorical data are *qualitative data*, or *Yes/No data.*)

3.2.3 Ordinal Data

Ordinal data mixes numerical and categorical data. The data fall into categories, but the numbers placed on the categories have meaning. For example, rating a restaurant on a scale from 0 (lowest) to 4 (highest) stars gives ordinal data. Ordinal data are often treated as categorical, where the groups are ordered when graphs and charts are made. However, unlike categorical data, the numbers do have mathematical meaning. For example, if you survey 100 people and ask them to rate a restaurant on a scale from 0 to 4, taking the average of the 100 responses will have meaning. This would not be the case with categorical data.

3.2.4 Constant and Variable Data

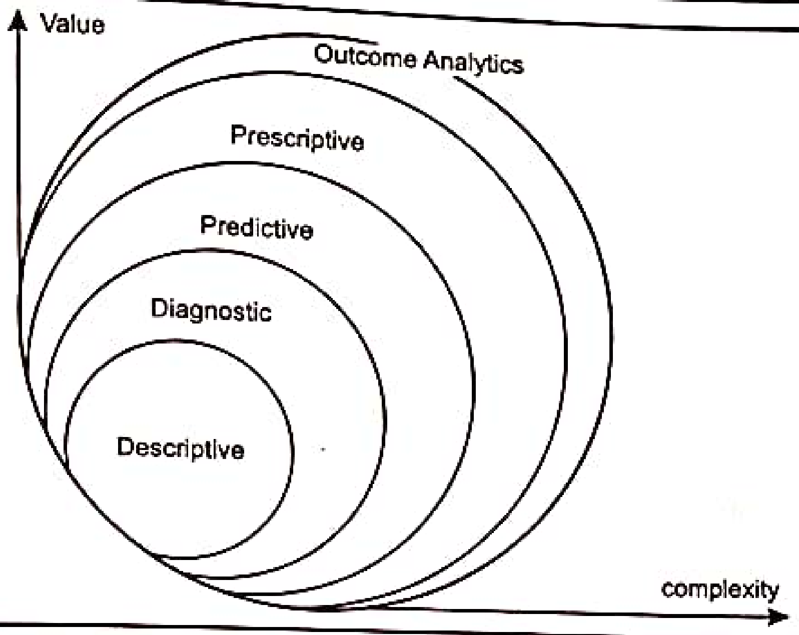
Variables conforming only to nominal or ordinal measurements cannot be reasonably measured numerically, sometimes they are grouped together as categorical variables, whereas ratio and interval measurements are grouped together as quantitative variables, which can be either discrete or continuous, due to their numerical nature. Such distinctions can often be loosely correlated with data type in computer science, in that dichotomous categorical variables may be represented with the Boolean data type, polytomous categorical variables with arbitrarily assigned integers in the integral data type, and continuous variables with the real data type involving floating point computation. But the mapping of computer science data types to statistical data types depends on which categorization of the latter is being implemented.

3.3 TYPES OF BIG DATA ANALYTICS

It is useful to distinguish between different kinds of analytics because the differences have implications for the technologies and architectures used for big data analytics.

Some types of analytics are better performed on some platforms than on others. Fig. 3.2 provides with the historical evolution of big data Analytics.

1.



Descriptive: *What's happening in my business?*

- Comprehensive, accurate and live data.
- Effective visualisation

Diagnostic: *Why is it happening?*

- Ability to drill down to the root-cause
- Ability to isolate all confounding information

Predictive: *What's likely to happen?*

- Business strategies have remained fairly consistent over time
- Historical patterns being used to predict specific outcomes using algorithms
- Decisions are automated using algorithms and technology

Prescriptive: *What do I need to do?*

- Recommended actions and strategies based on champion/challenger testing strategy outcomes.
- Applying advanced analytical techniques to make specific recommendations.

Fig. 3.2: Evolution Of Analytics Within the Data Science

In general we find *five types* of big data analytics in most current (Year 2018) literature[1] :

3.3.1 Descriptive: What is happening?

The first stage of business analytics is descriptive analytics, which still accounts for the majority of all business analytics today. This is the most common of all forms. In business it provides the analyst a view of key metrics and measures within the business.

Descriptive analytics looks at past performance and understands that performance by mining historical data to look for the reasons behind past success or failure. Most management reporting – such as sales, marketing, operations, and finance – uses this type of post-mortem analysis.

Descriptive analytics, such as reporting/OLAP, dashboards/scorecards, and data visualization, have been widely used for some time, and are the core applications of traditional BI. This technique is the most time-intensive and often produces the least value; however, it is useful for uncovering patterns within a certain segment of customers. Descriptive analytics provide insight into what has happened historically and will provide you with trends to dig into in more detail. It gains insight from historical data with reporting, scorecards, clustering etc.

An examples of this could be a monthly profit and loss statement. Similarly, an analyst could have data on a large population of customers. Understanding demographic information on their customers (e.g. 30% of our customers are self-employed) would be categorised as "descriptive analytics". Utilising effective visualisation tools enhances the message of descriptive analytics.One trend, however, is to include the findings from predictive analytics, such as forecasts of future sales, on dashboards/scorecards.

Descriptive Statistics Application

In applying statistics to a problem, it is common practice to start with a population or process to be studied. Populations can be diverse topics such as "all persons living in a country" or "every atom composing a crystal".

Ideally, statisticians compile data about the entire population (an operation called census). This may be organized by governmental statistical institutes. Descriptive statistics can be used to summarize the population data. Numerical descriptors include mean and standard deviation for continuous data types (like income), while frequency and percentage are more useful in terms of describing categorical data (like race).

When a census is not feasible, a chosen subset of the population called a **sample** is studied. Once a sample that is representative of the population is determined, data is collected for the sample members in an observational or experimental setting

Again, descriptive statistics can be used to summarize the sample data. However, the drawing of the sample has been subject to an element of **randomness**, hence the established numerical descriptors from the sample are also due to uncertainty.

Examples of descriptive analytics also include summary statistics, clustering and association rules used in market basket analysis.

Key points:

- Backward looking
- Focused on descriptions and comparisons
- Pattern detection and descriptions
- MECE (mutually exclusive and collectively exhaustive) categorization
- Category development based on similarities and differences (segmentation)

3.3.2 Diagnostic Inferential Analytics : Why is it happening?

This is the next step of complexity in data analytics is descriptive analytics. On assessment of the descriptive data, diagnostic analytical tools will empower an analyst to drill down and in so doing isolate the root-cause of a problem.

Well-designed business information (BI) dashboards incorporating reading of time-series data (i.e. data over multiple successive points in time) and featuring filters and drill down capability allow for such analysis.

Data scientists turn to this technique when trying to determine why something happened. It is useful when researching leading churn indicators and usage trends amongst your most loyal customers. Examples of diagnostic analytics include churn reason analysis and customer health score analysis.

Key points:

- Backward looking
- Focused on causal relationships and sequences
- Relative ranking of dimensions/variable based on inferred explanatory power)
- Target/dependent variable with independent variables/dimensions
- Includes both frequentist and Bayesian causal inferential analyses

3.3.3 Predictive Analytics : What is likely to happen?

Predictive analytics is all about forecasting. Whether it's the likelihood of an event happening in future, forecasting a quantifiable amount or estimating a point in time at which something might happen – these are all done through predictive models.

Predictive models typically utilise a variety of variable data to make the prediction. The variability of the component data will have a relationship with what it is likely to predict (e.g. the older a person, the more susceptible they are to a heart-attack – we would say that age has a linear correlation with heart-attack risk). These data are then compiled together into a score or prediction.

In a world of great uncertainty, being able to predict allows one to make better decisions. Predictive models are some of the most important utilised across a number of fields.

In this forecasting method , historical data is combined with rules, algorithms, and occasionally external data to determine the probable future outcome of an event or the likelihood of a situation occurring.

The most commonly used technique; predictive analytics use models to forecast what might happen in specific scenarios. Examples of predictive analytics include next best offers, churn risk and renewal risk analysis.

Key points:

- Forward looking
- Focused on non-discrete predictions of future states, relationship, and patterns
- Description of prediction result set probability distributions and likelihoods
- Model application
- Non-discrete forecasting (forecasts communicated in probability distributions)

3.3.4 Prescriptive Analytics : What do I need to do?

The next step up in terms of value and complexity is the prescriptive model. The prescriptive model utilises an understanding of what has happened, why it has happened and a variety of "what-might-happen" analysis to help the user determine the best course of action to take. Prescriptive analysis is typically not just with one individual action, but is in fact a host of other actions.

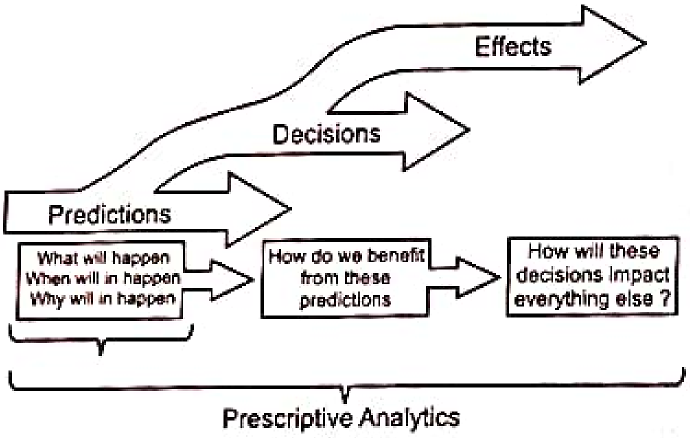


Fig. 3.3: Development of Prescriptive Analytics

Prescriptive Analytics extends beyond predictive analytics by specifying both the actions necessary to achieve predicted outcomes, and the interrelated effects of each decision

It is most valuable and most underused big data analytics technique, prescriptive analytics gives you a laser-like focus to answer a specific question. It helps to determine the best solution among a variety of choices, given the known parameters and suggests options for how to take advantage of a future opportunity or mitigate a future risk. It can also illustrate the implications of each decision to improve decision-making. Examples of prescriptive analytics for customer retention include next best action and next best offer analysis.

Key points:

- Forward looking
- Focused on optimal decisions for future situations
- Simple rules to complex models that are applied on an automated or programmatic basis
- Discrete prediction of individual data set members based on similarities and differences
- Optimization and decision rules for future events

A good example of this is a traffic application helping you choose the best route home and taking into account the distance of each route, the speed at which one can travel on each road and, crucially, the current traffic constraints.

Another example might be producing an exam time-table such that no students have clashing schedules.

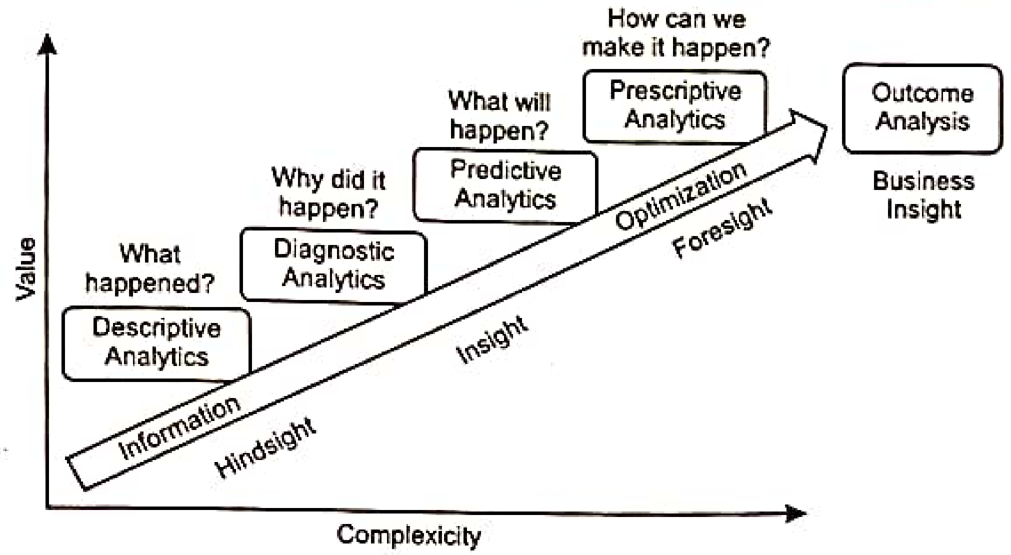


Fig. 3.4: Optmization of business data to Create Business Insights

3.3.5 Outcome Analytics

Also referred to as consumption analytics, this technique provides insight into customer behavior that drives specific outcomes. This analysis is meant to help you know your customers better and learn how they are interacting with your products and services. Creating increased value for customers is, today, a business imperative, as is the need to contain costs, and adopt new technologies to meet consumer expectations. Here's where Business Process Management (BPM) players can step in to develop customized analytics solutions that deliver targeted business outcomes. These outcomes can be varied – optimized marketing efforts, increased cash flow, accurate prediction of customer behavior and improved customer engagement, among others.

Key points:

1. Backward looking, Real-time and Forward looking
2. Focused on consumption patterns and associated business outcomes
3. Description of usage thresholds
4. Model application

3.4 COLLECTING DATA FOR SAMPLING AND DISTRIBUTION

Accurate data collection is essential to many business processes,[2][3][4] to the enforcement of many government regulations,[5] and to maintaining the integrity of scientific research.[6]

The problem with collecting data is that you do not generally know what distribution the data follows. So you have a sample, but no distribution to help figure it out. The true distribution is generally not knowable, but you could often find something workable distribution as per central limit theorem. It implies under fairly easy to satisfy conditions that some of the summary statistics you'd calculate from your sample do have a known distribution even if you do not know the distribution of your sample.

3.5 PROBABILITY

Probability is the *measure* of the likelihood that an *event* will occur. Probability is quantified as a number between 0 and 1, where, loosely speaking, 0 indicates **impossibility** and 1 indicates **certainty**. The higher the probability of an event, the more likely it is that the event will occur. A simple example is the tossing of a fair (unbiased) coin. Since the coin is fair, the two outcomes ("heads" and "tails") are both equally probable; the probability of "heads" equals the probability of "tails"; and since no other outcomes are possible, the probability of either "heads" or "tails" is 1/2 (which could also be written as 0.5 or 50%).

When dealing with experiments that are *random* and *well-defined* in a purely theoretical setting (like tossing a fair coin), probabilities can be numerically described by the number of desired outcomes divided by the total number of all outcomes. For example, tossing a fair coin twice will yield "head-head", "head-tail", "tail-head", and "tail-tail" outcomes. The probability of getting an outcome of "head-head" is 1 out of 4 outcomes or 1/4 or 0.25 (or 25%).

In probability theory, a probability distribution is a mathematical function that, stated in simple terms, can be thought of as providing the probabilities of occurrence of different possible outcomes in an experiment. For instance, if the random variable X is used to denote the outcome of a coin toss ("the experiment"), then the probability distribution of X would take the value 0.5 for X = heads, and 0.5 for X = tails (assuming the coin is fair).Normally in casinos, the coins are not fair and aredesired to bring the outcome result according to whim of the skilled but wicket person tossing the coin.

Throwing Dice

When a single die is thrown, there are six possible outcomes: **1, 2, 3, 4, 5, 3.** The probability of any one of them is 1/6

$$\text{Probability of an event happening} = \frac{\text{Number of ways it can happen}}{\text{Total number of outcomes}}$$

Example 3.1 (*a*): The chances of rolling a "4" with a die

Number of ways it can happen: 1 (there is only 1 face with a "4" on it)

Total number of outcomes: 6 (there are 6 faces altogether)

$$\text{So the probability} = \frac{1}{6}$$

Probability is Just a Guide

Probability does not tell us exactly what will happen, it is just a guide

Example 3.1 (*b*): Toss a coin 100 times, how many Heads will come up?

Probability says that heads have a ½ chance, so we can expect 50 Heads.

But when we actually try it we might get 48 heads, or 55 heads ... or anything really, but in most cases it will be a number near 50.

3.5 FREQUENCY DISTRIBUTION

A frequency distribution is a table that displays the frequency of various outcomes in a sample.. A frequency distribution is a table or graph that displays the frequency of various outcomes in a sample. Each entry in the table contains the frequency or count of the occurrences of values within a particular group or interval, and in this way, the table summarizes the distribution of values in the sample.

Frequency shows how often an event happens in any phenomena.

Example 3.2: Amit played football on

- Saturday Morning,
- Saturday Afternoon
- Thursday Afternoon

The frequency was 2 on Saturday, 1 on Thursday and for the whole week.

Frequency Distribution Table

By counting frequencies we can make a Frequency Distribution Table.

Example 3.3: Goals

Shyam's team has scored the following numbers of goals in recent games:

2, 3, 1, 2, 1, 3, 2, 3, 4, 5, 4, 2, 2, 3

Shyam put the numbers in order, then added up:

- how often 1 occurs (2 times),
- how often 2 occurs (5 times), ano so on.

<table>
<tr><th colspan="2">Scores:
1, 1, 2, 2, 2, 2, 2, 3, 3, 3, 3, 4, 4, 5</th></tr>
<tr><th>Score</th><th>Frequency</th></tr>
<tr><td>1</td><td>2</td></tr>
<tr><td>2</td><td>5</td></tr>
<tr><td>3</td><td>4</td></tr>
<tr><td>4</td><td>2</td></tr>
<tr><td>5</td><td>1</td></tr>
</table>

From the table (which is also called as *contingency table*), we can see interesting things such as getting 2 goals happens most often only once did they get 5 goals.

Bivariate (two way)

Bivariate (two way) frequency distributions are often presented as (two-way) contingency tables:

♦	Dance ♦	Sports ♦	TV ♦	Total ♦
Men	2	10	8	20
Women	16	6	8	30
Total	18	16	16	50

Fig. 3.4: A Two Way (Bivariate) Contingency Table

Managing and operating on frequency tabulated data is much simpler than operation on raw data. There are simple algorithms to calculate median, mean, standard deviation etc. from these tables.

Statistical hypothesis testing is founded on the assessment of differences and similarities between frequency distributions.

3.6 POPULATION AND PARAMETERS

In statistics, a population is a set of similar items or events which is of interest for some question or experiment.[7] A statistical population can be a group of actually existing objects (e.g. the set of all stars within the Milky Way galaxy) or a hypothetical and potentially infinite group of objects conceived as a generalization from experience (e.g. the set of all possible hands in a game of poker).[8]

We have seen that descriptive statistics provide information about our immediate group of data. For example, we could calculate the mean and standard deviation of the exam marks for the 100 students and this could provide valuable information about this group of 100 students. Any group of data like this, which includes all the data you are interested in, is called a **population.** A population can be small or large, as long as it includes all the data you are interested in. For example, if you were only interested in the exam marks of 100 students, the 100 students would represent your population. Descriptive statistics are applied to populations, and the properties of populations, like the mean or standard deviation, are called **parameters** as they represent the whole population (i.e., everybody you are interested in).

Statistical sampling is used quite often in statistics. sampling is the selection of a subset (a statistical sample) of individuals from within a statistical population to estimate characteristics of the whole population. Two advantages of sampling are that the cost is lower and data collection is faster than measuring the entire population. In this process we aim to determine something about a population. Since populations are typically large in size, we form a statistical sample by selecting a subset of the population that is of a predetermined size. By studying the sample we can use inferential statistics to determine something about the population.

A statistical sample of size *n* involves a single group of *n* individuals or subjects that have been randomly chosen from the population.

The geometric mean is relevant on those sets of data that are products or exponential in nature. This includes a variety of branches of natural sciences and social sciences.

Usages

In social sciences, we frequently encounter this in a number of ways. For example, the human population growth is expressed as a percentage, and thus when population growth needs to be averaged, it is the geometric mean that is most relevant.

In surveys and studies too, the geometric mean becomes relevant. For example, if a survey found that over the years, the economic status of a poor neighborhood is getting better, they need to quote the geometric mean of the development, averaged over the years in which the survey was conducted. The arithmetic mean will not make sense in this case either.

In economics, we see the percentage growth in interest accumulation. Thus if you are starting out with a sum of money that is compounded for interest, then the mean that you should look for is the geometric mean. Many such financial instruments like bonds yield a fixed percentage return, and while quoting their "average" return, it is the geometric mean that should be quoted.

3.10 PROBLEMS OF ESTIMATION : POPULATION OR SAMPLE

A common problem in statistics is to obtain information about the mean, μ, of a population.

For example, we might want to know

1. the mean age of people in the civilian labour force,
2. the mean cost of a wedding,
3. the mean gas mileage of a new-model car, or
4. the mean starting salary of liberal-arts graduates.

If the population is small, we can ordinarily determine μ exactly by first taking a census and then computing μ from the population data. If the population is large, however, as it often is in practice, taking a census is generally impractical, extremely expensive, or impossible.

Nonetheless, we can usually obtain sufficiently accurate information about μ by taking a sample from the population.

3.11 NORMAL DISTRIBUTION CURVE

In statistics, the theoretical curve that shows how often an experiment will produce a particular result. The curve is symmetrical and bell shaped, showing that trials will usually give a result near the average, but will occasionally deviate by large amounts. Normal distribution occurs very frequently in statistics, economics, natural and social