30% of the items they intend to buy. Forgetfulness can have significant cost impact for the online grocery stores. The customers may buy the forgotten items from a nearby store where they live, but since she/he is already in the store she/he may buy more items resulting in reduction in basket size in the future for online grocery stores such as bigbasket.com. Alternatively, the customer may place another order for forgotten items, but this time, the size of the basket is likely to be small and results in unnecessary logistics cost. Thus, the ability to predict the items that a customer may have forgotten to order can have a significant impact on the profits of online grocers such as bigbasket.com.

Another problem that online grocery customers face while ordering the items is the time taken to place an order. Unlike customers of Amazon or Flipkart, online grocery customers order several items each time; the number of items in an order may cross 100. Searching for all the items that a customer would like to order is a time-consuming exercise, especially when they order using smart phones. Thus, bigbasket created a 'smart basket' which is a basket consisting of items that a customer is likely to buy (recommended basket) reducing the time required to place the order (Abraham *et al.*, 2016).

The above two examples (Target's pregnancy test and 'did you forget' and smart basket feature at bigbasket.com) manifest the importance of business context in business analytics, that is, the ability to ask the right questions is an important success criteria for analytics projects.

1.3.2 | Technology

To find out whether a customer is pregnant or to find out whether a customer has forgotten to place an order for an item, we need data. In both the cases, the point of sale data has to be captured consisting of past purchases made by the customer. Information Technology (IT) is used for data capture, data storage, data preparation, data analysis, and data share. Today most data are unstructured data; data that is not in the form of a matrix (rows and columns) is called unstructured data. Images, texts, voice, video, click stream are few examples of unstructured data. To analyse data, one may need to use software such as R, Python, SAS, SPSS, Tableau, etc. Technology is also required to deploy the solution; for example, in the case of Target, technology can be used to personalize coupons that can be sent to individual customers. An important output of analytics is automation of actionable items derived from analytical models; automation of actionable items is usually achieved using IT.

1.3.3 | Data Science

Data Science is the most important component of analytics, it consists of statistical and operations research techniques, machine learning and deep learning algorithms. Given a problem, the objective of the data science component of analytics is to identify the most appropriate statistical model/machine learning algorithm that can be used. For example, Target's pregnancy prediction is a *classification problem* in which customers (or entities) are classified into different groups. In the case of pregnancy test, the classes are:

1. Pregnant
2. Not pregnant

There are several techniques available for solving classification problems such as logistic regression, classification trees, random forest, adaptive boosting, neural networks, and so on. The objective of the data science component is to identify the technique that is best based on a measure of accuracy. Usually, several models are developed for solving the problem using different techniques and a few models may be chosen for deployment of the solution.

Business analytics can be grouped into three types: **descriptive analytics, predictive analytics,** and **prescriptive analytics**. In the following sections, we shall discuss the three types of analytics in detail.

1.4 | DESCRIPTIVE ANALYTICS

"If the statistics are boring, then you've got the wrong numbers".

— Edward R. Tufte

Descriptive analytics is the simplest form of analytics that mainly uses simple descriptive statistics, data visualization techniques, and business related queries to understand past data. One of the primary objectives of descriptive analytics is innovative ways of data summarization. Descriptive analytics is used for understanding the trends in past data which can be useful for generating insights. Figure 1.5 shows visualization of relationship break-ups reported in Facebook.

It is clear from Figure 1.5 that spike in breakups occurred during spring break and in December before Christmas. There could be many reasons for increase in breakups during December (we hope it is

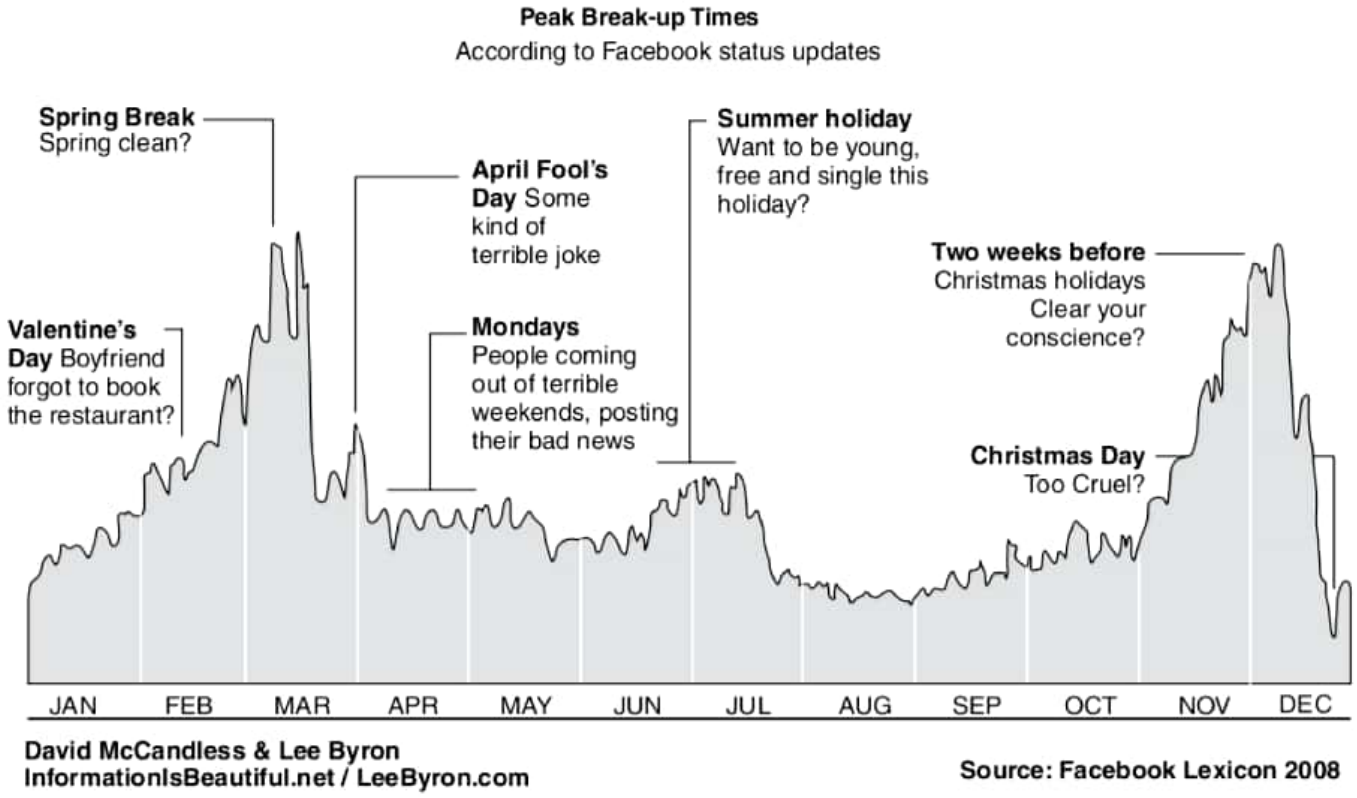


FIGURE 1.5 Peak breakup times according to Facebook status update. *Source*: David McCandless and Lee Bryon.

not a New Year resolution that they would like to change the partner). Many believe that since December is a holiday season, couples get a lot of time to talk to each other, probably that is where the problem starts. However, descriptive analytics is not about why a pattern exists, but about what the pattern means for a business. The fact that there is a significant increase in breakups during December we can deduce the following insights (or possibilities):

1. There will be more traffic to online dating sites during December/January.
2. There will be greater demand for relationship counsellors and lawyers.
3. There will be greater demand for housing and the housing prices are likely to increase in December/January.
4. There will be greater demand for household items.
5. People would like to forget the past, so they might change the brand of beer they drink.

Descriptive analytics using visualization identifies trends in the data and connects the dots to gain insights about associated businesses. In addition to visualization, descriptive analytics uses descriptive statistics and queries to gain insights from the data. The following are a few examples of insights obtained using descriptive analytics reported in literature:

1. Most shoppers turn towards the right side when they enter a retail store (Underhill, 2009, pages 77–79). Retailers keep products with higher profit on the right side of the store since most people turn right.
2. Married men who kiss their wife before going to work live longer, earn more and get into less number of accidents as compared to those who do not (Foer, 2006).
3. Correlated with Facebook relationship breakups, divorces spike in January. According to Caroline Kent (2015), January 3 is nicknamed 'divorce day'.
4. Men are more reluctant to use coupons as compared to women (Hu and Jasper, 2004). While sending coupons, retailers should target female shoppers as they are more likely to use coupons.

Trends obtained through descriptive analytics can be used to derive actionable items. For example, when Hurricane Charley struck the U.S. in 2004, Linda M. Dillman, Walmart's Chief Information Officer, wanted to understand the purchasing behaviour of their customers (Hays, 2004). Using data mining techniques, Walmart found that the demand for strawberry pop-tarts went up over 7 times during the hurricane compared to their normal sales rate; the pre-hurricane top-selling item was found to be beer. These insights were used by Walmart when the next hurricane — Hurricane Frances — hit the U.S. in August–September 2004; most of the items predicted by Walmart sold quickly. Although the high pre-hurricane demand for beer can be intuitively predicted, the demand for strawberry pop-tarts was a complete surprise.

Data visualization to understand hidden facts/trends has been in use for several centuries. Dr. John Snow's cholera spot map is an interesting application of data visualization. Cholera claimed millions of lives across the world during the 19th century. Medical practitioners did not have a clear understanding of the causes of the disease (Cameron and Jones, 1983). Between 1845 and 1856, over 700 articles were

published in London on the causes of cholera and how the epidemic could be prevented (Snow, 2002). However, none of them offered any real solution. The breakthrough in cholera epidemiology was made by Dr. John Snow based on the data of cholera outbreak in central London in 1854. Between 31 August and 10 September 1854, over 500 people died of cholera in London. John Snow marked the locations of the homes of those who had died during the epidemic and prepared a spot map (Figure 1.6)[2]. The spot map revealed that the highest number of deaths occurred in the Golden Square area (Snow, 1999). The most striking difference between this area and the other districts of London was the source of water (Brody *et al.*, 2000); thus, Snow established that water contamination was the main source of cholera.

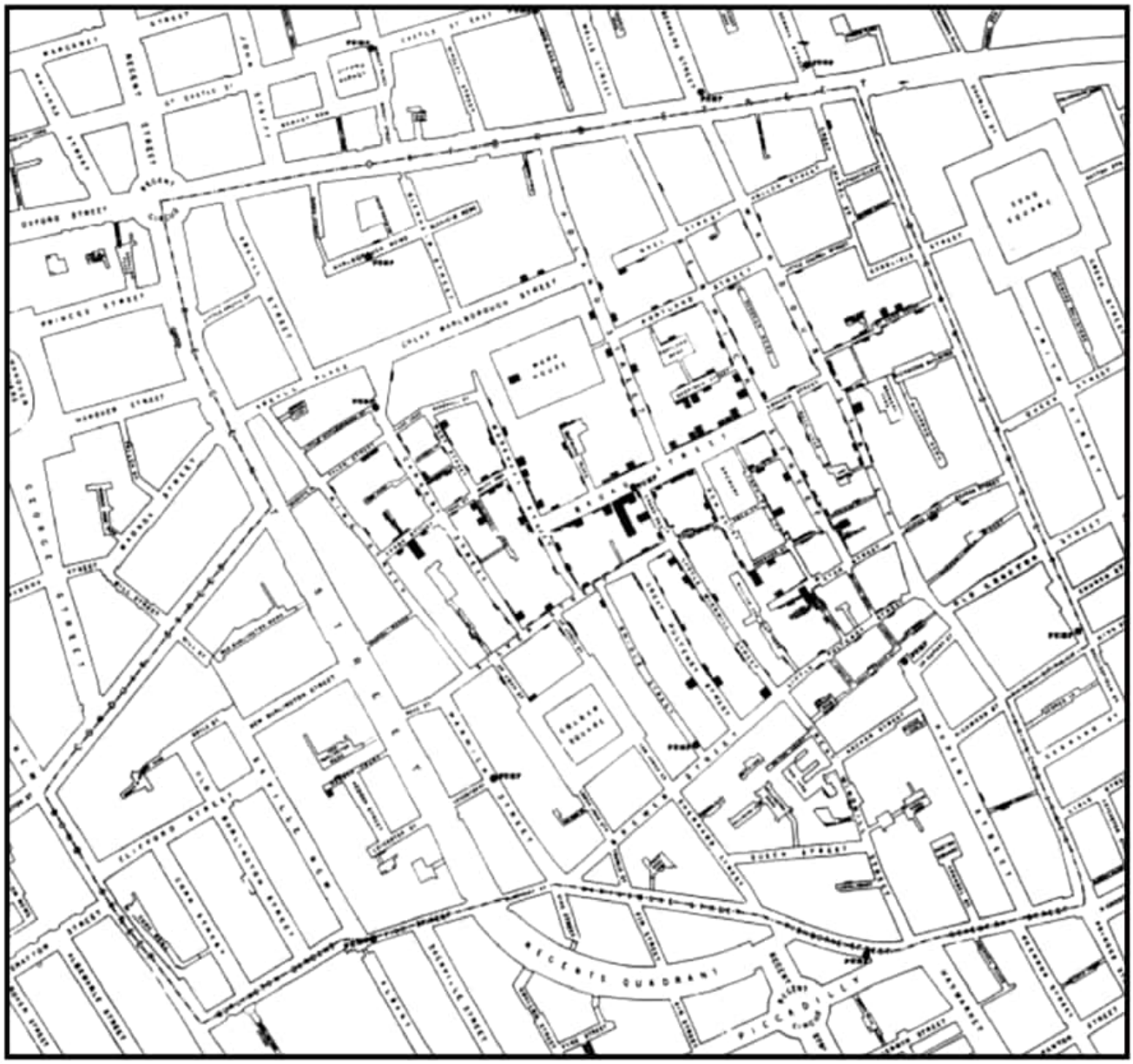


FIGURE 1.6 John Snow's spot map of cholera outbreak in London, 1854.

[2] Source: http://en.wikipedia.org/wiki/John_Snow_(physician)

Edward Tufte (2001), in his book titled *The Visual Display of Quantitative Information*, demonstrated how innovative visuals can be used to effectively communicate data. Google search keywords are used to predict demand for different apparel styles, jewellery, footwear, and so on to understand demand trends for many products. These trends help retailers take better decisions regarding procurement and inventory planning. Dashboards are created using innovative visuals form the core of business intelligence and are an important element of analytics. Tableau and Qlik Sense are popular visualization tools that are used by several organizations to create dashboards to monitor several key performance indicators relevant for the organization in real time. Indian companies such as Gramener[3] have used innovative data visualization tools to communicate hidden facts in the data. Descriptive analytics could be the initial stage of creating analytics capability.

Simple analysis of data can lead to business practices that result in financial rewards. For instance, companies such as RadioShack and Best Buy found a high correlation between the success of individual stores and the number of female employees in the sales team (Underhill, 2009). Underhill (2009) also reported that the conversion rate (percentage of people who purchased something) in consumer durable shops was higher among female shoppers than among male shoppers. Many organizations across the globe have to deal with fraudulent transactions. Sometimes, a simple query can lead to fraud detection. In 2014, China Eastern Airline found that a man had booked a first class ticket more than 300 times in a year and cancelled it before its expiry for full refund so that he could eat free food at the airport's VIP lounge (David K Li, 2014). In India, insurance frauds accounted for 2500–3500 crore in 2010 (Anon, 2013). It is always a good practice to start analytics projects with descriptive analytics.

1.5 | PREDICTIVE ANALYTICS

> If you torture the data long enough, it will confess.
>
> — **Ronald Coase**

In the analytics capability maturity model (ACMM), predictive analytics comes after descriptive analytics and is the most important analytics capability. It aims to predict the probability of occurrence of a future event such as forecasting demand for products/services, customer churn, employee attrition, loan defaults, fraudulent transactions, insurance claim, and stock market fluctuations. While descriptive analytics is used for finding what has happened in the past, predictive analytics is used for predicting what is likely to happen in the future. The ability to predict a future event such as an economic slowdown, a sudden surge or decline in a commodity's price, which customer is likely to churn, what will be the total claim from auto insurance customer, how long a patient is likely to stay in the hospital, and so on will help organizations plan their future course of action. Anecdotal evidence suggests that predictive analytics is the most frequently used type of analytics across several industries. The reason for this is that almost every organization would like to forecast the demand for the products that they sell, prices of the materials used by them, and so on. Irrespective of the type of business, organizations would like to forecast the demand for their products or services and understand the causes of demand fluctuations. The use of predictive analytics can reveal relationships that were previously unknown and are not intuitive.

[3] Source: https://gramener.com/

The most popular example of the application of predictive analytics is Target's pregnancy prediction model discussed earlier in the chapter. In 2002, Target hired statistician Andrew Pole; one of his assignments was to predict whether a customer is pregnant (Duhigg, 2012). At the outset, the question posed by the marketing department to Pole may look bizarre, but it made great business sense. Any marketer would like to identify the price-insensitive customers among the shoppers, and who can beat soon-to-be parents? A list of interesting applications of predictive analytics is presented in Table 1.2.

The examples shown in Table 1.2 represent a tiny fraction of the predictive analytics applications used in the industry. Companies such as Procter & Gamble use analytics as a competitive strategy — every critical management decision is made using analytics (Davenport, 2013). If one were to search for the reasons behind highly successful companies, one would usually find analytics being deployed as the competitive strategy. Google — without which many people think the world would end — uses Markov chains for page ranking (Hayes, 2013). Google also developed accurate prediction models that could predict events such as the outcome of political elections, the launch date of a product, or action(s) taken by competitors (Coles *et al.*, 2007). Davenport and Harris (2007) reported how companies such as Amazon, Capital One, Harrah's, and the Boston Red Sox have dominated their business by using analytics. The application of analytics is not restricted to big corporates only; many sports clubs have successfully used analytics to manage their clubs. The most famous application of analytics in sports is by Oakland Athletics, which used analytics to put together a team with the limited resources available

TABLE 1.2 List of predictive analytics applications

Organization	Predictive Analytics Model
Polyphonic HMI	Predicts whether a song will be a hit using machine learning algorithms. Their product 'Hit Song Science' uses mathematical and statistical techniques to predict the success of a song on a scale of 1 to 10 (Anon, 2003).
Okcupid	Predicts which online dating message is likely to get a response from the opposite sex (Siegel, 2013).
Amazon.com	Uses predictive analytics to recommend products to their customers. It is reported that 35% of Amazon's sales is achieved through their recommender system (Siegel, 2013, MacKinzie *et al.*, 2013).
Hewlett Packard (HP)	Developed a flight risk score for its employees to predict who is likely to leave the company (Siegel, 2013).
University of Maryland	Claimed that dreams can predict whether one's spouse will cheat (Whitelocks, 2013).
Flight Caster	Predicts flight delays 6 hours before the airline's alerts.
Netflix	Predicts which movie their customer is likely to watch next (Greene, 2006). 75% of what customer watch at Netflix is from product recommendations (MacKinzie *et al.*, 2013).
Capital One Bank	Predicts the most profitable customer (Davenport, 2007).
Google	Predicted the spread of H1N1 flu using the query terms (Carneiro and Mylonakis, 2010).
Farecast	Developed a model to predict airfare, whether it is likely to increase or decrease, and the amount of increase/decrease.[4]

[4] Source: http://www.crunchbase.com/company/farecast

for purchasing players (Lewis, 2003). Oakland Athletics had the third lowest payroll among the major league baseball teams in 2002. The manager of Oakland Athletics, Billy Beane, used statistical techniques to identify player qualities that made an impact on the match outcome and to also identify relatively cheaper skill. Oakland Athletics revised their team management strategy and with a payroll of USD 41 million, they were able to compete successfully in the league. In 2002, they won 20 games in a row.

1.6 | PRESCRIPTIVE ANALYTICS

Every decision has a consequence.
— **Damon Darrel**

Prescriptive analytics is the highest level of analytics capability which is used for choosing optimal actions once an organization gains insights through descriptive and predictive analytics. In many cases, prescriptive analytics is solved as a separate optimization problem. Prescriptive analytics assists users in finding the optimal solution to a problem or in making the right choice/decision among several alternatives. Operations Research (OR) techniques form the core of prescriptive analytics. Apart from operations research techniques, machine learning algorithms, metaheuristics, and advanced statistical models are used in prescriptive analytics. Note that actionable items can be derived directly after descriptive and predictive analytics model development; however, they may not be the optimal action. For example, in a Business to Business (B to B) sales, the proportion of sales conversions to sales leads could be very low. The sales conversion period could be very long, as high as 6 months to one year. Predictive analytics such as logistics regression can be used for predicting the propensity to buy a product and actionable items (such as which customer to target) can be derived directly based on predicted probability to buy or using lift chart. However, the values of the sale are likely to be different, as are the profits earned from different customers. Thus, targeting customers purely based on probability to buy may not result in an optimal solution. Use of techniques such as binary integer programming will result in optimal targeting of customers that maximize total expected profit. That is, while actionable items can be derived from descriptive and predictive analytics, use of prescriptive analytics ensures optimal actions (choices or alternatives). The link between different analytics capability is shown in Figure 1.7.

Ever since their introduction during World War II, OR models have been used in every sector of every industry. The list of prescriptive analytics applications is long and several companies across the world have benefitted from the use of prescriptive analytics tools. Coca-Cola Enterprises (CCE) is the largest distributor of Coca-Cola products. In 2005, CCE distributed 2 billion physical cases

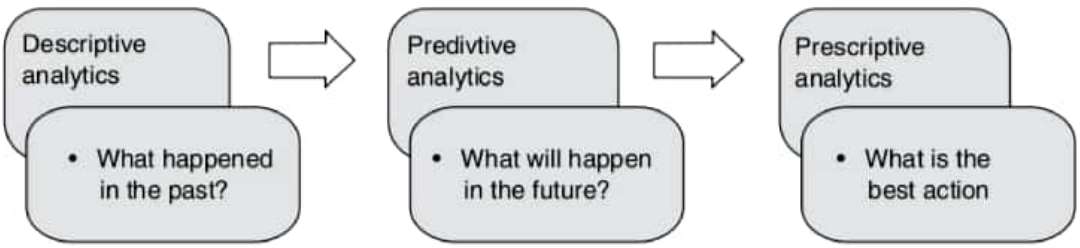


FIGURE 1.7 Link between different analytics capabilities.

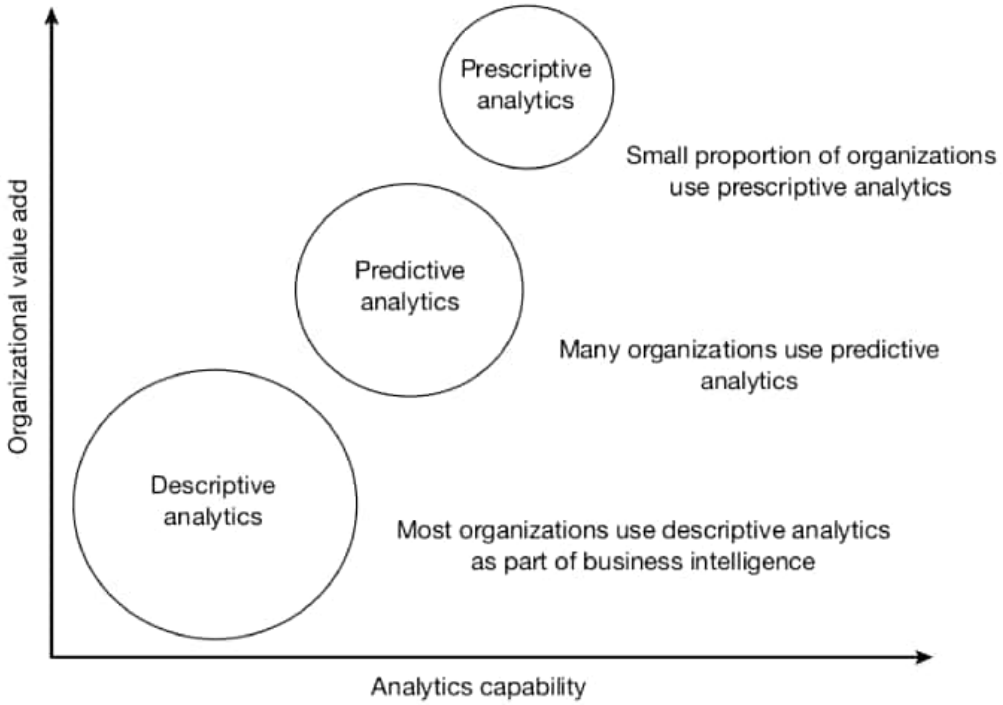


FIGURE 1.8 Analytics capability versus organizational value add.

containing 42 billion bottles and cans of Coca-Cola (Kant *et al.*, 2008). CCE developed an OR model that would meet several objectives such as improved customer satisfaction and optimal asset utilization for their distribution network of Coca-Cola products from 430 distribution centres to 2.4 million retail outlets. The optimization model resulted in cost savings of USD 54 million and improved customer satisfaction. The Akshaya Patra Midday Meal Routing and Transportation Algorithm (AMRUTA) was developed to solve the vehicle routing problem (discussed in Section 1.1); this was implemented at Akshaya Patra's Vasanthapura campus, resulting in savings of USD 75000 per annum (Mahadevan *et al.*, 2013). A major challenge for any e-commerce company is to improve the conversion of visits to transactions and order sizes. Hewlett Packard (HP) established HPDirect.com in 2005 to build online sales. HP Global Analytics developed predictive and prescriptive analytics techniques to improve sales. The analytical solutions helped HP to increase conversion rates and order sizes (Rohit *et al.*, 2013).

Inventory management is one of the problems that is most frequently addressed using prescriptive analytics. Samsung implemented a set of methodologies under the title '*Short Life and Low Inventory in Manufacturing*' (SLIM) to manage all the manufacturing and supply chain problems. Between 1996 and 1999, Samsung implemented SLIM in all its manufacturing facilities, resulting in a reduction in the manufacturing cycle time of random access memory devices from more than 80 days to less than 30 days. SLIM enabled Samsung to capture additional markets worth USD 1 billion (Leachman *et al.*, 2002). Product mix, marketing mix, travelling salesman problem, vehicle routing, facility location, manpower planning, capital budgeting, transportation and capacity management are a few frequently addressed prescriptive analytics problems. Figure 1.8 shows analytics capability and organizational value add; prescriptive analytics provides maximum value add to the organization since the benefits are realized during every period.

Descriptive Analytics

"The Purpose of Visualization is Insight – Not Pictures".
—**Ben Shneiderman**

LEARNING OBJECTIVES

- **LO2-1** Understand the basic concepts in descriptive analytics and how it is used in data-driven decision making.
- **LO2-2** Learn different variable types such as qualitative and quantitative along with scales of measurement such as nominal, ordinal, interval and ratio.
- **LO2-3** Understand data types such as cross-sectional data, time series data and panel data.
- **LO2-4** Understand the difference between population and sample and gain insights through fundamental concepts in statistics such as measures of central tendency, measures of variability and measures of shape.
- **LO2-5** Learn data visualization and various types of visual charts.
- **LO2-6** Understand the application of descriptive analytics in decision making.

ESSENCE OF DESCRIPTIVE ANALYTICS

Descriptive analytics is about finding "what has happened" by summarizing the data using innovative methods and analysing the past data using simple queries. Analysing past data can provide insights that can assist organizations to take appropriate decisions. Consider the Walmart example discussed in Chapter 1, where they found that during the hurricane season the demand for strawberry pop-tart increased seven times the normal season; this is a very good example for application of descriptive statistics. Based on this insight, Walmart ensured that there is enough stock of strawberry pop-tarts in the stores during a hurricane season. John Snow's spot map on Cholera outbreak in London and his final hypothesis that Cholera is water-borne disease is another classic example of application of descriptive analytics through data visualization. There are many such examples where simple analysis of the past data has revealed interesting facts such as difference in shopping behaviour of men and women, relationship freeze, etc. Descriptive analytics involves data summarization – using techniques such as pivot tables, descriptive statistics and data visualization that can be used for analysing past data to gain insights and hidden patterns.

IMPORTANT

Descriptive analytics is the starting point of analytics based solution to problems. It helps to understand the data and provide directions for predictive and prescriptive analytics. Business Intelligence (BI), which largely involves creating reports and business dashboard that lead to actionable insights, is essentially a descriptive analytics exercise.

2.1 | INTRODUCTION TO DESCRIPTIVE ANALYTICS

Descriptive analytics is the science of describing past data and thus capturing "what happened" in a given context. Primary objective of descriptive analytics is simple comprehension of data using data summarization, basic statistical measures and visualization. Various tools and techniques are used in describing the data. Descriptive statistics such as measures of central tendency, measures of variation and measures of shape can provide useful insights. Many different plots such as histogram, bar chart, pie-chart, box-plot, scatter plot and tree diagram can provide insights about past data and subsequently assist with further analysis by generating new hypotheses.

Descriptive analytics is an important part of reporting across several industries which enables top management to monitor key performance indicators and take decisions. Most companies generate reports and dashboards at regular intervals as part of business intelligence (BI) to communicate various aspects of the business to the top management, stakeholders, and the external world. Business reports include descriptive analytics in the form of tables, charts, and innovative diagrams such as Treemap. With the advent of mobile technology, many real-time reports are generated and are accessed by the top management in their mobile handsets enabling them to take quick actions if necessary. For example, a retailer such as Bigbazaar or Reliance retail in India may like to know the top 5 (in terms of revenue generated) products that are sold by region, by city, by store, etc. Such information would assist the management to plan their inventory, shelf space, pricing, etc. They can also monitor trend in revenue generated at regional, city, and store levels over the past several periods. Several companies use dashboards and scorecards to communicate KPIs that are relevant to them; one of the primary applications of descriptive analytics is designing effective dashboards and scorecards.

2.2 | DATA TYPES AND SCALES

Data is classified into different categories based on data structure and scale of measurement of the variables.

2.2.1 | Structured and Unstructured Data

Data at a macro-level can be classified as structured and unstructured data. Structured data means that the data is described in a matrix form with labelled rows and columns. Any data that is not originally in the matrix form with rows and columns is an unstructured data. For example, e-mails, click streams, textual data, images (photos and images generated by medical devices), log data, and videos. Machine-generated data such as images generated by satellite, magnetic resonance imaging (MRI), electrocardiogram (ECG) and thermography are few examples of unstructured data. There is an increasing trend in

the generation of unstructured data due to social media platforms such as Facebook and YouTube and analysis of unstructured data is important for effective management. Internet of things (IoT) is another source unstructured data.

The importance of unstructured data in decision making has increased many folds in the recent past due to its applications to different sectors of the industry. For example, analysing social media data is important for companies to understand the sentiments expressed by the customers about their products/services and take necessary remedial measures. Significant proportion of social media data is natural language (text) apart from images and videos. Apart from social media, machine-generated data are usually unstructured (e.g. data generated from medical devices such as ECG, MRI, etc.). High percentage of Big Data problems constitute unstructured data. One of the main challenges in analysing unstructured data is in the conversion of unstructured data to structured data, which then enables model development. Examples of structured and unstructured data are shown in Tables 2.1 and 2.2.

The data in Table 2.2 is a clickstream data (search behaviour of an internet user that captures the websites visited by the user). Clickstream data is useful for understanding the behaviour of internet users. Based on their surfing (internet browsing) behaviour, individuals are targeted with advertisement for products and services. The unstructured data as shown in Table 2.2 does not have matrix structure as in the case of structured data in Table 2.1. Before any analytics model can be built, unstructured data has to be converted into a structured data.

TABLE 2.1 Structured data consisting of nominal and ratio scales

No.	Gender	Age	Percentage SSC	Board SSC	Percentage HSC	Percentage Degree	Salary
1	M	23	62	Others	88	52	270000
2	M	21	76.33	ICSE	75.33	75.48	220000
3	M	22	72	Others	78	66.63	240000
4	M	22	60	CBSE	63	58	250000
5	M	22	61	CBSE	55	54	180000
6	M	23	55	ICSE	64	50	300000
7	F	24	70	Others	54	65	240000
8	M	22	68	ICSE	77	72.5	235000
9	M	24	82.8	CBSE	70.6	69.3	425000
10	F	23	59	CBSE	74	59	240000

TABLE 2.2 Unstructured data (sample clickstream data)

https://en.wikipedia.org/wiki/Clickstream
http://hortonworks.com/hadoop-tutorial/how-to-visualize-website-clickstream-data/
http://searchcrm.techtarget.com/definition/clickstream-analysis
https://www.qubole.com/blog/big-data/clickstream-data-analysis/

2.2.2 | Cross-sectional, Time Series, and Panel Data

Another important classification of data is based on the type of data collected. Based on the type of data collected, the data is grouped into the following three classes:

1. **Cross-Sectional Data:** A data collected on many variables of interest at the same time or duration of time is called cross-sectional data. For example, consider data on movies such as budget, box-office collection, actors, directors, genre of the movie during year 2017.
2. **Time Series Data:** A data collected for a single variable such as demand for smartphones collected over several time intervals (weekly, monthly, etc.) is called a time series data.
3. **Panel Data:** Data collected on several variables (multiple dimensions) over several time intervals is called panel data (also known as longitudinal data). Example of a panel data is data collected on variables such as gross domestic product (GDP), Gini index, and unemployment rate for several countries over several years.

2.3 | TYPES OF DATA MEASUREMENT SCALES

Structured data can be either numeric or alpha numeric and may follow different scales of measurement (level of measurement). It is important to understand the type of variables within the data with respect to the measurement scale since the model specification while building analytics models such as regression may depend on the scale of measurement.

2.3.1 | Nominal Scale (Qualitative Data)

Nominal scale refers to variables that are basically names (qualitative data) and also known as categorical variables. For example, variables such as marital status (single, married, divorced) and industry type (manufacturing, healthcare, banking and finance) fall under nominal scale. During data collection, it is usual to assign a numerical code to represent a nominal variable. For example, the data collector may have used number 1 to represent singles, 2 for married, and 3 for divorced category for categorical variable marital status. The codes 1, 2, and 3 used here do not have any value attached to them. That is, basic mathematical operations are meaningless in a nominal scale (e.g., subtraction: married – unmarried or ratio: married/unmarried are meaningless). While developing statistical models, nominal scale data are usually transformed before building the model. For example, when developing a regression model, categorical variables are converted using dummy variables before building the regression model (is discussed in Chapter 10).

2.3.2 | Ordinal Scale

Ordinal scale is a variable in which the value of the data is captured from an ordered set, which is recorded in the order of magnitude. For example, in many survey data, Likert scale is used. Likert scale is finite (usually a 5 point scale) and the data collector would have defined the order of preference. For example, assume that a feedback is collected on a training program using 5-point Likert scale in which 1 = Poor, 2 = Fair, 3 = Good, 4 = Very Good, and 5 = Excellent. In this case, we know that

5 is better than 4 and 4 is better than 3; however, the difference 5 – 4 (Excellent – Very Good) is meaningless.

2.3.3 | Interval Scale

Interval scale corresponds to a variable in which the value is chosen from an interval set. Variable such as temperature measured in centigrade (°C) or intelligence quotient (IQ) score are examples of interval scale. In interval scale, the ratios do not make sense. For example, 40°C is not twice hot as 20°C. Similarly, a person with an IQ score of 160 is not twice smarter than a person with an IQ score of 80. However, 40°C is 20°C more than 20°C, IQ score of 160 is 80 more than an IQ score of 80. In an interval scale, the reference is fixed arbitrarily, for example 0°C is fixed based on the freezing point of water.

2.3.4 | Ratio Scale

Any variable for which the ratios can be computed and are meaningful is called ratio scale. Most variables come under this type; for example: demand for a product, market share of a brand, sales, salary, and so on. If Ms Hawai Sundari's salary is 40,000 per month and Ms Dawai Sundari's salary is 90,000 per month then we can interpret that Dawai Sundari earns 2.25 times the salary of Hawai Sundari.

2.4 | POPULATION AND SAMPLE

Population is the set of all possible observations (often called cases, records, subjects or data points) for a given context of the problem. The size of the population can be very large in many cases. For example, in 2014, close to 834.08 million people were eligible to vote in the Indian general elections (Source: Election Commission of India). Thus, the population size of the voters in 2014 was 834.08 million which included all eligible voters. During every election, media and other organizations collect data to predict likely winner of election through opinion polls (and they rarely get it right due to complexities associated with collecting right sample). It is very difficult (also practically impossible) to collect data from all 834.08 million eligible voters about their choice of candidate, so the opinion polls are based on opinion expressed by a subset of voters called **sample**.

Population (also known as universal set) is the set of all possible data for a given context whereas sample is the subset taken from a population. In many analytical problems, we make inference about the population based on the sample data. There are many challenges in sampling (process of selecting an observation from the population). An incorrect sample may result in bias and incorrect inference about the population. Sampling is discussed in detail in Chapter 4.

2.5 | MEASURES OF CENTRAL TENDENCY

Measures of central tendency are the measures that are used for describing the data using a single value. **Mean, median** and **mode** are the three measures of central tendency and are frequently used to compare different data sets. Measures of central tendency help users to summarize and comprehend the data.

2.5.1 | Mean (or Average) Value

Mean is the arithmetical average value of the data and is one of the most frequently used measures of central tendency. Assume that the data has n observations in a sample, and let X_i be the value of the i^{th} observation. Then the mean is given by

$$\text{Mean} = \bar{X} = \frac{X_1 + X_2 + \cdots + X_n}{n} = \sum_{i=1}^{n} \frac{X_i}{n} \tag{2.1}$$

Symbol $\bar{X}$ is frequently used to represent the estimated value of the mean from a sample. If the entire population is available and if we calculate mean based on the entire population, then we get the population mean which is denoted by μ. Among the measures of central tendency, mean is the most frequently used measure since it uses all the observations (all X_i values) in the data set (either sample or population) to calculate the mean value. Table 2.1 has the salary of graduating students from a business school; the average salary is given by

$$\bar{X} = \frac{(270 + 220 + 240 + 250 + 180 + 300 + 240 + 235 + 425 + 240) \times 1000}{10} = 260000$$

The average (or mean) salary is 260000. Note that the average value need not be a part of the data set, that is, none of the graduating student's salary is 260000. In Microsoft Excel, function 'Average(array)' can be used for calculating the mean value of the data. Mean can be interpreted as the centre of gravity of the distribution of the data. An important property of mean is that the summation of deviation of observations from the mean is zero, that is

$$\sum_{i=1}^{n} \left(X_i - \bar{X}\right) = 0$$

Associated with the mean is a phenomenon often called "wisdom of crowd", according to which the collective wisdom of people is better than any individual person's knowledge. For example, in 1906, Francis Galton attended a contest in Plymouth, UK in which the villagers were asked to guess the weight of an Ox, the one who guessed the closest won the prize. Around 800 villagers participated in the contest. Francis Galton found that the average of all the weights entered was very close to the actual weight. In fact, the difference was less than a pound. Also, the average turned to be better than the guess by the winner of the contest (Surowiecki, 2004).

One should be careful about taking decisions based on the mean value of the data. There is a famous joke in statistics which says that, "*if someone's head is in freezer and leg is in the oven, the average body temperature would be fine, but the person may not be alive*". Making decisions solely based on mean value is not advisable. In capital asset procurement such as procurement of fighter aircraft and weapons, defence services across the world use mean time between failures (MTBF) as one of the measures of system reliability (performance). However, MTBF (which is the mean value of the time between failure data) in itself is not a useful measure to assess the reliability of the asset and not very useful in taking operational decisions. It has to be used along with other measures and measures of variability for better understanding of the data. Another issue with mean is, it is affected significantly by presence of

outliers. That is, presence of an outlier can change the mean value significantly. If the data is captured in frequencies, then Eq. (2.2) can be used for calculating the average:

$$\bar{X} = \sum_{i=1}^{n} \frac{f_i X_i}{f_i} \tag{2.2}$$

The frequency of age of students in Table 2.1 is given below:

Age	21	22	23	24
Frequency	1	4	3	2

The average age of students using Eq. (2.2) is given by

$$\bar{X} = \frac{1\times 21 + 4\times 22 + 3\times 23 + 2\times 24}{1+4+3+2} = 22.6$$

2.5.2 | Median (or Mid) Value

Median is the value that divides the data into two equal parts, that is, the proportion of observations below median and above median will be 50%. Easiest way to find the median value is by arranging the data in the increasing order and the median is the value at position $(n + 1)/2$ when n is odd. When n is even, the median is the average value of $(n/2)^{th}$ and $(n + 2)/2^{th}$ observation after arranging the data in the increasing order.

Consider the example of a bank. The number of deposits in a branch of a bank in a week is shown in Table 2.3.

TABLE 2.3 Number of daily deposits in a Bank

Day	1	2	3	4	5	6	7
Number of Deposits	245	326	180	226	445	319	260

The ascending order of the data in Table 2.3 is given by

180, 226, 245, 260, 319, 326 and 445

Now $(n + 1)/2 = (8/2) = 4$. Thus the median is the 4^{th} value in the data after arranging them in the increasing order; in this case it is 260. There are equal numbers of observation below and above 260. In Microsoft Excel, the function 'Median(array)' can be used for calculating the median of a data set.

Another example is the salary in Table 2.1 that can be arranged as follows:

180000, 220000, 235000, 240000, 240000, 240000, 250000, 270000, 300000, 425000

The 5^{th} and 6^{th} observations are 240000 and 240000 and the average is 240000. Thus, the median salary for the data in Table 2.1 is 240000. Median is much more stable than the mean value, that is adding a new observation may not change the median significantly. However, the drawback of median is that it is not calculated using the entire data like in the case of mean. We are simply looking for the midpoint instead of using the actual values of the data.

2.5.3 | Mode

Mode is the most frequently occurring value in the data set. For example, in the data 'salary' in Table 2.1, the value 240000 is appearing three times and is the mode since all other values are observed only once. In Microsoft Excel, the function 'Mode(array)' can be used for calculating mode. Mode is the only measure of central tendency which is valid for qualitative (nominal) data since the mean and median for nominal data are meaningless. For example, assume that a customer data with a retailer has the marital status of customer, namely, (a) Married, (b) Unmarried, (c) Divorced Male, and (d) Divorced Female. Mean and median are meaningless when we try to use them on a qualitative data such as marital status. On the other hand, mode will capture the customer type in terms of marital status that occurs most frequently in the database. In the bar chart (and histogram), mode is the tallest column. It is possible that a data set may not have any mode at all. For example, if each value in the data set appears only once, then there is no mode in the data set.

2.6 | PERCENTILE, DECILE, AND QUARTILE

Percentile, decile and quartile are frequently used to identify the position of the observation in the data set. Percentile score is frequently used in education to identify the position of a student in the group. Another frequent application of percentile is the percentile life used in asset management. Percentile, denoted as P_x, is the value of the data at which x percentage of the data lie below that value. For example, P_{10} denotes the value below which 10 percentage of the data lies. To find P_x, we have to arrange the data in the increasing order and the value of P_x is the position in the data calculated using Eq. (2.3):

$$\text{Position corresponding to } P_x \approx \frac{x(n+1)}{100} \tag{2.3}$$

where n is the number of observations in the data. Note that the value obtained from Eq. (2.3) can be non-integer, in which case we can either round it to the nearest integer or use an approximation which will be explained in Example 2.1. **Decile** corresponds to special values of percentile that divide the data into 10 equal parts. First decile contains first 10% of the data and second decile contains first 20% of the data and so on. Similarly, **Quartile** divides the data into 4 equal parts. The first quartile (Q_1) contains first 25% of the data, Q_2 contains 50% of the data and is also the median. Quartile 3 (Q_3) accounts for 75% of the data. In Microsoft Excel, the function 'Percentile(array, k)' provides P_x value. That is, Percentile(array, 0.1) will give 10th percentile.

EXAMPLE 2.1

Time between failures (in hours) of a wire cutter used in a cookie manufacturing oven is given in Table 2.4. The function of the wire-cut is to cut the dough into cookies of desired size.

TABLE 2.4 Time between failures of wire-cut (in hours)

2	22	32	39	46	56	76	79	88	93
3	24	33	44	46	66	77	79	89	99
5	24	34	45	47	67	77	86	89	99
9	26	37	45	55	67	78	86	89	99
21	31	39	46	56	75	78	87	90	102

(a) Calculate the mean, median, and mode of time between failures of wire-cuts.
(b) The company would like to know by what time 10% (ten percentile or P_{10}) and 90% (ninety percentile or P_{90}) of the wire-cuts will fail?
(c) Calculate the values of P_{25} and P_{75}.

Solution:

(a) Mean = 57.64, median = 56, and mode = 46, 89 and 99.

(b) Note that the data in Table 2.4 is arranged in increasing order in columns. The position of $P_{10} = 10 \times (51)/100 = 5.1$. We can round off 5.1 to its nearest integer which is 5. The corresponding value from table is 21 (10 percentage of observations in Table 2.4 have a value of less than or equal to 21). That is, by 21 hours, 10% of the wire-cuts will fail. In asset management (and reliability theory), this value is called P_{10} life.

Instead of rounding the value obtained from Eq. (2.3), we can use the following approximation:

$$\text{Position corresponding to } P_{10} = 10 \times (51)/100 = 5.1$$

Value at 5th position is 21. Value at position 5.1 is approximated as
21 + 0.1 × (value at 6th position – value at 5th position) = 21 + 0.1(1) = 21.1

$$\text{Position corresponding to } P_{90} = 90 \times 51/100 = 45.9$$

The value at position 45 is 90 and the value at position 45.9 is

90 + 0.9 (value at 46th position – value at 45th position) = 90 + 0.9 × (3) = 92.7

That is, 90% of the wire-cuts will fail by 92.7 hours.

(c) Position corresponding to P_{25} (1st Quartile or Q_1) = 25 × 51/100 = 12.75
Value at 12th position is 33, so
P_{25} = 33 + 0.75 (value at 13th position – value at 12th position) = 33 + 0.75 (1) = 33.75
Position corresponding to P_{75} (3rd Quartile or Q_3) = 75 × 51/100 = 38.25
Value at 38th position is 86, so
P_{75} = 86 + 0.25 (value at 39th position – value at 38th position) = 86 + 0.25 (0) = 86

2.7 | MEASURES OF VARIATION

One of the primary objectives of analytics is to understand the variability in the data. Predictive analytics techniques such as regression attempt to explain variation in the outcome variable (Y) using predictor variables (X). Variability in the data is measured using the following measures:

1. Range
2. Inter-Quartile Distance (IQD)
3. Variance
4. Standard Deviation

Let us discuss each of them in detail.

2.7.1 | Range

Range is the difference between maximum and minimum value of the data. It captures the data spread. In the data in Table 2.4, the range = 102 – 2 = 100.

2.7.2 | Inter-Quartile Distance (IQD)

Inter-quartile distance (IQD), also called inter-quartile range (IQR), is a measure of the distance between Quartile 1 (Q_1) and Quartile 3 (Q_3). For the data in Table 2.4, we calculated Q_1 as 33.75 and Q_3 as 86. Thus the IQD = 86 – 33.75 = 52.25. IQD is a useful measure for identifying outliers in the data. Outlier is an observation which is far away (on either side) from the mean value of the data. Values of data below Q_1 – 1.5 IQD and above Q_3 + 1.5 IQD are classified as outliers.

For the data in Table 2.4

$$Q_1 - 1.5 \text{ IQD} = 33.75 - 1.5 \times 52.25 = -44.625$$
$$Q_3 + 1.5 \text{ IQD} = 86 + 1.5 \times 52.25 = 164.375$$

In Table 2.4, there are no values either below – 44.625 or above 164.375, thus there are no outliers. Note that IQD is one of the approaches used for identifying outliers; we will discuss other approaches that are used for identifying outliers in Chapters 9 and 10. Also, using IQD for identifying outliers is appropriate only in the case of univariate data (data with one dimension). In the case of multivariate data, we use distance measures such as Mahalanobis distance to identify outliers (discussed in Chapters 9 and 10).

2.7.3 | Variance and Standard Deviation

Variance is a measure of variability in the data from the mean value. Variance for population, σ^2, is calculated using

$$\text{Variance} = \sigma^2 = \sum_{i=1}^{n} \frac{(X_i - \mu)^2}{n} \tag{2.4}$$

Note that, in Eq. (2.4), deviation from mean is squared since sum of deviations from mean will always add up to zero. The variance for the data in Table 2.4 is 818.0304 [using Eq. (2.4)]. In case of a sample, the Sample Variance (S^2) is calculated using

$$S^2 = \sum_{i=1}^{n} \frac{(X_i - \bar{X})^2}{n-1} \tag{2.5}$$

While calculating sample variance S^2, the sum of squared deviation $\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2$ is divided by $(n-1)$. This is known as Bessel's correction. For the data in Table 2.4, the sample standard variance is 834.7249. Microsoft Excel functions Var.P(array) and Var.S(array) are used for calculating population variance and sample variance, respectively. The population standard deviation (σ) and sample standard deviation (S) are given by

$$\sigma = \sqrt{\sum_{i=1}^{n} \frac{(X_i - \mu)^2}{n}} \tag{2.6}$$

For the data in Table 2.4, the standard deviation obtained using the Eq. (2.6) is 28.6012.

$$S = \sqrt{\sum_{i=1}^{n} \frac{(X_i - \bar{X})^2}{n-1}} \tag{2.7}$$

For the data in Table 2.4, the standard deviation obtained using the Eq. (2.7) is 28.8916. In Microsoft Excel, functions Stdev.P(array) and Stdev.S(array) are used for calculating population standard deviation and sample standard deviation respectively. There are two arguments for dividing the sum of squared deviations from mean by $(n-1)$ instead of n in Eqs. (2.5) and (2.7). One argument is that, when we take a sample and estimate the mean from the sample $\overline{X}$, we tend to underestimate the sum of squared deviations from the mean. For example, take a sample consisting of first 5 (first column) and last 5 (last column) observations from Table 2.4. The sample is given in Table 2.5.

TABLE 2.5 Sample of 10 observations from Table 2.4

2	3	5	9	21	93	99	99	99	102

The mean $\overline{X}$ for the sample in Table 2.5 is 53.2 and standard deviation [using Eq. (2.7)] is 47.9740. When we estimate the numerator, $(X_i - \mu)^2$, in Eq. (2.4) using $\overline{X}$, instead of μ, we will underestimate $(X_i - \mu)^2$ resulting in underestimation of standard deviation. The calculations of $(X_i - \overline{X})^2$ and $(X_i - \mu)^2$ for the sample in Table 2.5 are shown in Table 2.6.

TABLE 2.6 Underestimation of standard deviation in sample

Data	Standard deviation (using sample mean 53.2)	Standard deviation (using population mean 57.64)
2	2621.44	3095.81
3	2520.04	2985.53
5	2323.24	2770.97
9	1953.64	2365.85
21	1036.84	1342.49
93	1584.04	1250.33
99	2097.64	1710.65
99	2097.64	1710.65
99	2097.64	1710.65
102	2381.44	1967.81
Sample Mean = 53.2	$\sum(X_i - \bar{X})^2 = 20713.60$	$\sum(X_i - \mu)^2 = 20910.74$

In Table 2.6, we can see that the numerator in Eq. (2.4) is underestimated (20713.60) when we use the sample average against population average (20910.74). This will result in underestimation of the standard deviation, a phenomenon called **downward bias**. To overcome this bias, we divide $\sum(X_i - \bar{X})^2$ with $(n-1)$ instead of n.

Another argument of using Eq. (2.5) is through the concept of **degrees of freedom**. The following two definitions are used for degrees of freedom (Pandey and Bright, 2008):

1. Degrees of freedom is equal to the number of independent variables in the model (Trochim, 2005). For example, we can create any sample of size n with mean value of $\overline{X}$ by randomly selecting $(n-1)$ values. We need to fix just one out of n values. Thus the number of independent variables in this case is $(n-1)$.
2. Degrees of freedom is defined as the difference between the number of observations in the sample and number of parameters estimated (Walker 1940, Toothaker and Miller, 1996). If there are n observations in the sample and k parameters are estimated from the sample, then the degrees of freedom is $(n-k)$. While using Eq. (2.5) or Eq. (2.7), the value of $\overline{X}$ is estimated from the sample. Thus the degrees of freedom is $(n-1)$.

Whenever we estimate a parameter from a sample, we lose a degree of freedom. While estimating standard deviation from a sample, we tend to underestimate since mean is also estimated from the sample itself. The downward bias is addressed by dividing the sum of squared deviation from mean with $(n-1)$ instead of n.

2.7.4 | Chebyshev's Theorem

Chebyshev's theorem (also known as Chebyshev's inequality) is an empirical rule that allows us to predict proportion of observations that is likely to lie between an interval defined using mean and standard deviation. Probability of finding a randomly selected value in an interval defined by $\mu \pm k\sigma$ is $\geq 1-\frac{1}{k^2}$, that is

$$P(\mu - k\sigma \leq X \leq \mu + k\sigma) \geq 1 - \frac{1}{k^2} \tag{2.8}$$

Equation (2.8) is useful when the value of $k > 1$, otherwise it gives a trivial solution.

EXAMPLE 2.2

Amount spent per month by a segment of credit card users of a bank has a mean value of 12000 and standard deviation of 2000. Calculate the proportion of customers who are spending between 8000 and 16000.

Solution:

$$P(8000 \leq X \leq 16000) = P(\mu - 2\sigma \leq X \leq \mu + 2\sigma) \geq 1 - \frac{1}{2^2} = 0.75$$

That is, the proportion of customers spending between 8000 and 16000 is at least 0.75 (or 75%)

2.8 | MEASURES OF SHAPE – SKEWNESS AND KURTOSIS

Skewness is a measure of symmetry or lack of symmetry. A data set is symmetrical when the proportion of data at equal distance (measured in terms of standard deviation) from mean (or median) is equal. That is, the proportion of data between μ and $\mu - k\sigma$ is same as μ and $\mu + k\sigma$, where k is some positive constant. This implies that the distribution (or proportion) of the data on either side of mean (and median) is same. Measure of skewness can be used to identify whether the distribution is left skewed (longer tail on left side of the distribution) or right skewed (longer tail on the right side of the distribution). There are many different approaches to measuring skewness. **Pearson's moment coefficient of skewness** for a data set with n observations is given by

$$g_1 = \frac{\sum_{i=1}^{n} (X_i - \mu)^3 / n}{\sigma^3} \tag{2.9}$$

The value of g_1 will be close to 0 when the data is symmetrical. A positive value of g_1 indicates a positive skewness and a negative value indicates negative skewness. The formula in Eq. (2.9) is adjusted for sample size when skewness is calculated from a sample. The following formula is used usually for a sample with n observations (Joanes and Gill, 1998):

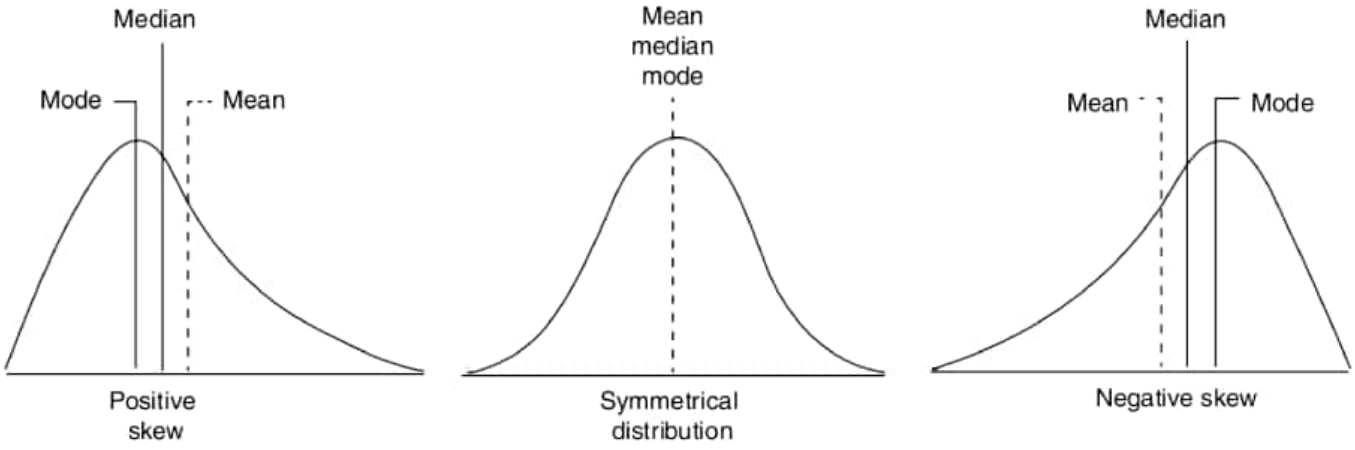


FIGURE 2.1 Skewness.

$$G_1 = \frac{\sqrt{n(n-1)}}{n-2} g_1 \tag{2.10}$$

The value of $\frac{\sqrt{n(n-1)}}{n-2}$ will converge to 1 as the value of n increases. For the data in Table 2.4, the value of G_1 is −0.232. Since the value of G_1 is negative, we can conclude that the data is left skewed. In Microsoft Excel, function 'SKEW(array)' can be used for calculating the value of skewness (G_1) calculated from a sample. In Figure 2.1, the positive skewed (right tailed), normal, and negative skewed (left tailed) distributions are shown.

Kurtosis is another measure of shape, aimed at shape of the tail, that is, whether the tail of the data distribution is heavy or light. Kurtosis is measured using the following equation:

$$\text{Kurtosis} = \frac{\sum_{i=1}^{n} \left(X_i - \bar{X}\right)^4 / n}{\sigma^4} \tag{2.11}$$

Kurtosis value of less than 3 is called **platykurtic distribution** and greater than 3 is called **leptokurtic distribution**. The kurtosis value of 3 indicates standard normal distribution (also called **mesokurtic**). The excess kurtosis is a measure that captures deviation from kurtosis of a normal distribution and is given by

$$\text{Excess Kurtosis} = \frac{\sum_{i=1}^{n} \left(X_i - \bar{X}\right)^4 / n}{\sigma^4} - 3 \tag{2.12}$$

For the data in Table 2.4, excess Kurtosis = −1.0968 (that is kurtosis is 1.9032). Figure 2.2 shows shapes of platykurtic, mesokurtic, and leptokurtic distributions. In Microsft Excel, 'KURT(array)' can be used for calculating the excess kurtosis.

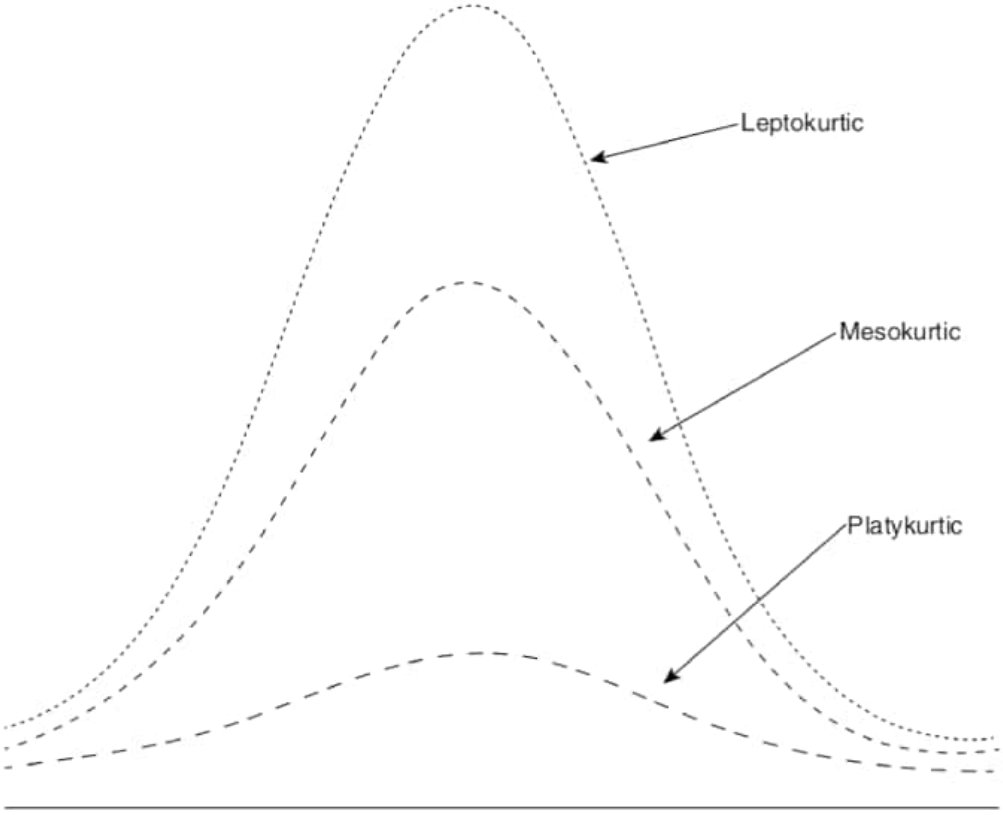


FIGURE 2.2 Leptokurtic, mesokurtic, and platykurtic distributions.

2.9 | DATA VISUALIZATION

Data visualization is an integral part of descriptive analytics and it assists decision makers with useful insights. There are many useful charts such as histogram, bar chart, pie-chart, box-plot that would assist data scientist with visualization of the data. In the recent years, tree maps and sunburst maps are very popular among analytics experts, which can create hierarchical visuals of data. It is always advisable to start an analytics project with data visualization.

2.9.1 | Histogram

Histogram is the visual representation of the data which can be used to assess the probability distribution (frequency distribution) of the data. It is a frequency distribution of data arranged in consecutive and non-overlapping intervals. Histograms are created for continuous (numerical) data. The following steps are used in constructing histograms:

1. Divide the data into finite number of non-overlapping and consecutive bins (intervals). The total number of bins to be used can be calculated using Eqs. (2.13) or (2.14).
2. Count the number of observations from the data that fall under each bin (interval).
3. Create a frequency distribution (bin in the horizontal axis and frequency in the vertical axis) using the information obtained in steps 1 and 2.

Histogram is very useful since it assists data scientist to identify the following:

1. The shape of the distribution and to assess the probability distribution of the data.
2. Measures of central tendency such as median and mode.
3. Measures of variability such as spread.
4. Measure of shape such as skewness.

Histograms are also useful in identifying the presence of outliers. One of the first steps in constructing histogram is identifying the number of bins. There are many different formulas used in literature and one of the simplest formula is

$$\text{Number of bins, } N = \frac{X_{max} - X_{min}}{W} \tag{2.13}$$

Here X_{max} and X_{min} are the maximum and minimum values of the data and W is desired the width of the bin (interval). Intervals in histograms are usually of equal size. Sturges (1926) proposed the following formula for calculating the number of bins:

$$\text{Number of bins, } N = 1 + 3.322 \log_{10}(n) \tag{2.14}$$

where n is the total number of observations in the data set. Figures 2.3 and 2.4 show the histogram of Bollywood movie budget in crores of rupees (1 crore = 10 million) and box-office collection, respectively, based on the data of 149 Bollywood movies (Data file: Bollywood Data.Xls).

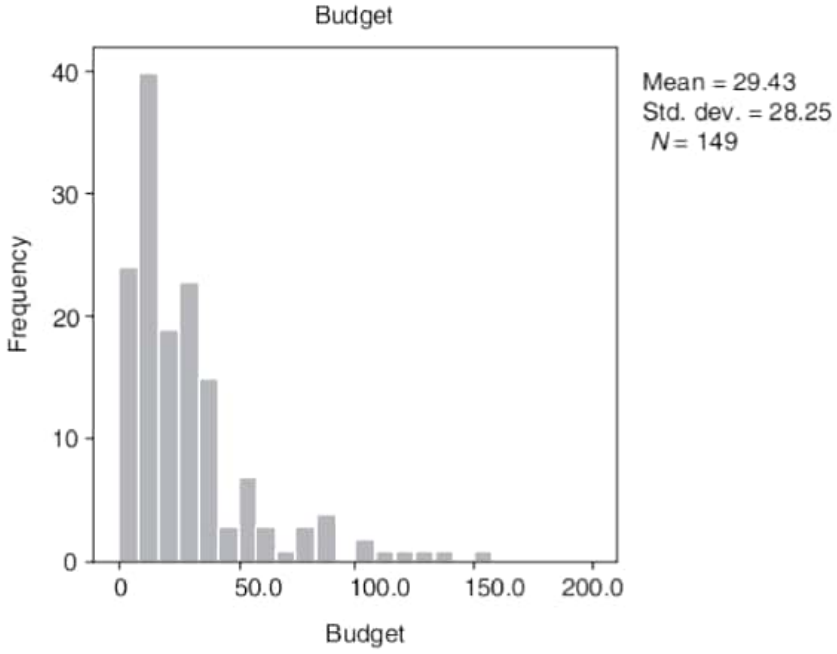


FIGURE 2.3 Histogram of Bollywood movie budget.

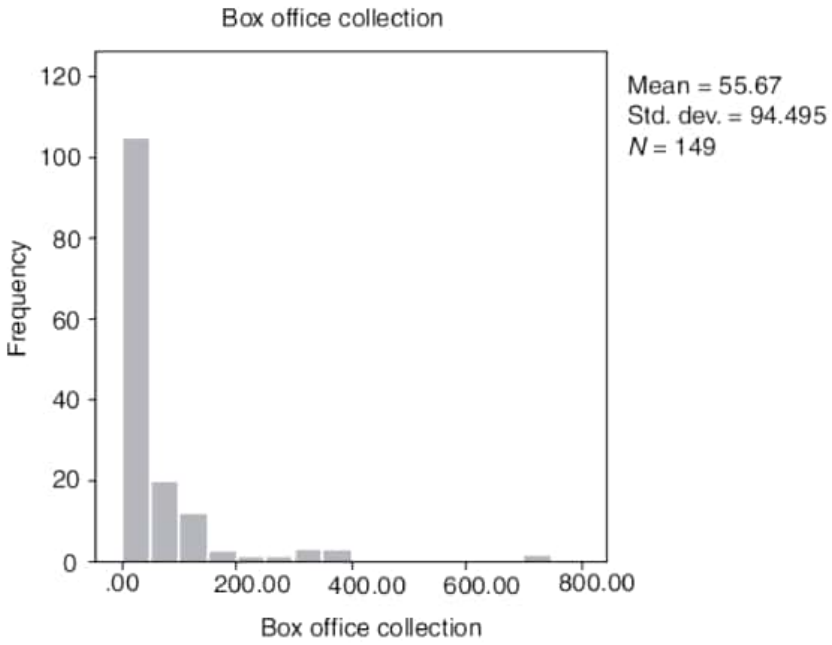


FIGURE 2.4 Histogram of Bollywood movie box-office collection.

From Figure 2.3, we can infer that the budget for large proportion of movies is less than 50 crores and it is a right-skewed distribution (that is, long tail on the right-hand side). In Figure 2.4, we can also see an outlier where the box-office collection is more than 700 crores (movie PK acted by Amir Khan and directed by Rajkumar Hirani). The cumulative histograms are called **Ogive curves**. The Ogive curve for Bollywood box-office collection is shown in Figure 2.5.

Usually, we superimpose normal distribution on the histogram to see how close the frequency distribution of the data is to a normal distribution. Figure 2.6 shows histogram of movie budget superimposed with normal distribution; it is obvious that the frequency distribution of budget is not a normal distribution.

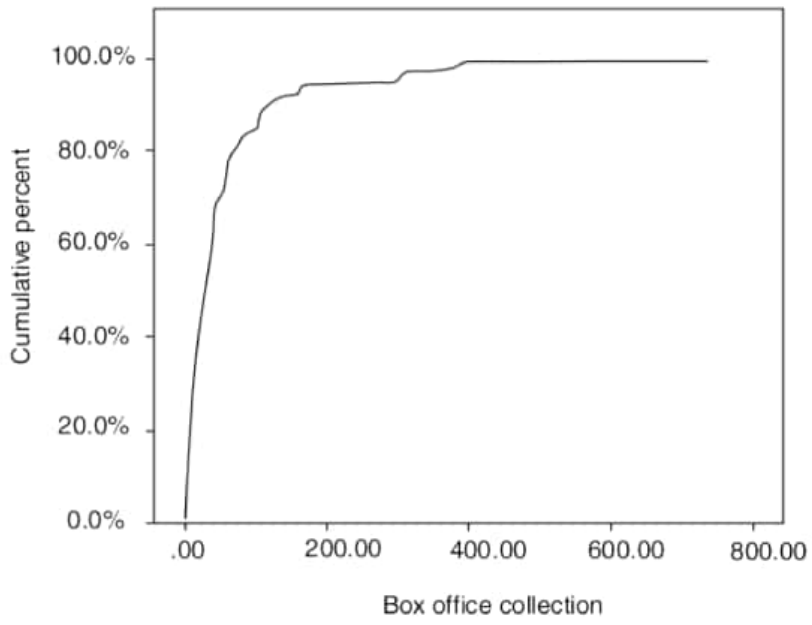


FIGURE 2.5 Ogive curve for box-office collection.

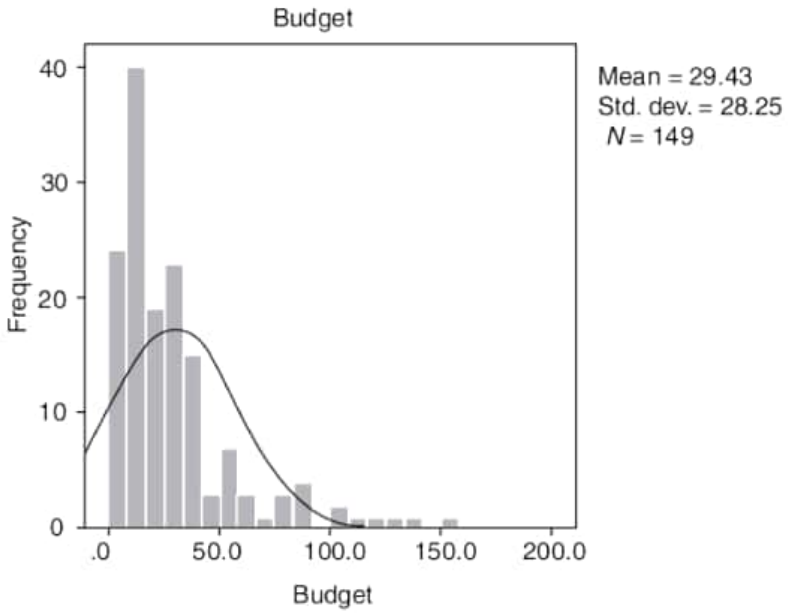


FIGURE 2.6 Histogram of Bollywood movie budget along with normal distribution frequency.

2.9.2 | Bar Chart

Bar chart is a frequency chart for qualitative variable (or categorical variable). Histograms cannot be used when the variable is qualitative. Bar chart can be used to assess the most-occurring and least-occurring categories within a data set. Figure 2.7 shows the bar chart for the movie genre (Data file: Bollywood Data.xlsx). From the bar chart, it is evident that genres, drama and comedy, are mostly preferred by the production houses in Bollywood.

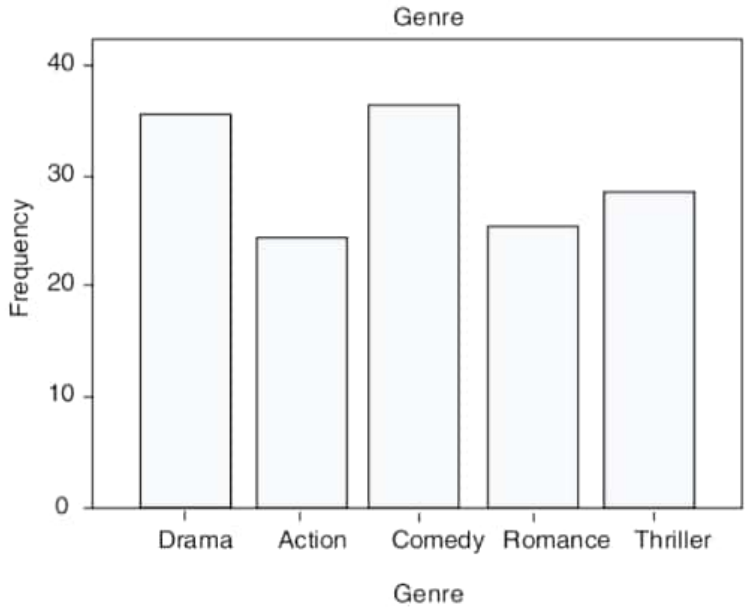


FIGURE 2.7 Bar chart for movie genre.

2.9.3 | Pie Chart

Pie chart is mainly used for categorical data and is a circular chart that displays the proportion of each category in the data set. The pie chart for the movie genre based on the Bollywood movie data set is shown in Figure 2.8.

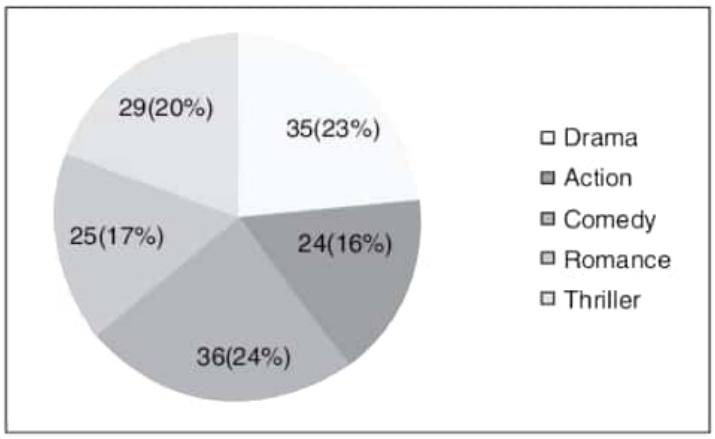


FIGURE 2.8 Pie chart for movie genre.

Pie chart helps to visualize the proportion (percentage) of each category as sector of a circle.

2.9.4 | Scatter Plot

Scatter plot is a plot of two variables that will assist data scientists to understand if there is any relationship between two variables. The relationship could be linear or non-linear. Scatter plot is also useful for assessing the strength of the relationship and to find if there are any outliers in the data. Figure 2.9 shows a scatter plot between the movie budget and movie box-office collection (in crores of rupees) plotted using the data set in file Bollywood Data.xlsx.

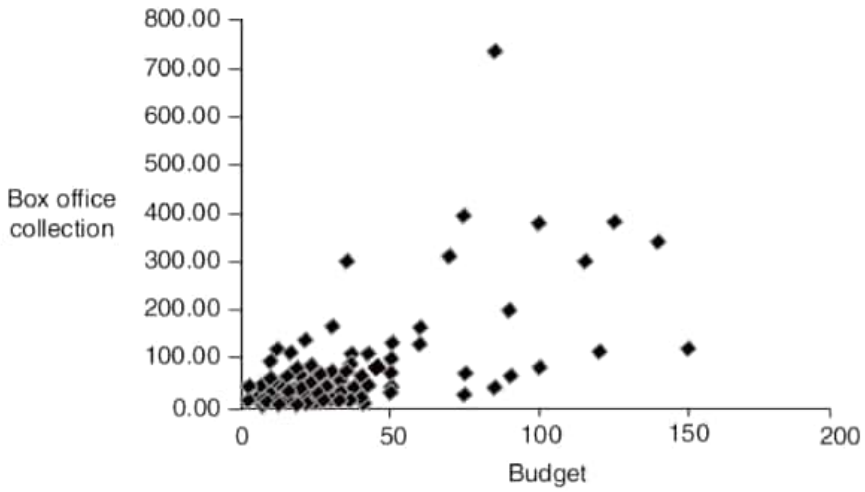


FIGURE 2.9 Scatter plot between movie budget and box-office collection.

Figure 2.9 shows a linear relationship between budget and box-office collection and existence of an outlier. Scatter plots are useful during regression model building to decide on the initial model, that is whether to consider a variable in a regression model or not.

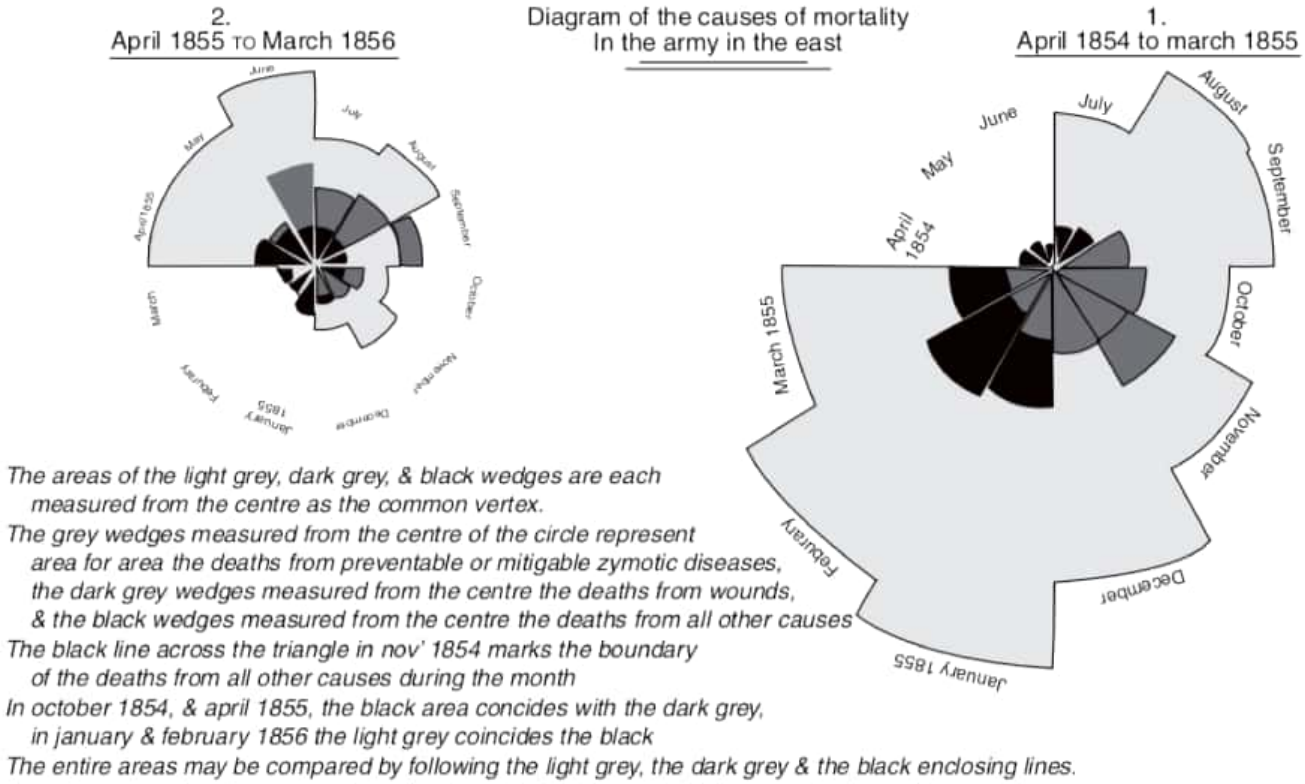
2.9.5 | Coxcomb Chart

Coxcomb chart (also known as polar area chart or roses) is an extension of pie chart made popular by Florence Nightingale (Lewi, 2006). In a Coxcomb chart, each area represents the magnitude of the category. The main difference between the regular pie chart and coxcomb chart is that in the case of pie chart the radius of each sector is same, whereas, in coxcomb chart the radius of the sector is adjusted to create the magnitude of the area.

Florence Nightingale collected data from Crimean war (war between British and French on one side and Russians on the other side) on causes of mortality among soldiers. She classified the causes into three categories:

1. Preventable diseases
2. Wounds sustained in the war
3. Other causes

In Figure 2.10 (originally prepared by Florence Nightingale), the largest area of the chart corresponds to the cause 'preventable diseases'.



The areas of the light grey, dark grey, & black wedges are each measured from the centre as the common vertex.
The grey wedges measured from the centre of the circle represent area for area the deaths from preventable or mitigable zymotic diseases, the dark grey wedges measured from the centre the deaths from wounds, & the black wedges measured from the centre the deaths from all other causes
The black line across the triangle in nov' 1854 marks the boundary of the deaths from all other causes during the month
In october 1854, & april 1855, the black area concides with the dark grey, in january & february 1856 the light grey coincides the black
The entire areas may be compared by following the light grey, the dark grey & the black enclosing lines.

FIGURE 2.10 Coxcomb chart on causes of mortality in the army prepared by Florence Nightingale.[1]

[1] Source: https://en.wikipedia.org/wiki/Florence_Nightingale#/media/File:Nightingale-mortality.jpg

2.9.6 | Box Plot (or Box and Whisker Plot)

Box plot (aka Box and Whisker plot) is a graphical representation of numerical data that can be used to understand the variability of the data and the existence of outliers. Box plot is designed by identifying the following descriptive statistics:

1. Lower quartile (1st Quartile), median and upper quartile (3rd Quartile).
2. Lowest and highest value.
3. Inter-quartile range (IQR).

The box plot is constructed using IQR, minimum and maximum values. The box plot for the data in Table 2.4 is shown in Figure 2.11.

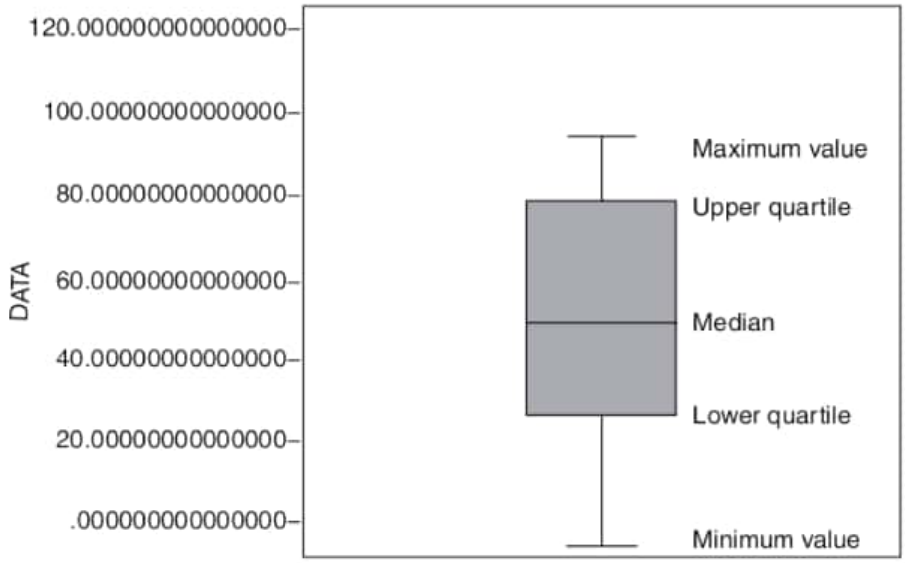


FIGURE 2.11 Box plot of the data in Table 12.3.

The length of the box is equivalent to IQR. It is possible that the data may contain values beyond Q_1 – 1.5 IQR and Q_3 + 1.5 IQR. The whisker of the box plot extends till Q_1 – 1.5 IQR (or minimum value) and Q_3 + 1.5 IQR (or maximum value); observations beyond these two limits are potential outliers. The box plot for the Bollywood movie budget is shown in Figure 2.12.

In Figure 2.12 position of the lowest whisker is 2 (since that is the minimum value). The value of lower quartile is 11 (lower line of the box), median is 24 (middle line in the box), and top quartile is 35 (upper line of the box). The top whisker is at Q_3 + 1.5 IQR = 71. All the observations beyond Q_3 + 1.5 IQR shown above the upper whisker are outliers.

2.9.7 | Treemap

Treemap is a hierarchical map made up of nested rectangles frequently used as part of business intelligence reports which helps organizations to understand the data hierarchically. To construct a treemap, the data should be hierarchical with several levels. The size of rectangle and colour are used for describing/differentiating the characteristics of the data. A sample Treemap is shown in Figure 2.13 in which

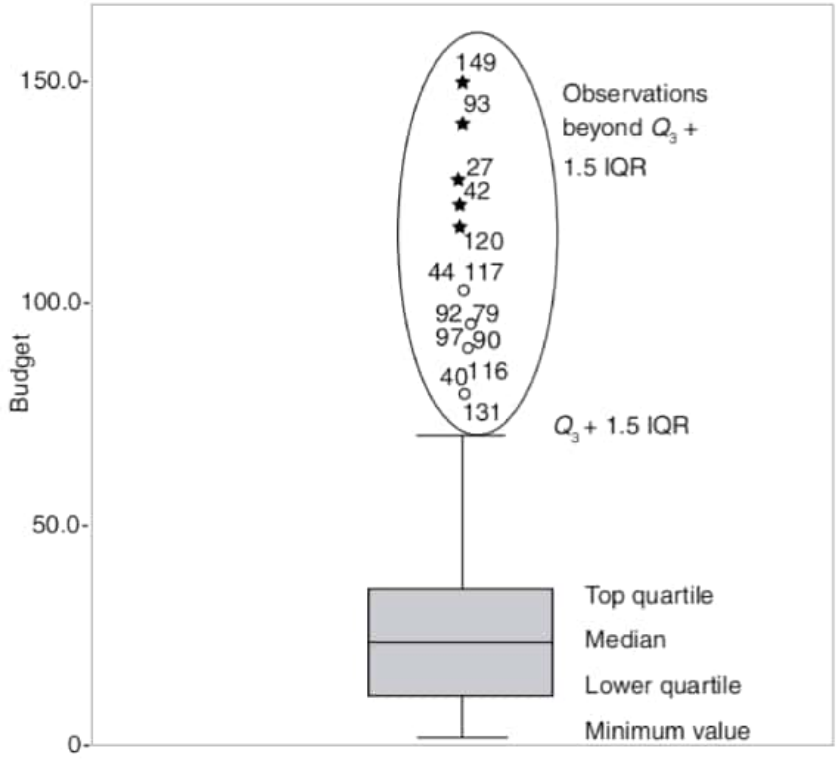


FIGURE 2.12 Box plot for Bollywood movie budget.

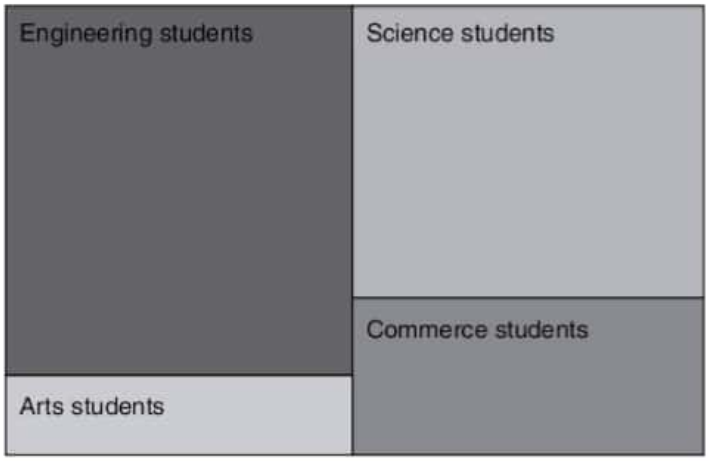


FIGURE 2.13 Treemap of student discipline at the undergraduate level.

the academic discipline at undergraduate level of students admitted into an MBA program is captured. Size of the area captures proportion of the students from that discipline. That is, in Figure 2.13 the area corresponding to engineering students is the largest indicating that the largest proportion of students come from engineering background and area corresponding to Arts students is the least indicating least number of students with arts background in the MBA program.

Each of the disciplines can be further analysed. For example, the engineering students can be further grouped according to the type of college (Tier 1, Tier 2, etc.).

EXAMPLE 3.4

Black boxes used in aircrafts are manufactured by three companies A, B and C. 75% are manufactured by A, 15% by B, and 10% by C. The defect rates of black boxes manufactured by A, B, and C are 4%, 6%, and 8%, respectively. If a black box tested randomly is found to be defective, what is the probability that it is manufactured by company A?

Solution:

Let $P(A)$, $P(B)$, $P(C)$ be events corresponding to the black box being manufactured by companies A, B, and C, respectively, and $P(D)$ be the probability of defective black box. We are interested in calculating the probability $P(A|D)$.

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

Now $P(D|A) = 0.04$ and $P(A) = 0.75$. Using Eq. (3.14):

$$P(D) = 0.75 \times 0.04 + 0.15 \times 0.06 + 0.10 \times 0.08 = 0.047$$

So

$$P(A \mid D) = \frac{0.04 \times 0.75}{0.047} = 0.6382$$

3.6| RANDOM VARIABLES

Random variable is a function that maps every outcome in the sample space to a real number. Random variables provide robustness required while developing probabilistic models since the outcome of a random experiment may be recorded in different format. Outcomes of experiments may be recorded in numerical and non-numerical terms. For example, consider bank transactions that are classified as either genuine (G) or fraud (F). Assume that the bank looks at last four transactions; the sample space in this case can be written as $S = \{GGGG, GGGF, GGFG, \ldots\}$. However, depending on the size of the bank, on any given day the number of transactions may run into several millions. Also, the bank would like to know the number of fraudulent transactions than the actual sequence of occurrence of fraudulent and genuine transactions. So, we need a variable that measures the number of fraudulent transactions. For every random experiment, we can define a function that maps the outcome to a real number. Random variable is defined as

> A function that assigns a real number to each sample point in the sample space S.

The sample space discussed earlier $S = \{GGGG, GGGF, GGFG, \ldots\}$ will be mapped to a real number set $S = \{0, 1, 2, 3, 4\}$, which is often the interest of the analyst. In the set $S = \{0, 1, 2, 3, 4\}$ the value represents the number of fraudulent transactions out of 4 transactions. Random variable is a robust and convenient way of representing the outcome of a random experiment. When the outcomes themselves are already expressed in terms of real numbers, it is possible to reassign a unique real number (or the original number itself) to the outcome. Random variables are usually denoted using capital letters, such as X, Y, and Z, whereas small letters, such as x, y, z, a, b, c, and so on, are used to denote particular values

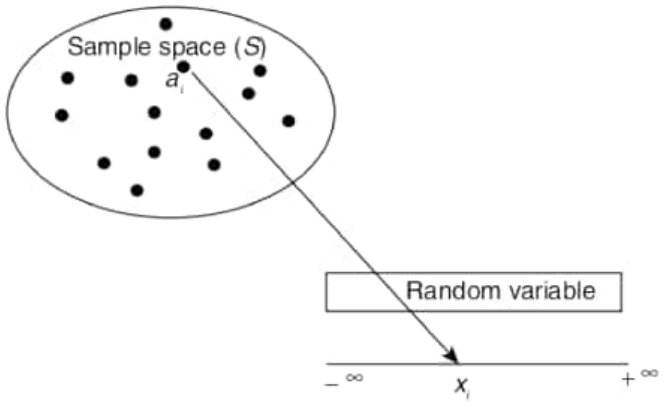


FIGURE 3.3 Random variable as a mapping from sample space to real number space.

of random variables written as $P(X = x)$. Figure 3.3 shows the mapping of the outcome of a random experiment to the real number. Once a random variable is defined, then the probability of events can be measured for various values that the random variable can take. For example, in the case of fraudulent transactions, we can now calculate probabilities such as

1. $P(X = 2)$, probability that the number of fraudulent transactions are exactly two.
2. $P(X > 2)$, probability that the number of fraudulent transactions are more than two.
3. $P(X < 2)$, probability that the number of fraudulent transactions are less than two.

Use of random variables provides us the flexibility required in modelling.

Random variables can be classified as discrete or continuous depending on the values that the random variable can take.

3.6.1| Discrete Random Variables

If the random variable X can assume only a finite or countably infinite set of values, then it is called a discrete random variable. There are very many situations where the random variable X can assume only finite or countably infinite set of values. Examples of discrete random variables are:

1. Credit rating (usually classified into different categories such as low, medium and high or using labels such as AAA, AA, A, BBB, etc.).
2. Number of orders received at an e-commerce retailer which can be countably infinite.
3. Customer churn [the random variables take binary values: (a) Churn and (b) Do not churn].
4. Fraud [the random variables take binary values: (a) Fraudulent transaction and (b) Genuine transaction].
5. Any experiment that involves counting (for example, number of returns in a day from customers of e-commerce portals such as Amazon, Flipkart; number of customers not accepting job offers from an organization).

In analytics, classification problems, an important class of problems, is an example of discrete random variable.

3.6.2 | Continuous Random Variables

A random variable X which can take a value from an infinite set of values is called a continuous random variable. Examples of continuous random variables are listed below:

1. Market share of a company (which take any value from an infinite set of values between 0 and 100%).
2. Percentage of attrition among employees of an organization.
3. Time to failure of engineering systems.
4. Time taken to complete an order placed at an e-commerce portal.
5. Time taken to resolve a customer complaint at call and service centers.

In many situations, a continuous variable may be converted to a discrete random variable for modelling purpose.

3.6.3 | Probability Mass Function and Cumulative Distribution Function of a Discrete Random Variable

For a discrete random variable, the probability that a random variable X taking a specific value x_i, $P(X = x_i)$, is called the probability mass function $P(x_i)$. That is, a probability mass function is a function that maps each outcome of a random experiment to a probability (Figure 3.4).

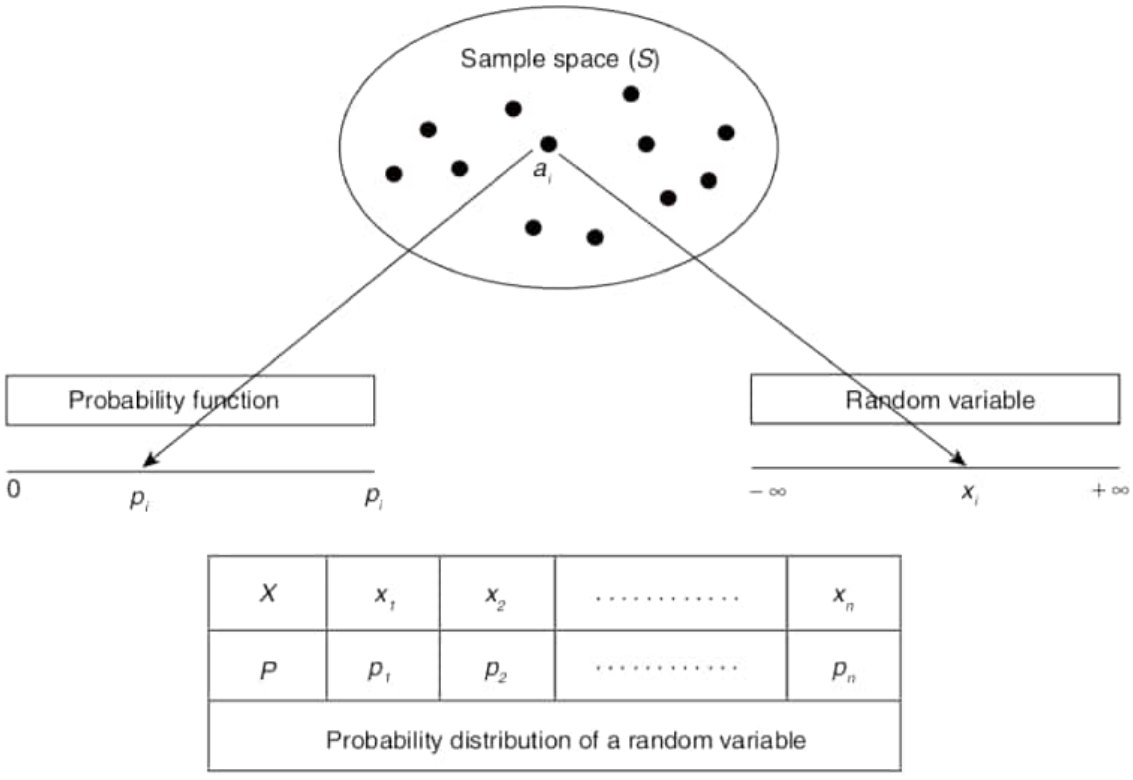


FIGURE 3.4 Probability mass function.

Consider the number of daily fraudulent transactions at a bank branch and the corresponding probabilities as described in Table 3.3. The values in Table 3.3 denote possible values of the random variable and the corresponding probability. That is, Table 3.3 describes how probability is distributed across different values of the random variable.

TABLE 3.3 Probability mass function

Random Variable X (X = number of fraudulent transactions)	$x_i = 0$	$x_i = 1$	$x_i = 2$	$x_i = 3$	$x_i = 4$
$P(X = x_i)$	0.20	0.15	0.25	0.25	0.15

From Table 3.3, we have the following information:

Probability that there will no fraudulent transaction on any given day, $P(X = 0) = 0.20$. Similarly, $P(X = 1) = 0.15$, $P(X = 2) = 0.25$, $P(X = 3) = 0.25$ and $P(X = 4) = 0.15$.

The probability mass function, $P(x_i)$ satisfies the following conditions:

1. $P(x_i) \geq 0$.
2. $\sum_{x_i} P(x_i) = 1$

Cumulative distribution function, $F(x_i)$, is the probability that the random variable X takes values less than or equal x_i. That is, $F(x_i) = P(X \leq x_i)$.

Based on the values given in Table 3.3,

$$F(2) = P(X \leq 2) = P(X = 0) + P(X = 1) + P(X = 2) = 0.60$$

3.6.4 | Expected Value, Variance, and Standard Deviation of a Discrete Random Variable

Expected value (or mean) of a discrete random variable is given by

$$E(X) = \sum_{i=1}^{n} x_i P(x_i) \tag{3.15}$$

where x_i is the specific value taken by a discrete random variable X and $P(x_i)$ is the corresponding probability, that is, $P(X = x_i)$. Expected value of a discrete random variable plays a crucial role in many contexts. For example, expected monetary value (EMV) forms the basis for selecting an alternative from several possible alternatives in a decision tree approach. EMV is calculated based on expected value.

Variance of a discrete random variable is given by

$$\text{Var}(X) = \sum_{i=1}^{n} \left[x_i - E(X)\right]^2 \times P(x_i) \tag{3.16}$$

Standard deviation of a discrete random variable is given by

$$\sigma = \sqrt{\text{Var}(X)} \tag{3.17}$$

For the data in Table 3.2, the expected value of the number of fraudulent transactions is given by

$$E(X)=\sum_x x_i P(x_i)=0\times 0.2+1\times 0.15+2\times 0.25+3\times 0.25+4\times 0.15=2$$

That is, the average number of fraudulent transactions is 2.

The variance of the random variable X is given by

$$\begin{aligned}\text{Var}(X)&=\sum_{i=1}^{n}[x_i-E(X)]^2\times P(x_i)\\&=(0-2)^2\times 0.2+(1-2)^2\times 0.15+(2-2)^2\times 0.25+(3-2)^2\times 0.25+(4-2)^2\times 0.15\\&=1.8\end{aligned}$$

The standard deviation, $\sigma=\sqrt{\text{Var}(X)}=\sqrt{1.8}=1.34$

3.7 | PROBABILITY DENSITY FUNCTION (PDF) AND CUMULATIVE DISTRIBUTION FUNCTION (CDF) OF A CONTINUOUS RANDOM VARIABLE

Since continuous random variables can take infinitely many values, their exact measurement is very difficult; even atomic clock comes with an error, although infinitesimally small. For this reason, the probability density function, $f(x_i)$, is defined as probability that the value of random variable X lies between an infinitesimally small interval defined by x_i and $x_i + \delta x$ and its mathematical expression is

$$f(x)=\lim_{\delta x\to 0}\frac{P(x_i\le X\le x_i+\delta x)}{\delta x} \tag{3.18}$$

Probability density function reflects how dense is the likelihood of a continuous random variable X taking a value in an infinitesimally small interval around value x. The cumulative distribution function (CDF) of a continuous random variable is defined by

$$F(a)=P(X\le a)=\int_{-\infty}^{a} f(x)dx \tag{3.19}$$

Cumulative distribution function $F(a)$ is the area under the probability density function (Figure 3.5) up to $X = a$.

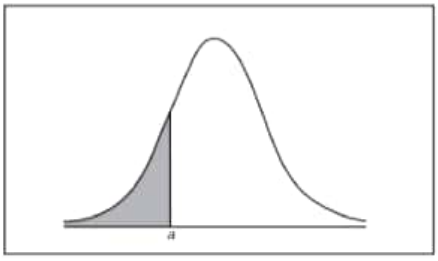


FIGURE 3.5 Cumulative distribution function $F(a)$.

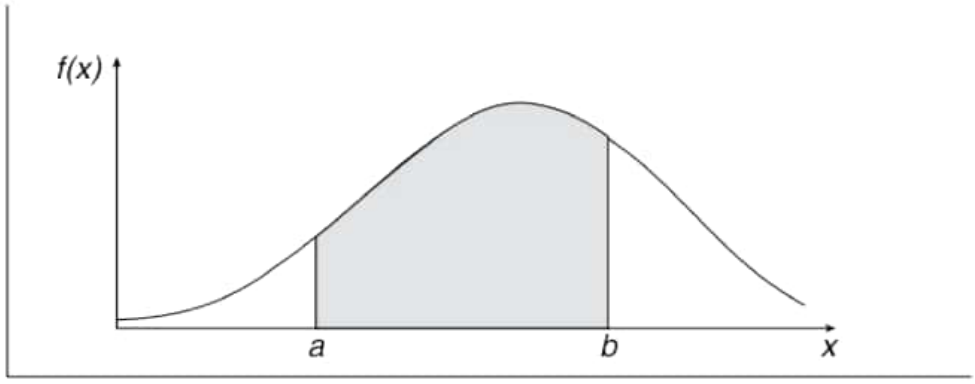


FIGURE 3.6 Area between values (*a*, *b*) under probability density function.

Probability density function and cumulative distribution function of a continuous random variable satisfy the following properties:

1. $f(x) \geq 0$
2. $F(\infty) = \int_{-\infty}^{+\infty} f(x)dx = 1$
3. $P(a \leq X \leq b) = \int_{a}^{b} f(x)dx = F(b) - F(a)$

The probability between two values a and b, $P(a \leq X \leq b)$, is the area between the values a and b under the probability density function (Figure 3.6).
The expected value of a continuous random variable, $E(X)$, is given by

$$E(X) = \int_{-\infty}^{+\infty} xf(x)dx \tag{3.20}$$

The variance of a continuous random variable, Var(X), is given by

$$\text{Var}(X) = \int_{-\infty}^{\infty} \left[x - E(X)\right]^2 f(x)dx \tag{3.21}$$

3.8 | BINOMIAL DISTRIBUTION

Binomial distribution is one of the most important discrete probability distribution due to its applications in several contexts. A random variable X is said to follow a Binomial distribution when

1. The random variable can have only two outcomes *success* and *failure* (also known as Bernoulli trials).
2. The objective is to find the probability of getting k successes out of n trials.
3. The probability of success is p and thus the probability of failure is $(1 - p)$.
4. The probability p is constant and does not change between trials.

Solution:

(a) The value of scale parameter $\lambda = (1/20)$. The probability the customer has to wait for less than 5 minutes is given by

$$F(5) = 1 - e^{-\frac{1}{20} \times 5} = 0.2211$$

(b) If a customer has been waiting for 20 minutes, the probability that the customer will wait for additional 20 minutes is given by

$$P(X > 20 + 20 \mid X > 20) = P(X > 20) = e^{-\left(\frac{1}{20}\right) \times 20} = e^{-1} = 0.3678$$

In the above equation, we have used the memoryless property of exponential distribution.

3.14 | NORMAL DISTRIBUTION

Normal distribution, also known as **Gaussian distribution**, is one of the most popular continuous distribution in the field of analytics especially due to its use in multiple contexts. Normal distribution is observed across many naturally occurring measures such as birth weight, height, intelligence, etc. The probability density function and the cumulative distribution function are given by

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}, \quad -\infty < x < +\infty \tag{3.41}$$

$$F(x) = \int_{-\infty}^{x} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{t-\mu}{\sigma}\right)^2} dt, \quad -\infty < x < +\infty \tag{3.42}$$

Here μ and σ are the mean and standard deviation of the normal distribution. Normal distribution with mean μ and standard deviation σ is denoted as $N(\mu, \sigma^2)$. Normal distribution is defined between $-\infty$ and $+\infty$. For a normal distribution, μ is the location parameter, which locates (center) the distribution on the horizontal axis and σ is the scale parameter, which defines the spread of the normal distribution. Normal distribution has no shape parameter since all normal distribution curves have bell shape and are symmetrical. Normal density curve and cumulative distribution curve are shown in Figures 3.15 and 3.16. In Microsoft Excel, the functions NORM.DIST(x, μ, σ, false) and NORM.DIST(x, μ, σ, true) can be used for calculating the probability density function and cumulative distribution function of a normal distribution with mean μ and standard deviation σ.

Historically, normal distribution was used in quantifying measurement errors associated with astronomical objects (Stahl, 2006). The error here is measured from the mean value and thus often called an *error function*. The curve can be also interpreted as 'small errors are more frequent than large errors'. No closed form solution exists for the cumulative distribution function of a normal distribution. Many functions that are numerical approximations are used for finding the value of cumulative distribution function of normal distribution.

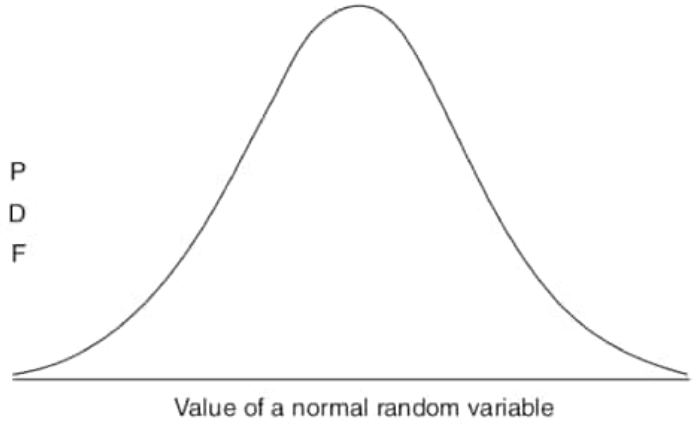


FIGURE 3.15 Probability density function of a normal distribution.

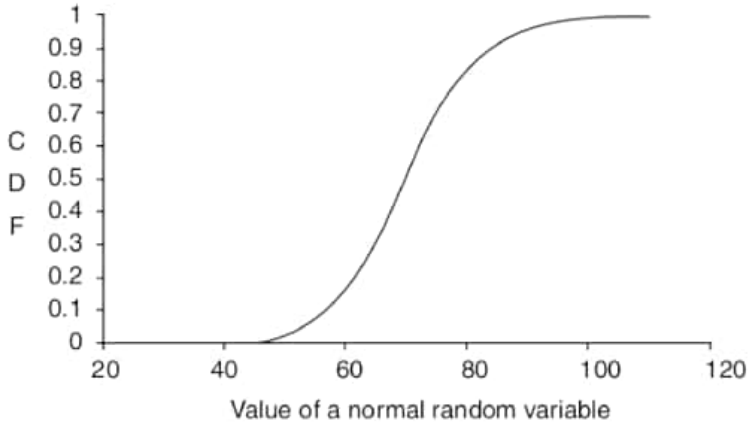


FIGURE 3.16 Cumulative distribution function of a normal distribution.

3.14.1 | Properties of Normal Distribution

1. Theoretical normal density functions are defined between $-\infty$ and $+\infty$.
2. It is a two parameter distribution, where the parameter μ is the mean (location parameter) and the parameter σ is the standard deviation (scale parameter).
3. All normal distributions have symmetrical bell shape around mean μ (thus it is also median). μ is also the mode of the normal distribution, that is, μ is the mean, median as well as the mode.
4. For any normal distribution, the areas between specific values measured in terms of μ and σ are given by:

Value of Random Variable	Area under the Normal Distribution (CDF)
$\mu-\sigma \le X \le \mu+\sigma$ (area between one sigma from the mean)	0.6828
$\mu-2\sigma \le X \le \mu+2\sigma$ (area between two sigma from the mean)	0.9545
$\mu-3\sigma \le X \le \mu+3\sigma$ (area between three sigma from the mean)	0.9973

5. Any linear transformation of a normal random variable is also normal random variable. That is, if X is a normal random variable, then the linear transformation $AX + B$ (where A and B are two constants) is also a normal random variable.
6. If X_1 and X_2 are two independent normal random variables with mean μ_1 and μ_2 and variance σ_1^2 and σ_2^2, respectively, then $X_1 + X_2$ is also a normal distribution with mean $\mu_1 + \mu_2$ and variance $\sigma_1^2 + \sigma_2^2$.
7. Sampling distribution of mean values of a large sample drawn from a population of any distribution is likely to follow a normal distribution. This result is known as the *central limit theorem* and will be discussed in detail in Chapter 4.

3.14.2 | Standard Normal Variable

A normal random variable with mean $\mu = 0$ and $\sigma = 1$ is called the standard normal variable and usually represented by Z. The probability density function and cumulative distribution function of a standard normal variable are given by

$$f(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} \tag{3.43}$$

$$F(z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx \tag{3.44}$$

By using the following transformation, any normal random variable X can be converted into a standard normal variable:

$$Z = \frac{X - \mu}{\sigma} \tag{3.45}$$

The random variable X can be written in the form of a standard normal random variable using the relationship

$$X = \mu + \sigma Z \tag{3.46}$$

Thus, any normal random variable X can be expressed using the standard normal random variable Z. No closed form solution exists for the cumulative standard normal distribution; however, there are several approximate formulas available for calculating CDF of a standard normal distribution (Yerukala and Boiroju, 2015). A simple approximation of standard normal CDF is given by Tocher (1963) as follows:

$$P(Z \le z) = F(z) \approx \frac{e^{2kz}}{1 + e^{2kz}} \tag{3.47}$$

where $k = \sqrt{2/\pi}$.

Another more accurate approximation is provided by Byrc (2002):

$$P(Z \le z) = F(z) = 1 - \left(\frac{z^2 + A_1 z + A_2}{\sqrt{2\pi} \times z^3 + B_1 z^2 + B_2 z + 2A_2} \right) \times e^{-z^2/2} \tag{3.48}$$

where A_1, A_2, B_1, and B_2 are constants given by

$$A_1 = 5.575192695;\ A_2 = 12.77436324;\ B_1 = 14.38718147;\ B_2 = 31.53531977$$

Note that, any normal random variable can be converted into standard normal random variable and the above approximations can be used for finding the CDF value.

In Microsoft Excel®, the normal PDF value is given by the function Normdist(x, μ, σ, false) and CDF value is given by Normdist(x, μ, σ, true), where x is the value of the normal random variable, μ is the mean of the distribution, and σ is the corresponding standard deviation. The CDF value of a standard normal distribution can be obtained using the function NORMSDIST(Z) in Microsoft Excel.

EXAMPLE 3.12

According to a survey on use of smart phones in India, the smart phone users spend 68 minutes in a day on average in sending messages and the corresponding standard deviation is 12 minutes. Assume that the time spent in sending messages follows a normal distribution.

(a) What proportion of the smart phone users are spending more than 90 minutes in sending messages daily?
(b) What proportion of customers are spending less than 20 minutes?
(c) What proportion of customers are spending between 50 minutes and 100 minutes?

Solution:

It is given that $\mu = 68$ minutes and $\sigma = 12$ minutes.

(a) Proportion of customers spending more than 90 minutes is given by

$$P(X \geq 90) = 1 - P(X \leq 90) = 1 - F(90)$$

The standard normal random variable value for $X = 90$ is given by

$$Z = \frac{x - \mu}{\sigma} = \frac{90 - 68}{12} = 1.8333$$

That is, $F(X = 90) = F(Z = 1.8333)$. From standard normal distribution table, we can get the value of $F(Z)$ for $Z = 1.8333$. The area under the standard normal distribution curve for $Z = 1.8333$ is 0.9666. Thus,

$$P(X \geq 90) = 1 - P(X \leq 90) = 1 - F(90) = 1 - 0.9666 = 0.0334$$

Alternatively, using Excel, we get

$$P(X \geq 90) = 1 - P(X \leq 90) = 1 - \text{Normdist}(90, 68, 12, \text{true}) = 0.0334$$

(b) Proportion of customers spending less than 20 minutes is

$$P(X \leq 20) = F(20)$$

Using Excel function, we have Normdist(20, 68, 12, true) = 3.1671×10^{-5}

(c) Proportion of customers spending between 50 and 100 minutes is given by

$$\begin{aligned} P(50 \leq X \leq 100) &= F(100) - F(50) \\ &= \text{Normdist}(100, 68, 12, \text{true}) - \text{Normdist}(50, 68, 12, \text{true}) \\ &= 0.9293 \end{aligned}$$

EXAMPLE 3.13

At Die Another Day (DAD) hospital, nurses are given an additional bonus of INR 1,00,000 if they stay for more than 36 months with DAD hospital. The average stay of nurses follows a normal distribution with an average of 28 months and the corresponding standard deviation is 4.8 months. Calculate

(a) The expected number of nurses who will be given bonus and the value of bonus that will be given if 50 new nurses join DAD hospital in the current month,

(b) What will be the additional amount paid if DAD hospital changes the policy that they will give bonus if the stay exceeds 24 months? What assumptions are made in this case?

Solution:

a) Expected number of nurses and the value of bonus:

Expected number of nurses who will be getting bonus = $50 \times P(X \geq 36)$

$$P(X \geq 36) = 1 - \text{Normdist}(36, 28, 4.8, \text{true}) = 0.04779$$

Expected number of nurses who will be getting bonus = $50 \times 0.04799 = 2.389518$

Expected value of bonus given = $50 \times P(X \geq 36) \times 100{,}000$ = INR 238951.76

(b) The additional bonus given is

$$50 \times 1{,}00{,}000 \times \left[\text{Normdist}(36, 28, 4.8, \text{true}) - \text{Normdist}(24, 28, 4.8, \text{true})\right] = 3749406$$

The major assumption here is that the policy change is unlikely to change the attrition behaviour of the nurses, which may not be true. Since the nurses now know that if they stay for 24 months, they will get the bonus, the distribution parameter values are likely to change.

Sampling and Estimation

"To Clarify *add* data."
— Edward R Tufte

LEARNING OBJECTIVES

- **LO4-1** Understand the need for sampling and the importance of appropriate sampling.
- **LO4-2** Understand the difference between population parameters and sample statistic.
- **LO4-3** Learn different types of sampling techniques and limitations of each sampling approach.
- **LO4-4** Learn about estimation of parameters and sampling distribution.
- **LO4-5** Understand Central Limit Theorem (CLT) and its importance in hypothesis testing.
- **LO4-6** Learn method of moments and maximum likelihood estimator (MLE) and its applications in estimation of parameters of probability distributions.

ESSENCE OF SAMPLING

Sampling is a process of selecting subset of observations from a population to make inference about various population parameters such as mean, proportion, standard deviation, etc. It is an important step in inferential statistics since an incorrect sample may lead to wrong inference about the population. Sampling process itself has several steps and each step is important to ensure that the ideal sample is used for estimation of population parameters and for making inferences about the population. Under Big Data context, we may use almost the entire population; however, in most cases we will still be dependent on samples to make inference.

IMPORTANT

Sampling is necessary when it is difficult or expensive to collect data on the entire population. The inference about the population is made based on the sample that was collected; incorrect sample may lead to incorrect inference about the population.

4.1 | INTRODUCTION TO SAMPLING

The population of India was close to 1.32 billion in July 2016 according to United Nations (Source: Wikipedia[1]). Census India collects information such as demography, literacy, housing, economic activity of the individuals, etc. Thousands of people are used for collecting data on such huge population,

[1] Source: https://en.wikipedia.org/wiki/Demographics_of_India

which is very expensive and thus carried out only once in 10 years. Many organizations cannot afford to collect data on the entire population and in many cases it is expensive and time-consuming. In many cases, all members of the population are not known for sampling purpose. For example, assume that an organization is interested to conduct a study on diabetic patients. According to Lancet, the estimated diabetic patients in India in 2014 were 64.5 million (Mascarenhas, 2016). It is impossible to collect data from all diabetic patients, since many diabetic patients themselves may not be aware that they are diabetic, especially during early stages. Even when all the members of the population are known, collecting data can be expensive and thus in many cases we have to settle for samples. Sampling is necessary because even when the entire population is available, using the entire population for estimation of a population parameter may not be feasible. Consider reliability estimation of engineering systems, the original equipment manufacturer (OEM) may carry out some destructive testing (in which the item is destroyed in the test, for example crash testing of cars) to understand the causes of failure and estimate the population parameter such as mean time between failure (MTBF); for obvious reasons OEM cannot use the entire population of systems for destructive testing.

Sampling is an important activity in analytics, especially when inferences are made about the population based on the sample. Incorrect sampling can result in incorrect estimation of population parameters and wrong inference about the population. One of the frequently used examples to demonstrate the importance of sampling is the opinion poll conducted by the Literary Digest in 1936 USA presidential election (Squire, 1988). Alfred Landon, the Republican Governor of Kansas, was contesting against the incumbent Franklin D Roosevelt. Literary Digest predicted that Landon would get 55%, Roosevelt 41%, and Lemke 4% votes. However, the actual votes polled were 61% for Roosevelt and 37% for Landon, resulting in a prediction error of 20% for Roosevelt (Squire, 1988). The main reason for such a huge error was attributed to the sampling method employed by Literary Digest, in spite of the fact that their sample size was about 2.4 million. There were two major issues with the sampling processes used by Literary Digest. The first one was the selection of voters for the poll (sampling frame); the names were taken from telephone directories, subscribers of the magazines, club members, automobile registry, etc. (Squire, 1988). In 1936, this meant that they were selecting middle and upper middle class voters since telephone ownership was a rarity in 1936. Thus, the sampling framework used had an in-built selection bias of individuals for the purpose of study. The second was that they contacted close to 10 million voters; however, only 2.4 million responded (which in itself is a very large sample). But such low response can result in non-responsive bias. Cahalan (1989) reported that based on his survey 67% Roosevelt supporters and 52% Landon supporters claimed that they had not received Literary Digest ballot. Literary Digest went bankrupt after this prediction. The election result was correctly predicted by George Gallup, a pioneer of survey sampling with just 50,000 samples (Squire, 1988). This example clearly demonstrates the importance of selection of items in the sample and a large sample does not improve prediction if the sampling process is incorrect. Even in the age of analytics and big data, the winner of the 2016 US presidential election was incorrectly predicted by media. Most media outfits predicted comfortable win for Ms Hillary Clinton. Most of these incorrect predictions can be attributed to incorrect sampling procedures (mostly biased data collection and even biased reporting).

4.1.1 | When Jesus Christ Became 2nd Best in the World

Respondent bias is another source problem in survey sampling. There was an internet poll conducted in 1998 to find the 'most influential figure of the last 2000 years'. Jamie Pollock who played as defensive midfielder for the football team Manchester City won the title leaving Jesus Christ and Carl Marx to second and third place, respectively (Szczepanik, 2016). Apparently the poll was rigged by the supporters of the club Queens Park Rangers (QPR) who voted multiple times for Jamie Pollock. The story goes like this: On 25th April 1998, Manchester City was playing against QPR in Division 1 of English football and both teams were in danger of being relegated to division 2 (third tier of English football). When the goal was 1–1, Jamie Pollock scored an own goal giving upper hand to QPR. Although the match ended in a 2–2 draw, Manchester City was relegated to Division 2 for the first time and QPR managed stay in Division 1 (Moxley 2012, Szczepanik 2016). Collecting survey and ensuring that the sample is unbiased is one of the major challenges in analytics. Summers (1969) grouped bias in survey research in to several categories such as (a) sampling bias, (b) non-responsive bias, (c) respondent bias, (d) instrumentation bias, etc.

4.2 | POPULATION PARAMETERS AND SAMPLE STATISTIC

In many real-life problems, the population can be very large making it impossible to collect every feature of each case in the population. Measures such as mean and standard deviation calculated using the entire population are called *population parameters*. The population parameters mean and standard deviation are usually denoted using symbols μ and σ, respectively. Since calculating population parameters in most practical situations is almost impossible, we depend on samples to estimate the population parameters. Population parameters estimated from sample are called *sample statistic* or *statistic*. The sample statistic is denoted using symbols $\bar{X}$ (for mean) and S (or s for standard deviation). Statisticians also use hat symbol ($\hat{X}$ for mean and $\hat{\sigma}$ for standard deviation) for statistic. Since inferences about population are made using a sample, statistic plays an important role in hypothesis testing.

In addition to sample mean and standard deviation, we frequently estimate population proportion (p), which is proportion of cases in the data belonging to a specific category. Assume that a bank classifies its customers based on the risk categories: (1) Low, (2) Medium, and (3) High. We would like to know what proportions of the population belong to categories 1, 2, and 3 which are denoted by p_1, p_2, and p_3, respectively. The corresponding estimates will be denoted by $\hat{p}_1, \hat{p}_2$ and $\hat{p}_3$. Assume that n_1, n_2, and n_3 are the number of cases under categories 1 (low risk), 2 (medium risk), and 3 (high risk), respectively. The estimates of proportions are given by

$$\hat{p}_1 = \frac{n_1}{n_1 + n_2 + n_3}, \quad \hat{p}_2 = \frac{n_2}{n_1 + n_2 + n_3} \text{ and } \hat{p}_3 = \frac{n_3}{n_1 + n_2 + n_3}$$

As discussed in the case of 1936 American presidential election, the sample selection plays an important role in unbiased estimate of the proportions. The same applies to estimation of any parameter, such as scale, shape, and location parameters of probability distributions. Later in the chapter, we will discuss the method of moments and maximum likelihood estimation which can be used for estimating parameters of probability distribution.

4.3 | SAMPLING

The process of identifying a subset from a population of elements (aka observations or cases) is called sampling process or simply sampling. The following steps are used in any sampling process:

1. **Identification of target population that is important for a given problem under study.** For example, assume that we are interested in studying attrition among young professionals in India. The definition 'young professionals' in India is vague; we need a clear identification of the target population. A better definition of the population in this case would be to study the attrition among IT professionals in the age group 25–35 years in India. It is important to clearly define the target population for correct inference.
2. **Decide the sampling frame.** Sampling frame defines the source (or method/procedure) used for identifying the elements of the target population. Choice of sampling frame is important for accuracy of the study. Literary Digest used the telephone directory as one of the sampling frame which turned out to be an incorrect sampling frame. One may use more than one sampling frame (Literary Digest did use more than one sampling frame). Sampling framework will also include features of individual entities. One of the challenges at this stage of sampling process is that in many situations, sampling frame itself may not exist (Kiregyera, 1982). In such cases, the researcher has to define sampling frame using which individual data may be collected. To analyse attrition among IT professionals, sources such as LinkedIn and job portals Naukri and Monster can be used. However, these frames may not have important variables (features) that are required such as information related to salary and other data captured during exit interview. So, ideally to understand the attrition behaviour one has to use the data captured by many human resource departments across multiple companies.
3. **Determine the sample size:** Determining sample size for data collection is important since collecting data can be expensive and at the same time insufficient sample results in lack of precision in estimation of the parameters. The sample size for analytics projects is determined using factors such as effect size, standard deviation, desired level of confidence, and margin of error. We will discuss the formula for calculating the sample size in section 4.8. An important point to note here is that even in the days of big data in which many business contexts produce huge quantity of data, we still have several scenarios for which sufficient data may not be available (especially when the event itself is rare such as occurrence of Tsunami). Several rules of thumb are often used for determining sample size. For multivariate models such as multiple linear regression, logistic regression, and factor analysis, the thumb rule such as 10 times or 20 times the number of independent variables are used (Norman *et al.*, 2012). That is, if there are 10 variables, then a sample size of 200 in most cases would be acceptable (Norman *et al.*, 2012).
4. **Sampling method:** Sampling method is the technique used for selecting individual cases in the sample from the target population using the sampling frame. At a higher level, sampling method is classified into two major categories: **probabilistic sampling** and **non-probabilistic sampling**. Probabilistic sampling is further classified as random sampling, stratified sampling, etc. Bootstrap aggregating (also known as Bagging) and Boosting are two popular sampling methods used in machine learning algorithms.

4.4 | PROBABILISTIC SAMPLING

In a probability sampling, the individual observations in the sample are selected according to a probability distribution. Assume that the population has a total of N cases, and we are interested in creating a sample of size n. There are ${}^{N}C_{n}$ $[= N!/\{n! \times (N-n)!\}]$ different ways for creating such a sample. For example, if $N = 100$ and $n = 30$, there will more than 2.93×10^{25} possible samples. Based on how each case in the sample is selected forms the basis of different sampling methods.

4.4.1 | Random Sampling

Random sampling is one of the most popular and frequently used sampling methods. Shewhart (1931) defines random sample as a 'sample drawn under conditions such that the law of large number applies'. That is, in random sampling, every case in the population has equal probability of getting selected in a sample. Random sampling is usually carried out **without replacement**, that is, an observation which is selected in the sample is removed from the population for subsequent selection. However, random samples can also be created **with replacement**, that is, an observation which is selected for inclusion in the sample can again be considered since it is replaced (not removed) in the population.

Selection of cases in a sample can be implemented using several procedures. The easiest one is to label all cases in the population sequentially and generate uniform random numbers (integers) to select the cases from the population. For example, assume that the population has 10 cases (patients and their length of stay measured in days for treatment at a hospital) as shown in Table 4.1.

Now, to generate a sample of size 5 cases, we can generate 5 integer random numbers that follow uniform distribution between 1 and 10 and use those cases in the sample. Uniform random integer numbers between two values can be generated using Microsoft Excel function RANDBETWEEN(lower value, upper value). Table 4.2 shows the random numbers generated using RANDBETWEEN(1, 10) and the corresponding samples (length of stay of patients selected in the sample).

One disadvantage of the above procedure is that the random numbers are likely to repeat, and thus are not ideal for generating samples without replacement. For sampling without replacement, one may generate one case in the sample sequentially, removing and reordering the population at each step. Simple random sampling is used when the population is homogenous.

TABLE 4.1 Patients and length of stay (LoS) in days

Patient	1	2	3	4	5	6	7	8	9	10
LoS	4	20	12	13	15	17	16	20	9	17

TABLE 4.2 Random sample of size 5 using uniform random numbers

Random Numbers					Corresponding Sample (LoS value)				
3	4	5	1	8	12	13	15	4	20
1	7	9	1	3	4	16	9	4	12
8	4	7	3	5	20	13	16	12	15

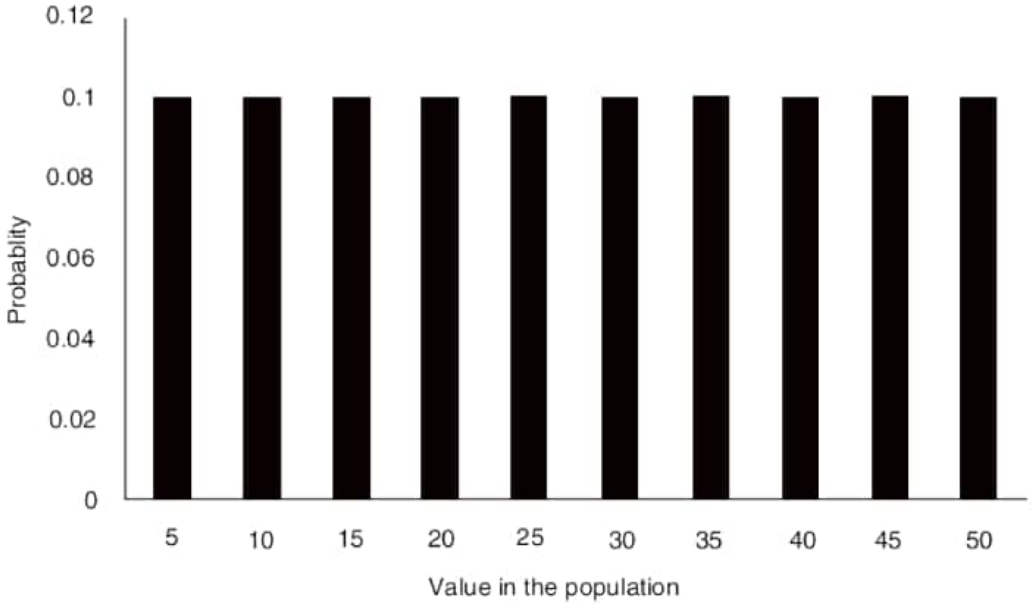


FIGURE 4.1 Probability density function of the population data provided in Table 4.3.

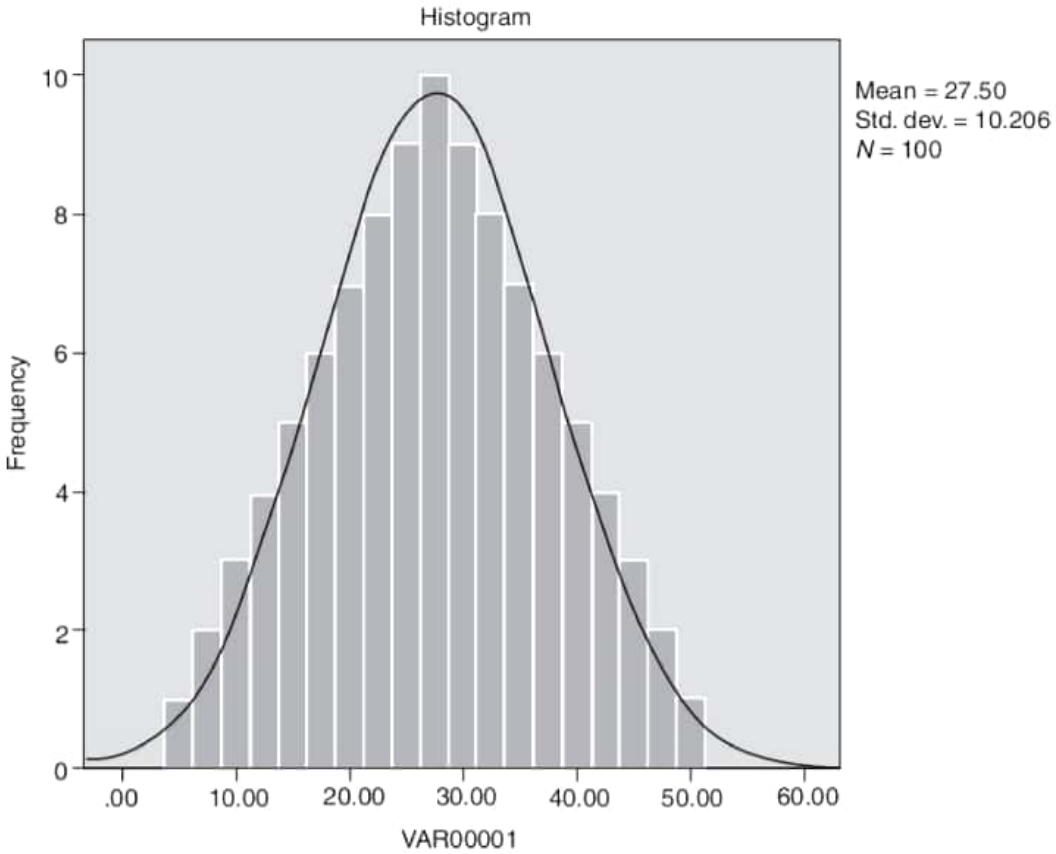


FIGURE 4.2 Histogram of sampling distribution of means.

The histogram of sample mean of all samples of size 2 is shown in Figure 4.2. It is interesting to note that the probability density function of the sampling distribution follows a normal distribution and its mean is 27.5, exactly same as the mean of the population data in Table 4.3.

4.7 | CENTRAL LIMIT THEOREM (CLT)

Central limit theorem is one of the most important theorems in statistics due to its applications in testing of hypothesis. Let $S_1, S_2, \ldots, S_k$ be samples of size n drawn from an independent and identically distributed population with mean μ and standard deviation σ. Let $\bar{X}_1, \bar{X}_2, \ldots, \bar{X}_k$ be the sample means (of the samples $S_1, S_2, \ldots, S_k$). According to the CLT, the distribution of $\bar{X}_1, \bar{X}_2, \ldots, \bar{X}_k$ follows a normal distribution with mean μ and standard deviation $\sigma / \sqrt{n}$ for large value of n. That is, the sampling distribution of mean will follow a normal distribution with mean μ (same as the mean of the population) and standard deviation $\sigma / \sqrt{n}$. The use of the term central limit theorem dates back to the work by George Polya (Fischer, 2010). Several proofs exist for central limit theorem starting from a proof by Laplace in 18th century and further works by Poisson and Cauchy (Fischer, 2010).

In simple terms, central limit theorem states that for a large sample drawn from a population with mean μ and standard deviation σ, the sampling distribution of mean, $\bar{X}$, follows an approximate normal distribution with mean μ and standard deviation (standard error) $\sigma / \sqrt{n}$ irrespective of the distribution of the population (Thomas, 1984).

Alternative version of CLT can be stated as follows:

Let $X_1, X_2, \ldots, X_n$ be n random variables that are independent and identically distributed with mean μ and standard deviation σ. Then for large n, mean

$$\bar{X} = \frac{X_1 + X_2 + \ldots + X_n}{n}$$

follows a normal distribution with mean μ stand error $\sigma / \sqrt{n}$. Independent and identical distribution (IID) implies that the random variables are mutually independent and the random variables follow the same probability distribution.

Implications of central limit theorem:

1. The variable $\frac{X - \mu}{\sigma / \sqrt{n}}$ will be a standard normal distribution (mean = 0, standard error = 1).
2. If $S_n = X_1 + X_2 + \ldots + X_n$, then $E(S_n) = n\mu$ and Standard error is $\sigma\sqrt{n}$. The random variable $\frac{S_n - n\mu}{\sigma\sqrt{n}}$ is a standard normal variate.
3. Regardless of the population distribution, the sampling distribution of large sample ($n > 30$) will follow the normal distribution with mean same as population mean and standard error $\sigma / \sqrt{n}$.

IMPORTANT

Central limit theorem is the basis for hypothesis tests such as Z-test and t-test. In many cases, we will have access to only a sample and the inference about the population has to be made based on sample statistic.

IMPORTANT

An important assumption of CLT is that the random variables have to be independent and identically distributed.

4.7.1 | Central Limit Theorem for Proportions

The central limit theorem for proportion is stated as follows:

If we have a population in which a characteristic (for example, smart phone users) has a proportion of p, then the sampling distribution of the proportion (that is $\hat{p}$ calculated from several samples of size n) will follow a normal distribution with mean p and standard deviation $\sqrt{p(1-p)/n}$.

Central limit theorem for proportions can be stated as follow:

If $X_1, X_2, \ldots, X_n$ are counts from a Bernoulli trials with probability of success p, $E(X_i) = p$ and $\text{Var}(X_i) = p \times (1 - p)$, then the sampling distribution of probability of success (say $\hat{p}$) follows an approximate normal distribution with mean p and standard error $\sqrt{p(1-p)/n}$, where n is the sample size. The variable $\dfrac{\hat{p} - p}{\sqrt{p(1-p)/n}}$ converges to a standard normal distribution.

EXAMPLE 4.1

It is believed that college students in Bangalore spend on average 80 minutes daily on texting using their mobile phones and the corresponding standard deviation is 25 minutes. Data from a sample of 100 students were collected for calculating the amount of time spent in texting. Calculate the probability that the average time spent by this sample of students will exceed 84 minutes.

Solution:

Using the central limit theorem, the mean of the sampling distribution is 80 and the corresponding standard deviation is $25/\sqrt{100} = 2.5$. The probability that the sample average is more than 84 minutes is given by

$$P\left(Z > \frac{84-80}{2.5}\right) = P(Z > 1.6) = 0.05479$$

EXAMPLE 4.2

The value of insurance claims received at an insurance company follows exponential distribution with mean INR 4200. If a sample of 100 claims is taken from the population, calculate the probability that the total claim will exceed INR 5,00,000.

Solution:

According to CLT, the summation of random variables follows a normal distribution with mean $n\mu$ and standard error $\sigma\sqrt{n}$. Note that for an exponential distribution mean and standard deviation are same

The probability that the total claim will exceed INR 5,00,000 is

$$P\left(Z > \frac{5,00,000 - n\mu}{\sigma\sqrt{100}}\right)$$

In this case $n = 100$, $\mu = \sigma = 4200$, and Z is the standard normal variate. So

$$P(Z > 5,00,000) = P\left(Z > \frac{5,00,000 - 100 \times 4200}{4200 \times \sqrt{100}}\right) = P(Z > 1.90476) = 0.02841$$

That is, there is 2.8% chance that the total claim will exceed INR 5,00,000.

4.8 | SAMPLE SIZE ESTIMATION FOR MEAN OF THE POPULATION

From the central limit theorem, we know that the sampling distribution of mean follows a normal distribution with mean μ and standard deviation $\sigma / \sqrt{n}$. Then the standard normal variate of the sampling distribution of mean is given by

$$Z = \frac{\bar{X} - \mu}{\sigma / \sqrt{n}} \tag{4.1}$$

Note that the difference between the sample mean and the population mean, $\bar{X} - \mu$, is error in estimation of the population mean. Equation (4.1) can be written as

$$n = \left[\frac{Z_{\alpha/2} \times \sigma}{D}\right]^2 \tag{4.2}$$

where $Z_{\alpha/2}$ known as the critical value is the value of Z for which the area under standard normal distribution is $\alpha/2$ (that is $F(Z) = \alpha/2$) or $(1 - \alpha)$ is the desired confidence level in estimating the population mean and $D = \bar{X} - \mu$ is the error in estimating the population mean. For example, if $\alpha = 0.05$, then we have 95% confidence that the error is less than D when the sample size n is as given by Eq. (4.2). For $\alpha = 0.05$, the value of $Z_{\alpha/2} = Z_{0.025} = -1.96$ or $|Z_{\alpha/2}| = 1.96$. Note that the formula can be used only when the standard deviation is known.

Confidence Intervals

"Confidence comes not from always being right but from not fearing to be wrong ".

— **Peter McIntyre**

LEARNING OBJECTIVES

- **LO5-1** Learn the difference between point estimate and interval estimate. Understand the need for interval estimate.
- **LO5-2** Understand the concept of confidence interval and confidence level.
- **LO5-3** Learn to calculate confidence interval for population mean when population standard deviation is either known or unknown.
- **LO5-4** Understand confidence interval for population proportion and variance.
- **LO5-5** Gain insights from confidence interval and confidence level.

CONFIDENCE INTERVALS

When there is an uncertainty around measuring the value of an important population parameter, it is advisable to find the range in which the value of the parameter is likely to fall rather than predicting a single estimate (point estimate). Confidence interval is the range in which the value of a population parameter is likely to lie with certain probability. Confidence interval provides additional information about the population parameter that will be useful in decision making.

The objective of confidence interval is to provide both location and precision of population parameters.

5.1 | INTRODUCTION TO CONFIDENCE INTERVAL

We estimate population parameters such as mean, proportion, standard deviation, scale, shape, and location parameters of the probability distribution from a sample using techniques such as method of moments and maximum likelihood estimation (MLE). Point estimate obtained through techniques such as methods of moments and MLE is a unique value. The quality of estimated parameter values is

measured using factors such as biasness, consistency, and efficiency (discussed in Chapter 4). The accuracy of the point estimate of population parameters is very difficult to establish; hence we prefer *interval estimate* over point estimate. An interval estimate is defined as follows:

> An interval estimate of a population parameter such as mean and standard deviation is an interval or range of values within which the true parameter value is likely to lie with certain probability.

The interval estimate is stated between two values. For example, confidence interval for population mean may be stated as $30 \leq \mu \leq 50$ (that is, the population mean lies between values 30 and 50). The interval estimate may or may not contain the true parameter values. Thus, we associate a confidence (probability) with interval estimate that predicts the probability of finding true parameter value in the interval. For example, we may state that there is a 95% confidence that the interval contains the population mean. 95% confidence also implies that there is a 5% chance that the interval may not contain the actual population mean. Depending on the context of the problem, one may increase confidence level to 98% or 99%. Confidence level is defined as follows:

> Confidence level, usually written as $(1 - \alpha)100\%$, on the interval estimate of a population parameter is the probability that the interval estimate will contain the true population parameter. When $\alpha = 0.05$, 95% is the confidence level and 0.95 is the probability that the interval estimate will have the population parameter.

The value of α is called *significance*. The value of α signifies that the chance of not observing the true population mean in the interval estimate is 1 out of 20. Alternatively, 95% confidence implies that in 19 out of 20 cases, the true population mean will be within the interval estimate. The confidence interval is defined as follows:

> Confidence interval is the interval estimate of the population parameter estimated from a sample using a specified confidence level.

95% is most frequently used confidence level, although 90% and 99% are also used frequently. The choice of α depends on the context of the problem. When high accuracy for the estimate is required, low value of α is chosen.

5.2 | CONFIDENCE INTERVAL FOR POPULATION MEAN

Let $X_1, X_2, \ldots, X_n$ be the sample means of samples $S_1, S_2, \ldots, S_n$ that are drawn from an independent and identically distributed population with mean μ and standard deviation σ. From central limit theorem we know that the sample means X_i follow a normal distribution with mean μ and standard deviation $\sigma / \sqrt{n}$. The variable $Z = \dfrac{X_i - \mu}{\sigma / \sqrt{n}}$ follows a standard normal variable.

Equation (5.2) is valid for large sample sizes, irrespective of the distribution of the population. Equation (5.2) is equivalent to

$$P(\bar{X} - Z_{\alpha/2} \times \sigma / \sqrt{n} \le \mu \le \bar{X} + Z_{\alpha/2} \times \sigma / \sqrt{n}) = 1 - \alpha \quad (5.3)$$

That is, the probability that the population mean takes a value between $\bar{X} - Z_{\alpha/2} \times \sigma / \sqrt{n}$ and $\bar{X} + Z_{\alpha/2} \times \sigma / \sqrt{n}$ is 1 – α. The absolute values of $Z_{\alpha/2}$ for various values of α are shown in Table 5.1.

TABLE 5.1 Value of $|Z_{\alpha/2}|$ for different values of α

α	$\lvert Z_{\alpha/2} \rvert$	Confidence interval for population mean when population standard deviation is known
0.1	1.64	$\bar{X} \pm 1.64 \times \sigma / \sqrt{n}$
0.05	1.96	$\bar{X} \pm 1.96 \times \sigma / \sqrt{n}$
0.02	2.33	$\bar{X} \pm 2.33 \times \sigma / \sqrt{n}$
0.01	2.58	$\bar{X} \pm 2.58 \times \sigma / \sqrt{n}$

EXAMPLE 5.1

A sample of 100 patients was chosen to estimate the length of stay (LoS) at a hospital. The sample mean was 4.5 days and the population standard deviation was known to be 1.2 days.

(a) Calculate the 95% confidence interval for the population mean.

(b) What is the probability that the population mean is greater than 4.73 days?

Solution:

(a) 95% confidence interval for population mean: We know that $\bar{X}$ = 4.5 and σ = 1.2 and thus $\sigma / \sqrt{n} = 1.2 / \sqrt{100} = 0.12$.

The 95% confidence interval is given by

$$(\bar{X} - Z_{\alpha/2} \times \sigma / \sqrt{n},\ \bar{X} + Z_{\alpha/2} \times \sigma / \sqrt{n}) = (4.5 - 1.96 \times 0.12,\ 4.5 + 1.96 \times 0.12)$$
$$= (4.2648,\ 4.7352)$$

The Excel function CONFIDENCE(α, σ, n) [or CONFIDENCE.NORM(α, σ, n)], where α is the significance, σ is the population standard deviation, and n is the sample size, returns the value $Z_{\alpha/2} \times \sigma / \sqrt{n}$. For current problem CONFIDENCE(0.05, 1.2, 100) = 0.235196. The corresponding confidence interval is

$$(4.5 - 0.235196,\ 4.5 + 0.235196) = (4.2648,\ 4.7352)$$

(b) Note that 4.73 is the upper limit of the 95% confidence interval from part (a), thus the probability that the population mean is greater than 4.73 is approximately 0.025.

EXAMPLE 5.2

Amount of time (measured in hours) spent by 20 students on an online course is given in Table 5.2. Assuming that the population of time spent follows a normal distribution and standard deviation is 3.1 hours, calculate the 90% confidence interval for the mean time spent by the students.

TABLE 5.2 Sample time spent by students on an online course

4.7	9.3	8	7.4	9.2	1.7	7.2	8.6	9	6.9
9.2	11.2	7.6	4.9	5.3	2.8	12.3	10.6	5.7	3.8

Solution: The estimate mean from the sample is $\bar{X} = 7.27$ and the sampling distribution's standard deviation is $\sigma / \sqrt{n} = 3.1 / \sqrt{20} = 0.6932$.
The 90% confidence interval is given by

$$(\bar{X} - Z_{\alpha/2} \times \sigma / \sqrt{n},\ \bar{X} + Z_{\alpha/2} \times \sigma / \sqrt{n}) = (7.27 - 1.64 \times 0.6932,\ 7.27 + 1.64 \times 0.6932)$$
$$= (6.1332,\ 8.4068)$$

5.3 | CONFIDENCE INTERVAL FOR POPULATION PROPORTION

The central limit theorem for population proportion is stated as follows:

If $X_1, X_2, \ldots, X_n$ are from Bernoulli trials with probability of success p, that is, $E(X_i) = p$ and $\text{Var}(X_i) = p \times q$ (where $q = 1 - p$), then the sampling distribution of probability of success (say $\hat{p}$) for a large sample size follows an approximate normal distribution with mean p and standard error $\sqrt{pq/n}$, where n is the sample size. The variable $\dfrac{\hat{p} - p}{\sqrt{pq/n}}$ converges to a standard normal distribution. Note that the standard deviation of the sampling distribution of proportions depends on the value of p which is unknown. However, for large sample size, the estimate value $\hat{p}$ will converge to the actual value p. As a rule of thumb, we set the value of n such that $n \times p \times q \geq 10$ (few authors suggest that $n \times p \times q$ should be at least 15).

The $(1 - \alpha)100\%$ confidence interval for population proportion p is given by

$$\hat{p} - Z_{\alpha/2}\sqrt{\frac{\hat{p} \times \hat{q}}{n}} \leq p \leq \hat{p} + Z_{\alpha/2}\sqrt{\frac{\hat{p} \times \hat{q}}{n}} \tag{5.4}$$

EXAMPLE 5.3

A retail store was interested in finding the proportion of customers who pay through cash (as against credit or debit card) for the merchandize they buy at the store. From a sample of 100 customers, it was found that 70 customers paid by cash. Calculate the 95% confidence interval for proportion of customers who pay by cash.

Solution: In this case, $n = 100$, $\hat{P} = 70/100 = 0.7$ and $\hat{q} = 1 - \hat{p} = 0.3$. Since $n \times \hat{p} \times \hat{q} = 100 \times 0.7 \times 0.3 = 21 \geq 10$, we can use the confidence interval equation provided in Eq. (5.4).

$$\hat{p} - Z_{\alpha/2}\sqrt{\frac{\hat{p} \times \hat{q}}{n}} \leq p \leq \hat{p} + Z_{\alpha/2}\sqrt{\frac{\hat{p} \times \hat{q}}{n}}$$

$$\Rightarrow 0.7 - 1.96\sqrt{\frac{0.7 \times 0.3}{100}} \leq p \leq 0.7 + 1.96\sqrt{\frac{0.7 \times 0.3}{100}} = 0.6102 \leq p \leq 0.7898$$

That is, the 95% confidence interval for p is (0.6102, 0.7898). That is, we are 95% confident that the interval (0.6102, 0.7898) contains the true population proportion of the customers who pay by cash.

5.4 | CONFIDENCE INTERVAL FOR POPULATION MEAN WHEN STANDARD DEVIATION IS UNKNOWN

When the standard deviation of the population is unknown then we will not be able to use the formula stated in Eq. (5.2). William Gossett (Student, 1908) proved that if the population follows a normal distribution and the standard deviation is calculated from the sample, then the statistic given in Eq. (5.5) will follow a t-distribution with $(n-1)$ degrees of freedom.

$$t = \frac{\bar{X} - \mu}{S/\sqrt{n}} \tag{5.5}$$

Here S is the standard deviation estimated from the sample (standard error). The t-distribution is very similar to standard normal distribution; it has a bell shape and its mean, median, and mode are equal to zero as in the case of standard normal distribution. The major difference between the t-distribution and the standard normal distribution is that t-distribution has broad tail compared to standard normal distribution. However, as the degrees of freedom increases the t-distribution converges to standard normal distribution.

The $(1 - \alpha)100\%$ confidence interval for mean from a population that follows normal distribution when the population mean is unknown is given by

$$\bar{X} \mp t_{\alpha/2, n-1} \times \frac{S}{\sqrt{n}} \tag{5.6}$$

In Eq. (5.6), the value $t_{\alpha/2,n-1}$ is the value of t under t-distribution for which the cumulative probability $F(t) = 0.025$ when the degrees of freedom is $(n-1)$. Here the degrees of freedom is $(n-1)$ since standard deviation is estimated from the sample. The absolute values of $t_{\alpha/2,n-1}$ for different values of α are shown in Table 5.3 along with corresponding $Z_{\alpha/2}$ values.

TABLE 5.3 Values of $|t_{\alpha/2,n-1}|$ and $|Z_{\alpha/2}|$ for different degrees of freedom (*df*)

α	$\lvert t_{\alpha/2,10}\rvert$	$\lvert t_{\alpha/2,50}\rvert$	$\lvert t_{\alpha/2,500}\rvert$	$\lvert Z_{\alpha/2}\rvert$
0.1	1.812	1.675	1.647	1.64
0.05	2.228	2.008	1.964	1.96
0.02	2.763	2.403	2.333	2.33
0.01	3.169	2.677	2.585	2.58

It is evident from Table 5.3 that the values of $t_{\alpha/2,n-1}$ and $Z_{\alpha/2}$ converge for higher degrees of freedom. In fact, as the sample size nears 100, the *t*-distribution gets very close to a normal distribution. The values of $t_{\alpha/2,n-1}$ can be obtained using the function T.INV(α/2, $n-1$) in Microsoft Excel [another excel function T.INV.2T(α, $n-1$) also returns $t_{\alpha/2,n-1}$ value]. In Excel, TINV(α, $n-1$) returns critical values for two-tailed test (concept of two tailed test will be discussed in Chapter 6). If we have to calculate one-tailed critical value at significance α, then the corresponding Excel function is TINV(2α, $n-1$). Note that, different versions of Microsoft Excel has different functions to calculate inverse value of *t* distribution.

EXAMPLE 5.4

An online grocery store is interested in estimating the basket size (number of items ordered by the customer) of its customer order so that it can optimize its size of crates used for delivering the grocery items. From a sample of 70 customers, the average basket size was estimated as 24 and the standard deviation estimated from the sample was 3.8. Calculate the 95% confidence interval for the basket size of the customer order.

Solution: We know that $n=70$, $\bar{X}=24$, $S=3.8$ and $t_{0.025,69}=1.995$ [using TINV(0.05, 69) in Microsoft Excel].
The confidence interval for size of basket using Eq. (5.6) is given by

$$\bar{X} \pm t_{\alpha/2,n-1}\frac{S}{\sqrt{n}} = 24 \pm 1.995\frac{3.8}{\sqrt{70}} = 24 \pm 0.9061$$

Thus the 95% confidence interval for the size of the basket is (23.09, 24.91).

5.5 | CONFIDENCE INTERVAL FOR POPULATION VARIANCE

Let S_1^2, S_2^2, ..., S_k^2 be the sample variance estimated from samples of size *n* drawn from a normal distribution with variance σ^2. Then the random variable defined by

$$\frac{(n-1)\times S_i^2}{\sigma^2} \tag{5.7}$$

follows a χ^2-distribution with $(n-1)$ degrees of freedom. Note that Eq. (5.7) is valid only when the samples are drawn from a normal population; it is not valid otherwise. We can use Eq. (5.7) to derive confidence interval for variance when the samples are drawn from a normal distribution. The $(1-\alpha)100\%$ confidence interval for variance, σ^2, is given by (Tate and Klett, 1959 and Cohen, 1972)

$$\left[\frac{(n-1)S^2}{\chi^2_{\alpha/2,n-1}}, \quad \frac{(n-1)S^2}{\chi^2_{1-\alpha/2,n-1}}\right] \tag{5.8}$$

The confidence interval for standard deviation, σ, is given by

$$\left[\sqrt{\frac{(n-1)S^2}{\chi^2_{\alpha/2,n-1}}}, \quad \sqrt{\frac{(n-1)S^2}{\chi^2_{1-\alpha/2,n-1}}}\right] \tag{5.9}$$

where $\chi^2_{\alpha/2,n-1}$ is the value of chi-square distribution with $n-1$ degrees of freedom where $\alpha/2$ is the right side area, $\chi^2_{1-\alpha/2,n-1}$ is the value of chi-square distribution with $n-1$ degrees of freedom where $1-\alpha/2$ is the right side area.

EXAMPLE 5.5

Time taken to manufacture an aircraft door is a random variable due to several manual processes and assembly of more than 1000 parts to make the aircraft door. The sources of variability in door assembly include factors such as non-availability of parts, manpower, and machine tools. It is known that the time to assemble a door follows a normal distribution. The variance of the time taken to manufacture the door was estimated to be 324 hours based on a sample of 50 doors. Calculate a 95% confidence interval for the variance in manufacturing aircraft door.

Solution: We know that $n = 50$, $S^2 = 324$, $\chi^2_{0.025,49} = 70.22$, $\chi^2_{0.975,49} = 31.55$ [the value of χ^2 can be calculated using Microsoft Excel function CHIINV($\alpha/2$, df) or CHISQ.INV.RT($1-\alpha/2$, df)].

The 95% confidence interval for variance is given by

$$\left[\frac{(n-1)S^2}{\chi^2_{\alpha/2,n-1}}, \quad \frac{(n-1)S^2}{\chi^2_{1-\alpha/2,n-1}}\right] = \left[\frac{49\times 324}{70.22}, \quad \frac{49\times 324}{31.55}\right] = [226.09, \ 503.20]$$

The 95% confidence interval for standard deviation is [15.04, 22.43].

SUMMARY

1. The point estimate of population parameters give unique value, however, data scientists would like to know the range of values the population parameter is likely to take. Interval estimates provide better insights about the population parameter.
2. Confidence level $(1-\alpha)\times 100\%$ is the probability that the true population parameter value will lie within the confidence interval.
3. The choice of confidence level or significance α will depend on the context and accuracy required. For higher accuracy, the value of α will be lower.
4. Confidence intervals are derived using the central limit theorem of sampling distribution.

Hypothesis Testing

"Beware of the problem of testing too many hypotheses; the more you torture the data, the more likely they are to confess, but confessions obtained under duress may not be admissible in the court of scientific opinion."

—**Stephen M Stigler**

LEARNING OBJECTIVES

- **LO6-1** Understand hypothesis test and its importance in analytics.
- **LO6-2** Learn to setup a hypothesis test, understand the concept of null and alternative hypotheses.
- **LO6-3** Understand the link between central limit theorem and test statistic in one-sample *Z*-test and *t*-test.
- **LO6-4** Understand the concept of significance (α), probability value (*p*-value), Type I and Type II errors.
- **LO6-5** Understand simple one-sample hypothesis test for population mean when population variance is either known or unknown.
- **LO6-6** Learn to conduct a two-sample hypothesis test and its applications in analytics.
- **LO6-7** Understand the role of non-parametric tests such as chi-square test of independence.
- **LO6-8** Learn goodness of fit tests and their application in identifying best probability distribution to describe a data set.

HYPOTHESIS TESTING

Hypothesis testing is one of the most important concepts in analytics and also a concept which many students of statistics and analytics find it difficult to understand. Hypothesis is a claim made by a person/organization. The claim is usually about population parameters such as mean or proportion and we seek evidence from a sample for the support of the claim (for example, claim could be that the average salary of analytics experts is at least USD 1,00,000). Hypothesis testing is a process used for either rejecting or retaining a null hypothesis.

IMPORTANT

The objective of hypothesis testing is to either reject or retain a null hypothesis. In many cases, for example, in regression models, one would like to reject the null hypothesis to establish statistically significant relationship between the dependent and the independent variables. However, in goodness of fit tests, that are used for checking whether the data follows a specific distribution or not, we would like to retain the null hypothesis.

6.1 | INTRODUCTION TO HYPOTHESIS TESTING

6.1.1 | Blackout Babies

On 9 November 1965 there was a power failure that resulted in blackout for approximately 12 hours in New York and surrounding areas. Nine months later, in August 1966, New York Times published a series of three articles in which it claimed that the birth rates in August 1966 was higher than normal based on interviews with city doctors (Izenman and Zabell, 1981). The babies were nicknamed 'blackout babies'. The articles published by the New York Times raised an interesting question on whether power failures result in procreation? Izenman and Zabell (1981) using time series data analysis claimed that there is not enough evidence to suggest that the 1965 power failure resulted in increased birth rate nine months after the blackout. Many claims were made about the impact of power cuts on baby booms and mothers since then (Anon, 2009 and Fetzer *et al.*, 2013).

Hypothesis is a claim or belief, hypothesis testing is a statistical process of either rejecting or retaining a claim or belief or association related to a business context, product, service, processes, etc. Hypothesis testing consists of two complementary statements called **null hypothesis** and **alternative hypothesis**, and only one of them is true. Hypothesis testing is one of the most important concepts in analytics due to its role in inferential statistics. Hypothesis testing is an integral part of many predictive analytics techniques such as multiple linear regression and logistic regression. It plays an important role in providing evidence of an association relationship between an outcome variable and predictor variables.

In business, many claims are made by organizations. Few examples of such claims are listed below:

1. Children who drink the health drink Complan (a health drink owned by the company Heinz in India) are likely to grow taller.
2. If you drink Horlicks, you can grow taller, stronger, and sharper (3 in 1).
3. Using fair and lovely (fair and handsome) cream can make one fair and lovely (fair and handsome).
4. Wearing perfume (such as Axe) will help to attract opposite gender (known as Axe effect).
5. Women use camera phone more than men (Freier, 2016).
6. Beautiful people are likely to have girl child (Miller and Kanazawa, 2007). This is one of my favorite hypotheses since I have a daughter I can claim that I am good looking.
7. Married people are happier than singles (Anon, 2015), especially those who married their best friend (many married people may not agree!).
8. Vegetarians miss few flights (Siegel, 2016).
9. Smokers are better sales people.

There are many such claims and beliefs; many business rules and strategies are generated based on these hypotheses. The question is how can we check whether these are actually true. Hypothesis testing is used for checking the validity of the claim using evidence found in a sample data.

6.2 | SETTING UP A HYPOTHESIS TEST

In this section, we will discuss the steps involved in hypothesis testing. Data analysis in general can be classified as **exploratory data analysis** or **confirmatory data analysis**. In exploratory data analysis, the idea is to look for new or previously unknown hypothesis or suggest hypotheses. In the case of confirmatory data analysis, the objective is to test the validity of a hypothesis (confirm whether the hypothesis is true or not) using techniques such as hypothesis testing and regression. According to Tukey (1977), exploratory data analysis is similar to a detective work suggesting hypotheses whereas confirmatory data analysis looks for evidence in support of hypotheses using techniques such as hypothesis testing. The following steps are used in hypothesis testing:

1. Describe the hypothesis in words. Hypothesis is described using a population parameter (such as mean, standard deviation, proportion, etc.) about which a claim (hypothesis) is made. Few sample claims (hypothesis) are:

   (a) Average time spent by women using social media is more than men.
   (b) On average women upload more photos in social media than men.
   (c) Customers with more than one mobile handsets are more likely to churn.

2. Based on the claim made in step 1, define null and alternative hypotheses. Initially we believe that the null hypothesis is true. In general, null hypothesis means that there is no relationship between the two variables under consideration (for example, null hypothesis for the claim 'women use social media more than men' will be 'there is no relationship between gender and the average time spent in social media'). Null and alternative hypotheses are defined using a population parameter.
3. Identify the test statistic to be used for testing the validity of the null hypothesis. Test statistic will enable us to calculate the evidence in support of null hypothesis. The test statistic will depend on the probability distribution of the sampling distribution; for example, if the test is for mean value and the mean is calculated from a large sample and if the population standard deviation is known, then the sampling distribution will be a normal distribution and the test statistic will be a Z-statistic (standard normal statistic).
4. Decide the criteria for rejection and retention of null hypothesis. This is called **significance value** traditionally denoted by symbol α. The value of α will depend on the context and usually 0.1, 0.05, and 0.01 are used. Significance value α is the Type I error (discussed in Section 6.4).
5. Calculate the p-value (probability value), which is the conditional probability of observing the test statistic value when the null hypothesis is true. In simple terms, p-value is the evidence in support of the null hypothesis.
6. Take the decision to reject or retain the null hypothesis based on the p-value and significance value α. The null hypothesis is rejected when p-value is less than α and the null hypothesis is retained when p-value is greater than or equal to α.

6.2.1 | Description of Hypothesis

Hypotheses are claims that are usually stated in simple words initially as listed below:

1. Average annual salary of machine learning experts is different for males and females.
2. On an average people with Ph.D. in analytics earn more than people with Ph.D. in engineering.
3. The average box-office collection of comedy genre movies is more than that of action movies.
4. Average life of vegetarians is more than meat eaters.
5. Proportion of married people defaulting on loan repayment is less than proportion of singles defaulting on loan repayment.

6.2.2 | Null and Alternative Hypothesis

Null hypothesis, usually denoted as H_0 (H zero and H naught), refers to the statement that there is no relationship or no difference between different groups with respect to the value of a population parameter. Null hypothesis is the claim that is assumed to be true initially. That is at the beginning we assume that the null hypothesis is true and try to retain it unless there is strong evidence against null hypothesis.

Alternative hypothesis, usually denoted as H_A (or H_1), is the complement of null hypothesis. Alternative hypothesis is what the researcher believes to be true and would like to reject the null hypothesis.

The null and alternative hypotheses for the sample hypotheses stated in Section 6.2.1 are described in Table 6.1.

TABLE 6.1 Hypothesis statement to definition of null and alternative hypothesis

S. No.	Hypothesis Description	Null and Alternative Hypothesis
1	Average annual salary of machine learning experts is different for males and females. (In this case, the null hypothesis is that there is no difference in male and female salary of machine learning experts)	$H_0: \mu_m = \mu_f$ $H_A: \mu_m \neq \mu_f$ μ_m and μ_f are average annual salary of male and female machine learning experts, respectively.
2	On average people with Ph.D. in analytics earn more than people with Ph.D. in engineering.	$H_0: \mu_a \leq \mu_e$ $H_A: \mu_a > \mu_e$ μ_a = Average annual salary of people with Ph.D. in analytics. μ_e = Average annual salary of people with Ph.D. in engineering. It is essential to have the equal sign in null hypothesis statement.

IMPORTANT

Hypothesis test checks the validity of the null hypothesis based on the evidence from the sample. At the beginning of the test, we assume that the null hypothesis is true. Since the researcher may believe in alternative hypothesis, she/he may like to reject the null hypothesis. However, in many cases (such as goodness of fit tests), we would like to retain or fail to reject the null hypothesis.

6.2.3 | Test Statistic

Test statistic is the standardized difference between the estimated value of the parameter being tested calculated from the sample(s) and the hypothesis value (that is, standardized difference between $\bar{X}$ and μ in the case of testing mean) in order to establish the evidence in support of the null hypothesis. Test statistic is the standardized value used for calculating the p-value (probability value) in support of null hypothesis. Since test statistic is a standardized value, it measures the standardized distance (measured in terms of number of standard deviations) between the value of the parameter estimated from the sample(s) and the value of the null hypothesis.

The p-value is the conditional probability of observing the statistic value when the null hypothesis is true. For example, consider the following research hypothesis: Average annual salary of machine learning experts is at least 100,000. The corresponding null hypothesis is H_0: $\mu_m \leq 100{,}000$. Assume that estimated value of the salary from a sample is 1,10,000 (that is $\bar{X} = 1{,}10{,}000$) and assume that the standard deviation of population is known and standard error of the sampling distribution is 5000 (that is, $\sigma / \sqrt{n} = 5000$, where n is the sample size using which $\bar{X} = 1{,}10{,}000$ was calculated). The standardized distance between estimated salary from hypothesis salary is (1,10,000 – 1,00,000)/5000 = 2. That is, the standardized distance between estimated value and the hypothesis value is 2 and we can now find the probability of observing this statistic value from the sample if the null hypothesis is true (that is if $\mu_m \leq$ 100,000). A large standardized distance between the estimated value and the hypothesis value will result in a low p-value. Note that the value 2 is actually the value under a standard normal distribution since it is calculated from $\dfrac{\bar{X} - \mu}{\sigma / \sqrt{n}}$. Standard normal distribution and the p-value corresponding to $Z = 2$ are shown in Figure 6.1.

Probability of observing a value of 2 and higher from a standard normal distribution is 0.02275. That is, if the population mean is 1,00,000 and the standard error of the sampling distribution is 5000 then probability of observing a sample mean greater than or equal to 1,10,000 is 0.02275. The value 0.02275 is the p-value, which is the evidence in support of the statement in the null hypothesis.

$$p\text{-value} = P(\text{Observing test statistics value} \mid \text{null hypothesis is true}) \tag{6.1}$$

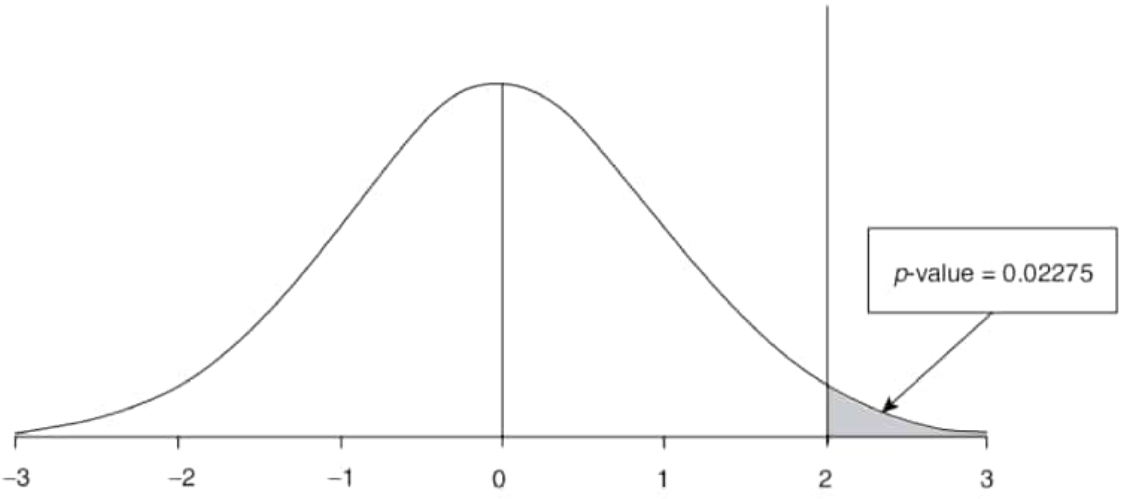


FIGURE 6.1 Standard normal distribution and p-value.

IMPORTANT

Note that the p-value is a conditional probability. It is the conditional probability of observing the statistic value given that the null hypothesis is true. P-value is the evidence in support of null hypothesis.

6.2.4 | Decision Criteria – Significance Value

Primary task in hypothesis testing is to take a decision to either reject or fail to reject (retain) the null hypothesis, thus we need a criteria to take the decision. Significance level, usually denoted by α, is the criteria used for taking the decision regarding the null hypothesis (reject or retain) based on the calculated p-value. The significance value α is the maximum threshold for p-value. The decision to reject or retain will depend on whether the calculated p-value crosses the threshold value α or not. The decision criteria is shown in Table 6.2.

The chosen value of α may depend on the context of the problem. Usually $\alpha = 0.05$ is used by researchers (recommended by Fisher, 1956); however, values such as 0.1, 0.02, and 0.01 are also frequently used. The value of α chosen is very low (0.05) for reason that we start the process of hypothesis testing with an assumption that null hypothesis is true. Unless there is strong evidence against this assumption, we will not reject the null hypothesis. The value of statistic in the sampling distribution for which the probability is α is called the **critical value**. In a right-tailed test, if the calculated statistic value is greater than the critical value (p-value will be less than α-value) then we reject the null hypothesis, whereas, if the statistic value is less than the critical value then we retain the null hypothesis. In case of left-tailed test, if the calculated statistic value is less than the critical value (p-value will be less than α-value) then we reject the null hypothesis, whereas, if the statistic value is greater than the critical value then we retain the null hypothesis. The areas beyond the critical values are known as **rejection region**.

IMPORTANT

The significance value α is the threshold conditional probability of rejecting a null hypothesis when it is true. It is the value of Type I error.

$$\text{Significance value } \alpha = P(\text{Rejecting a null hypothesis} \mid \text{null hypothesis is true}) \quad (6.2)$$

6.3 | ONE-TAILED AND TWO-TAILED TEST

Consider the following three hypotheses:

1. Salary of machine learning experts on average is at least US $100,000.
2. Average waiting time at the London Heathrow airport security check is less than 30 minutes.
3. Average annual salaries of male and female MBA students are different at the time of graduation.

TABLE 6.2 Decision making under hypothesis testing

Criteria	Decision
p-value $< \alpha$	Reject the null hypothesis
p-value $\geq \alpha$	Retain (or fail to reject) the null hypothesis

STATEMENT 1 Salary of machine learning experts on average is at least US $100,000:

The null and alternative hypotheses in this case are given by

$$H_0: \mu_m \leq 100{,}000$$

$$H_A: \mu_m > 100{,}000$$

where μ_m is the average annual salary of machine learning experts. Note that the equality symbol is always part of the null hypothesis since we have to measure the difference between estimated value from the sample and the hypothesis value. In this case, reject or retain decision will depend on the direction of deviation of the estimated parameter value from the hypothesis value. Figure 6.2 shows the rejection region on the right side of the distribution. Since the rejection region is only on one side this is a one-tailed test (right tailed test). Specifically, since the alternative hypothesis in this case is $\mu_m > 100{,}000$, this is called right-tailed test.

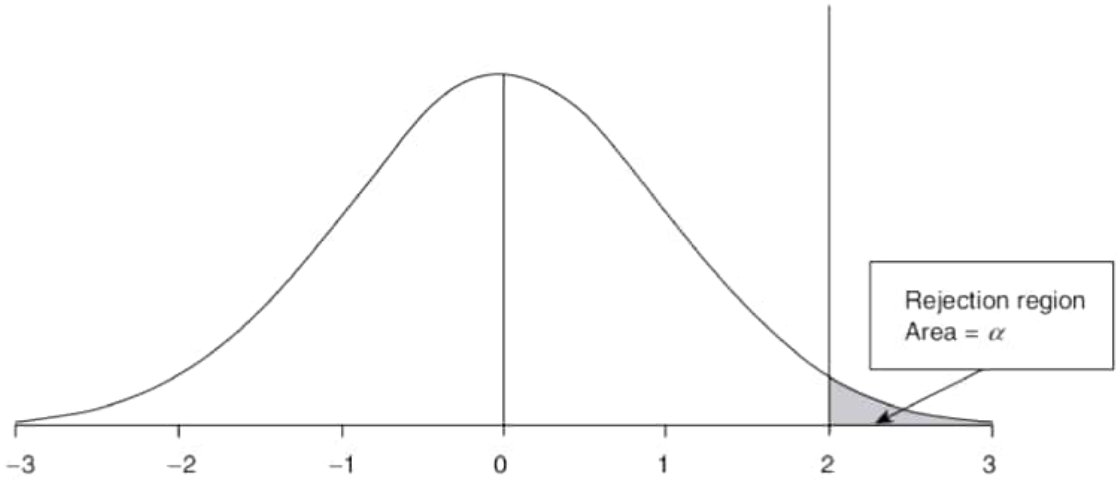


FIGURE 6.2 Right-tailed hypothesis test's rejection region.

STATEMENT 2 Average waiting time at the London Heathrow airport security check is less than 30 minutes:

The null and alternative hypotheses in this case are given by

$$H_0: \mu_w \geq 30$$
$$H_A: \mu_w < 30$$

where μ_w is the average waiting time at London Heathrow security check. In this case, reject region will be on the left side (known as left-tailed test) of the distribution as shown in Figure 6.3.

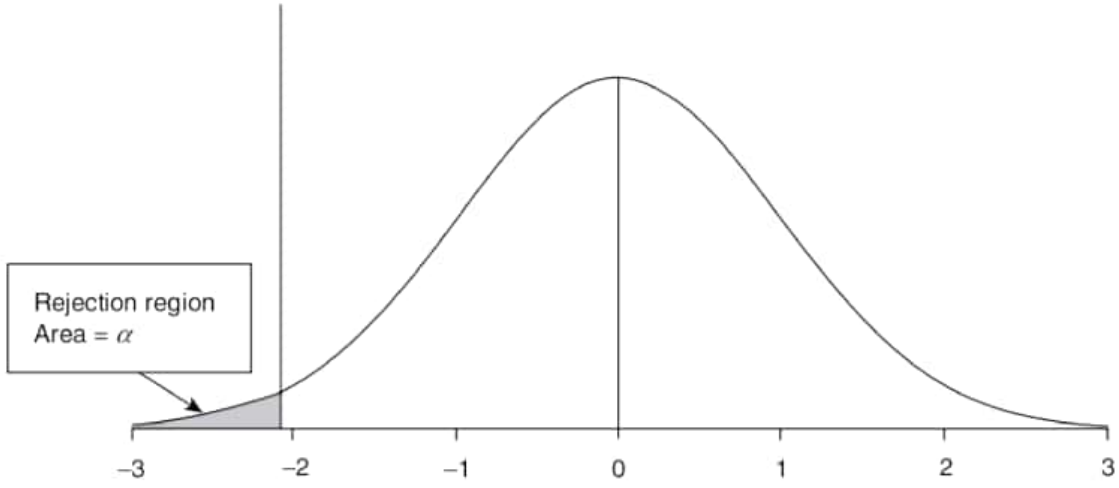


FIGURE 6.3 Rejection region in case of left-sided test.

STATEMENT 3 Average salary of male and female MBA students at graduation is different:

The null and alternative hypotheses in this case are given by

$$H_0: \mu_m = \mu_f$$

$$H_A: \mu_m \neq \mu_f$$

where μ_m and μ_f are the average salaries of male and female MBA students, respectively, at the time of graduation. In this case, the rejection region will be on either side of the distribution and if the significance level is α then the rejection region will be $\alpha/2$ on either side of the distribution. Since the rejection region is on either side of the distribution, it will be a two-tailed test. Figure 6.4 shows the rejection region of a two-tailed test.

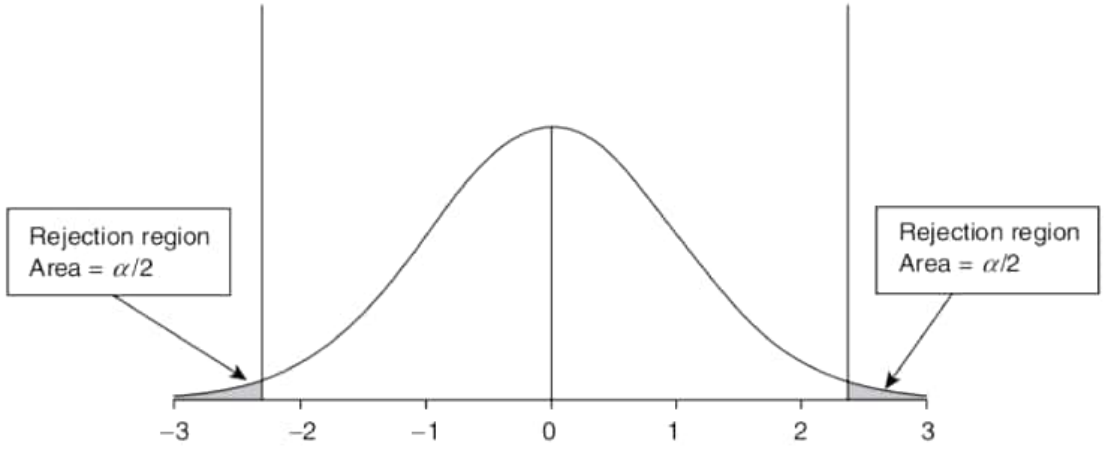


FIGURE 6.4 Rejection region in case of two-tailed test.

6.4 | TYPE I ERROR, TYPE II ERROR, AND POWER OF THE HYPOTHESIS TEST

In hypothesis test we end up with the following two decisions:

1. Reject null hypothesis.
2. Fail to reject (or retain) null hypothesis.

Type I and Type II errors are defined as follows:

1. **Type I Error:** Conditional probability of rejecting a null hypothesis when it is true is called Type I Error or False Positive (falsely believing that the claim made in alternative hypothesis is true). The significance value α is the value of Type I error. Mathematically, Type I error can be defined as follows:

$$\text{Type I Error} = \alpha = P(\text{Rejecting null hypothesis} \mid H_0 \text{ is true}) \tag{6.3}$$

It is important to understand the difference between the p-value and the significance value α. Probability value (p-value) is the evidence for the null hypothesis whereas significance value α is the error based on repetitive sampling. Hubbard *et al.* (2003) state that the p-value in a hypothesis test refers to probability of observing the data given a null hypothesis, whereas the significance level α refers to incorrect rejection of null hypothesis when it is true under **repeated trials**.

2. **Type II Error:** Conditional probability of failing to reject a null hypothesis (or retaining a null hypothesis) when the alternative hypothesis is true is called Type II Error or False Negative (falsely believing that there is no relationship). Usually Type II error is denoted by the symbol β. Mathematically, Type II error can be defined as follows:

$$\text{Type II Error} = \beta = P(\text{Retain null hypothesis} \mid H_0 \text{ is false}) \tag{6.4}$$

The value $(1 - \beta)$ is known as the power of hypothesis test. That is, the power of the test is given by

$$\text{Power of the test} = 1 - \beta = 1 - P(\text{Retain null hypothesis} \mid H_0 \text{ is false}) \tag{6.5}$$

Alternatively the power of test = $1 - \beta = P(\text{Reject null hypothesis} \mid H_0 \text{ is false})$

Description of Type 1 error, Type 2 error, and the power of test is given in Table 6.3.

6.5 | HYPOTHESIS TESTING FOR POPULATION MEAN WITH KNOWN VARIANCE: Z-TEST

Z-test (also known as one-sample Z-test) is used when a claim (hypothesis) is made about the population parameter such as population mean or proportion when population variance is known. In this section, we will be discussing the hypothesis testing for the population mean when the population variance is known. Since the hypothesis test is carried out with just one sample, this test is also known as **one-sample Z-test**. According to the central limit theorem (CLT) for sampling distribution of mean, we

TABLE 6.3 Description of type I error, type II error, and the power of test

Actual Value of H_0	Decision made about Null Hypothesis Based on the Hypothesis Test	
	Reject H_0	Retain H_0
H_0 is true	Type I error $P(\text{Reject } H_0 \mid H_0 = \text{true}) = \alpha$	Correct Decision $P(\text{Retain } H_0 \mid H_0 = \text{true}) = (1 - \alpha)$
H_0 is false	Correct Decision (Power of test) $P(\text{Reject } H_0 \mid H_0 = \text{false}) = 1 - \beta$	Type II Error $P(\text{Retain } H_0 \mid H_0 = \text{false}) = \beta$

know that the sampling distribution of mean from an independent and identically distributed population for large sample follows a normal distribution with mean μ and standard deviation $\sigma / \sqrt{n}$. The standardized value $\dfrac{\bar{X} - \mu}{\sigma / \sqrt{n}}$ follows a standard normal distribution. Z-test uses CLT to conduct a hypothesis test for population mean when the population variance is known; the test statistics for Z-test is given by

$$Z\text{-statistic} = \frac{\bar{X} - \mu}{\sigma / \sqrt{n}} \tag{6.6}$$

The critical value in this case will depend on the significance value α and whether it is a one-tailed or two-tailed test. The critical value for different values of α is shown in Table 6.4.

In Excel, the function NORMSINV(α) [and NORM.S.INV(α)] can be used for finding critical Z-value for left-tailed test. NORMSINV(1 − α) [and NORM.S.INV(1 − α)] will give the critical Z-value for the right-tailed test. NORMSINV(α/2) and NORM.S.INV(1 − α/2) will give critical Z-values for two-tailed test. The decision criteria for rejection or retention of the null hypothesis is described in Table 6.5.

IMPORTANT

One-sample Z-test is used when

1. *Testing the value of population mean when population standard deviation is known.*
2. *The population is a normal distribution and the population variance is known.*
3. *The sample size is large and the population variance is known. That is, the assumption of normal distribution can be relaxed for large samples ($n > 30$).*

TABLE 6.4 Critical value for different values of α

	Approximate Critical Values		
α	Left-Tailed Test	Right-Tailed Test	Two-Tailed Test
0.1	−1.28	1.28	−1.64 and 1.64
0.05	−1.64	1.64	−1.96 and 1.96
0.01	−2.33	2.33	−2.58 and 2.58

TABLE 6.5 Condition for rejection of null hypothesis H_0

Type of Test	Condition	Decision
Left-tailed test	Z-statistic < Critical value	Reject H_0
	Z-statistic ≥ Critical value	Retain H_0
Right-tailed test	Z-statistic > Critical value	Reject H_0
	Z-statistic ≤ Critical value	Retain H_0
Two-tailed test	$\|Z\text{-statistic}\| > \|\text{Critical Value}\|$	Reject H_0
	$\|Z\text{-statistic}\| \leq \|\text{Critical Value}\|$	Retain H_0

EXAMPLE 6.1

An agency based out of Bangalore claimed that the average monthly disposable income of families living in Bangalore is greater than INR 4200 with a standard deviation of INR 3200. From a random sample of 40,000 families, the average disposable income was estimated as INR 4250. Assume that the population standard deviation is INR 3200. Conduct an appropriate hypothesis test at 95% confidence level ($\alpha = 0.05$) to check the validity of the claim by the agency.

Solution:

IMPORTANT

In contexts such as this, we set alternative hypothesis as the statement that we would like to prove.

Claim: Average disposable income is more than INR 4200

Let μ and σ denote the mean and standard deviation in the population. The corresponding null and alternative hypotheses are

$$H_0: \mu \leq 4200$$
$$H_A: \mu > 4200$$

Since we know the population standard deviation, we can use the Z-test. The corresponding Z-statistic is given by

$$Z = \frac{\bar{X} - \mu}{\sigma / \sqrt{n}} = \frac{4250 - 4200}{3200 / \sqrt{40000}} = 3.125$$

This is a right-tailed test. The corresponding Z-critical value at $\alpha = 0.05$ for right-tailed test is approximately 1.64 [in Excel NOMRSINV(1 − α) that is NORMSINV(0.95) gives the critical value for the right-tailed test]. Since the calculated Z-statistic value is greater than the Z-critical value, we reject the null hypothesis. The corresponding

p-value = 0.00088 [p-value in Excel is given by 1 – NORMSDIST(Z-statistic value), that is 1 – NORMSDIST(3.125) in this case]. The critical value, Z-statistic value, and the corresponding p-value are shown in Figure 6.5.

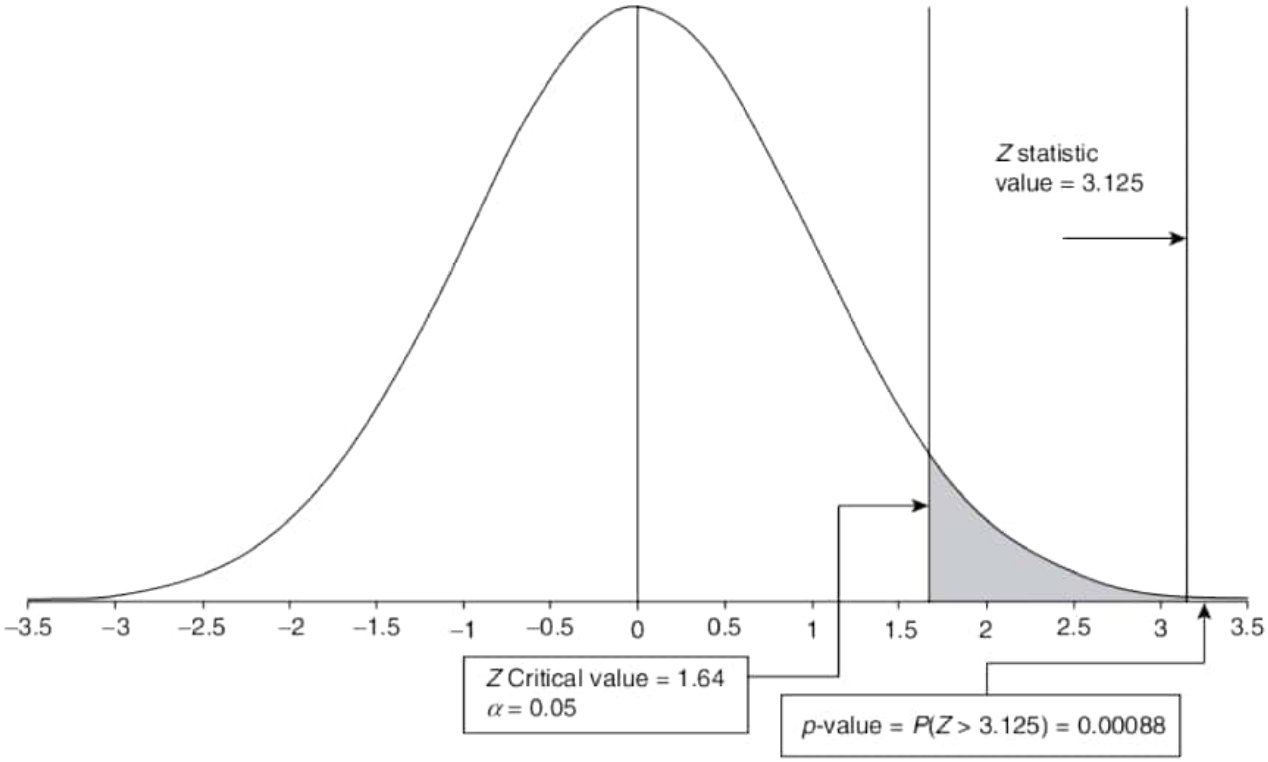


FIGURE 6.5 Critical value, Z-statistic value, and corresponding p-value.

IMPORTANT

Z-statistic measures the standardized difference between estimated value of mean and the hypothesis value of mean. $Z = 3.125$ implies that the sample mean is at 3.125 standard deviations away from the hypothesized population mean given that the null hypothesis is true

EXAMPLE 6.2

A passport office claims that the passport applications are processed within 30 days of submitting the application form and all necessary documents. Table 6.6 shows processing time of 40 passport applicants. The population standard deviation of the processing time is 12.5 days. Conduct a hypothesis test at significance level $\alpha = 0.05$ to verify the claim made by the passport office.

TABLE 6.6 Passport processing time

16	16	30	37	25	22	19	35	27	32
34	28	24	35	24	21	32	29	24	35
28	29	18	31	28	33	32	24	25	22
21	27	41	23	23	16	24	38	26	28

Solution:

Null and alternative hypotheses in this case are given by

$$H_0: \mu \geq 30$$
$$H_A: \mu < 30$$

From the data in Table 6.6, the estimated sample mean is 27.05 days.

The standard deviation of the sampling distribution $\sigma/\sqrt{n} = 12.5/\sqrt{40} = 1.9764$.

The value of Z-statistic is given by

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} = \frac{27.05 - 30}{12.5/\sqrt{40}} = -1.4926$$

The critical value of left-tailed test for $\alpha = 0.05$ is -1.644. Since the critical value is less than the Z-statistic value, we fail to reject the null hypothesis. The p-value for $Z = -1.4926$ is 0.06777 which is greater than the value of α. That is, there is no strong evidence against null hypothesis so we retain the null hypothesis, which is $\mu \geq 30$. Figure 6.6 shows the calculated Z-statistic value and the rejection region.

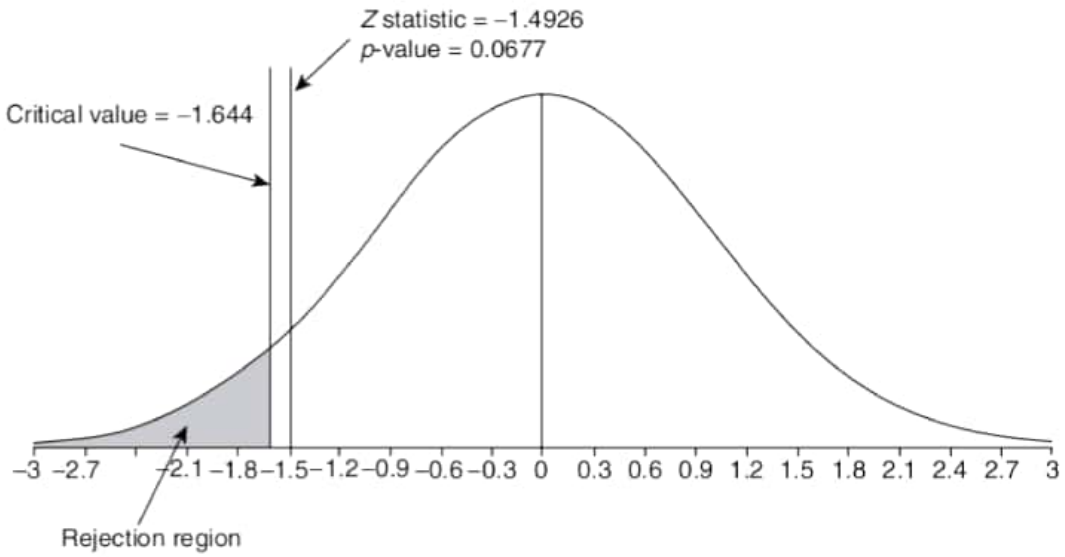


FIGURE 6.6 Left-tailed test for Example 6.2.

EXAMPLE 6.3

According to the company IQ Research, the average Intelligence Quotient (IQ) of Indians is 82 derived based on a research carried out by Professor Richard Lynn, a British Professor of Psychology, using data collected from 2002 to 2006 (Source: IQ Research[1]). The population standard deviation of IQ is estimated as 11.03. Based on a sample of 100 people from India, the sample IQ was estimated as 84.

(a) Conduct an appropriate hypothesis test at $\alpha = 0.05$ to validate the claim of IQ Research (that average IQ of Indians is 82).
(b) Ministry of education believes that the IQ is more than 82. If the actual IQ (population mean) of Indians is 86, calculate the Type II error and the power of hypothesis test.

Solution:

(a) Hypothesis test: It is given that $\mu = 82$, $\sigma = 11.03$, $n = 100$, and $\bar{X} = 84$.

The null and alternative hypotheses in this case are:

$$H_0: \mu = 82$$
$$H_A: \mu \neq 82$$

Since the direction of alternative hypothesis is both ways, we have a two-tailed *t*-test. The test statistics is given by

$$Z = \frac{\bar{X} - \mu}{\sigma / \sqrt{n}} = \frac{84 - 82}{11.03 / \sqrt{100}} = 1.8132$$

For a two-tailed test, the critical values at $\alpha/2 = 0.025$ are −1.96 and 1.96 [in Excel NORMSINV(0.025) = −1.96 and NORMSINV(1 − 0.025) = 1.96]. Since the calculated *Z*-statistic value is within the critical values, we fail to reject the null hypothesis (retain the null hypothesis). Figure 6.7 shows the rejection regions and the *Z*-statistic value in this case. Since the *Z*-statistic value is 1.8132 and falls on the right tail, we first calculate normal distribution beyond 1.8132 which is equal to 0.0348. Since this is a two-tailed test, the *p*-value is twice the area to the right side of the *Z*-statistic value, which is = 0.0698, that is the *p*-value in this case is 0.0698.

[1] Source: https://iq-research.info/en/page/average-iq-by-country

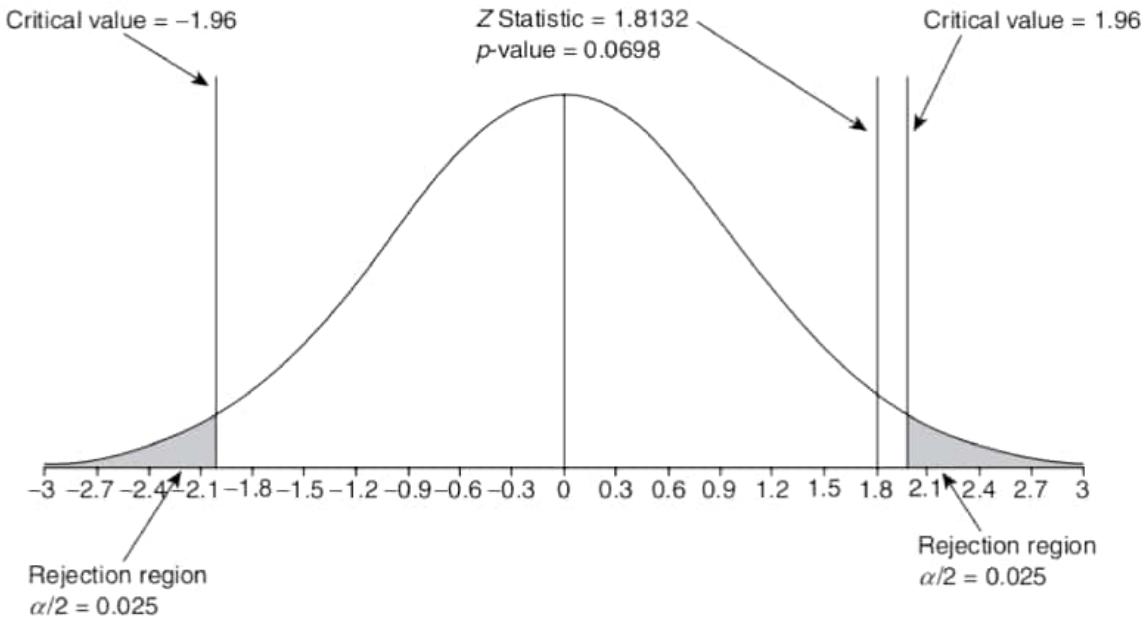


FIGURE 6.7 *Z*-statistic, critical values, and the rejection region for Example 6.3.

IMPORTANT

In a two-tailed test, the p-value is two times the tail area.

(a) **Calculating Type II Error and Power of Test:** In this case, the null and alternative hypotheses are

$$H_0: \mu \leq 82$$
$$H_A: \mu > 82$$

Note that ministry of education believes that the average IQ is 86 (thus we have to carry out a right-tailed test). Type II error is the conditional probability of retaining a null hypothesis when it is false, that is $P(\text{retaining } H_0 \mid H_0 \text{ is false})$.

The mean and standard deviation of Z-statistic in null hypothesis are 82 and 1.103, respectively. For the standard normal distribution the critical value for a right tailed test when $\alpha = 0.05$ is 1.644. The corresponding critical value for the normal distribution $N(82, 1.103)$ is

$$X_{\text{critical}} = \mu + Z_\alpha \times \sigma / \sqrt{n} = 82 + 1.644 \times 1.103 = 83.8133$$

That is, under normal distribution $N(82, 1.103)$, the region beyond 83.8133 is the rejection region (rejection of null hypothesis).

Now consider the normal distribution $N(86, 1.103)$. Area under this normal distribution may take values below 83.8133 which is region of retaining the null hypothesis, although the actual mean in this case is 86. Thus, we will be retaining the null hypothesis when it is incorrect resulting in Type II error, β (Figure 6.8).

For the normal distribution $N(86, 1.103)$, the probability of the variable taking value less than 83.8133 (the critical value) is given by

$$P(X \leq 83.8133) = P\left(Z \leq \frac{83.8133 - 86}{1.103}\right) = 0.0237$$

That is, the Type II error $\beta = 0.0237$

The power of test, $1 - \beta = 1 - 0.0237 = 0.9763$

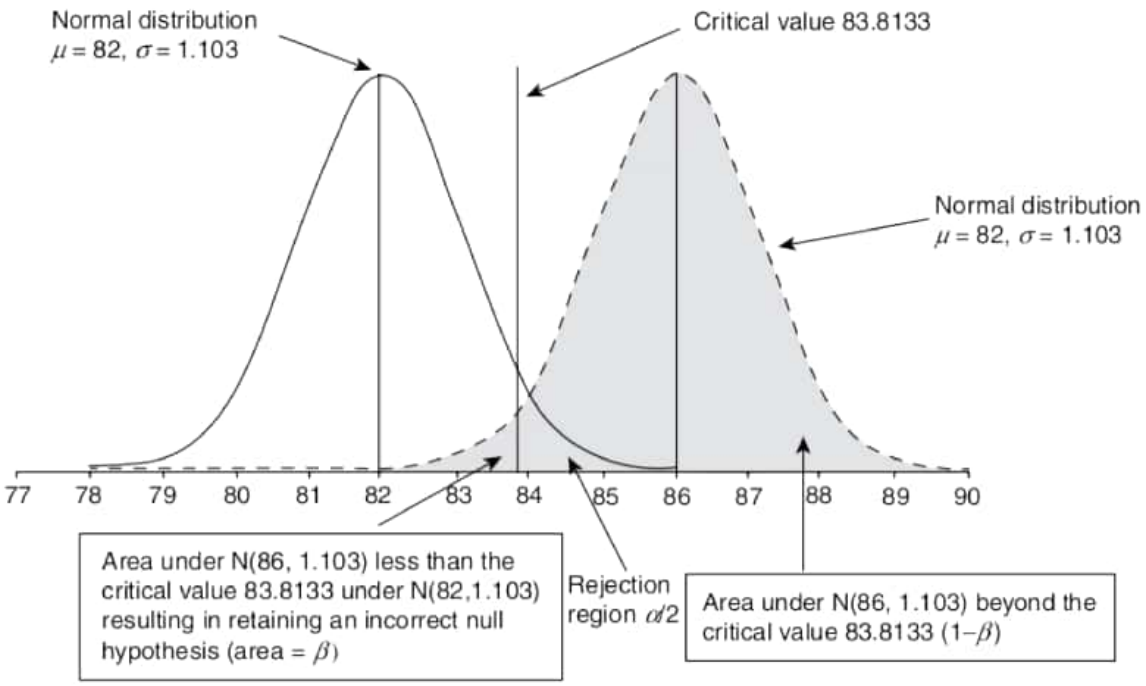


FIGURE 6.8 Type II error and power of hypothesis test.

6.5.1 | Power of Test and the Power Function

The power of the test $1 - \beta$ is the conditional probability of rejecting the null hypothesis when the alternative hypothesis is true. For different values of the actual value of population mean, we can calculate the power $(1 - \beta)$. The plot between different mean values and $(1 - \beta)$ is called the **power function** and is shown in Figure 6.9.

Figure 6.9 shows the change in power of test as the actual value of mean changes.

6.7 | HYPOTHESIS TEST FOR POPULATION MEAN UNDER UNKNOWN POPULATION VARIANCE: *t*-TEST

We use the fact that a sampling distribution of a sample from a population that follows normal distribution with unknown variance follows a *t*-distribution with $(n-1)$ degrees of freedom. In many cases the population variance (and thus the standard deviation) will not be known. In such cases we will have to estimate the variance using the sample itself. Let S be the standard deviation estimated from the sample of size n. Then the statistic $\frac{\bar{X}-\mu}{S/\sqrt{n}}$ will follow a *t*-distribution with $(n-1)$ degrees of freedom if the sample is drawn from a population that follows a normal distribution. Here 1 degree of freedom is lost since the standard deviation is estimated from the sample. Thus, we use the *t*-statistic (hence the test is called *t*-test) to test the hypothesis when the population standard deviation is unknown.

$$t\text{-statistic} = \frac{\bar{X}-\mu}{S/\sqrt{n}} \tag{6.8}$$

The t-test is used when the population follows a normal distribution and the population standard deviation σ is unknown and is estimated from the sample. t-test is a robust test for violation of normality of the data as long as the data is close to symmetry and there are no outliers.

EXAMPLE 6.5

Aravind Productions (AP) is a newly formed movie production house based out of Mumbai, India. AP was interested in understanding the production cost required for producing a Bollywood movie. The industry believes that the production house will require at least INR 500 million (50 crore) on average. It is assumed that the Bollywood movie production cost follows a normal distribution. Production cost of 40 Bollywood movies in millions of rupees are shown in Table 6.7. Conduct an appropriate hypothesis test at $\alpha = 0.05$ to check whether the belief about average production cost is correct.

TABLE 6.7 Production cost of Bollywood movies

601	627	330	364	562	353	583	254	528	470
125	60	101	110	60	252	281	227	484	402
408	601	593	729	402	530	708	599	439	762
292	636	444	286	636	667	252	335	457	632

Solution:

It is given that the production cost of Bollywood movies follows a normal distribution; however, the standard deviation of the population is not known and we need

to estimate the standard deviation value from the sample. Thus, we have to use the t-test for testing the hypothesis. From the sample data in Table 6.7 we get the following values:

$$n = 40,\ \bar{X} = 429.55, \text{ and } S = 195.0337$$

The null and alternative hypotheses are

$$H_0: \mu \leq 500$$

$$H_A: \mu > 500$$

The corresponding test statistic is

$$t\text{-statistic} = \frac{\bar{X} - \mu}{S/\sqrt{n}} = \frac{429.55 - 500}{195.0337/\sqrt{40}} = -2.2845$$

Note that this is a one-tailed test (right-tailed) and the critical t-value at $\alpha = 0.05$ under right-tailed test, $t_{critical} = 1.6848$ [in Excel TINV(2α, df) will return right-tailed critical value at significance of α, in this example $\alpha = 0.05$, the corresponding critical t-value using Excel function is TINV(0.1, 39) = 1.6848, that is the critical value is 1.6848]. Since t-statistic value is less than the critical t-value, we retain the null hypothesis. The t-statistic value and critical value for the t-test are shown in Figure 6.10.

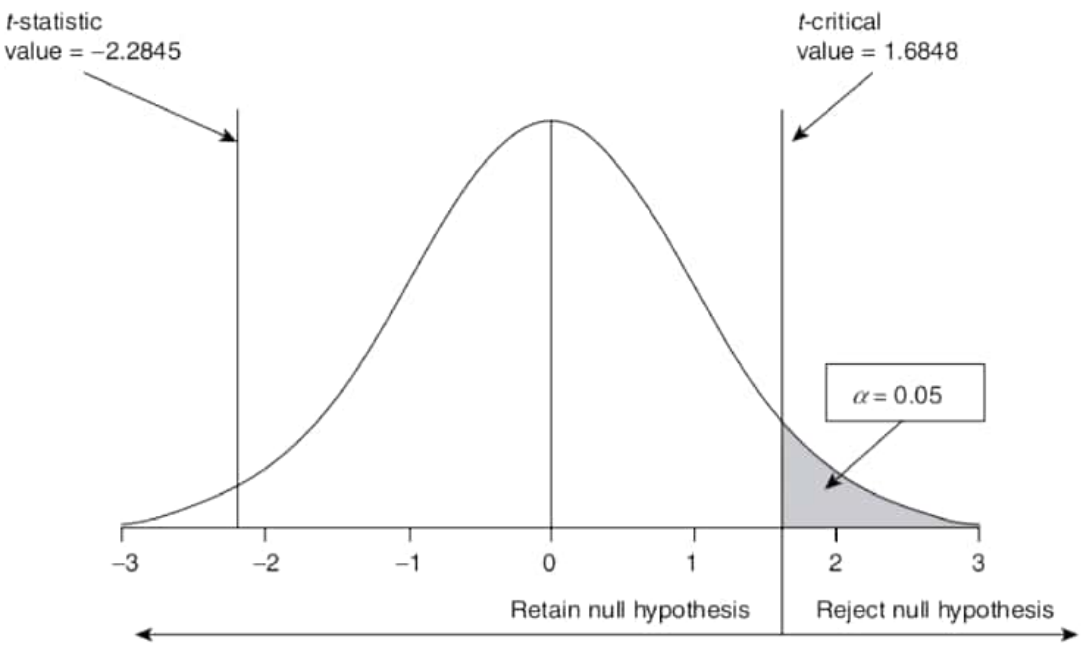


FIGURE 6.10 Critical value, t-statistic value for t-test in Example 6.5.

EXAMPLE 6.6

According to statistics released by the Department of Civil Aviation, the average delay of flights is equal to 16.8 minutes, flight delays are assumed to follow a normal distribution. However, from a sample of 50 flights, the average delay was estimated to be 19.5 minutes and the sample standard deviation was 6.6 minutes. Conduct a hypothesis test to disprove the claim that the average delay is equal to 16.8 minutes at $\alpha = 0.01$.

Solution:

Given $n = 50$, $\bar{X} = 19.5$, $S = 6.6$

Null and alternative hypotheses are

$$H_0: \mu = 16.8$$
$$H_A: \mu \neq 16.8$$

The corresponding t-statistic value is

$$t = \frac{\bar{X} - \mu}{S/\sqrt{n}} = \frac{19.5 - 16.8}{6.6/\sqrt{50}} = 2.8927$$

The critical t-value for two-tailed t-test when $\alpha = 0.01$ and degrees of freedom = 49 is 2.67 [in Excel, TINV(0.01, 49) = 2.68 or T.INV.2T(0.01, 49) = 2.68]. Since the calculated t-statistic value is greater than the t-critical value, we reject the null hypothesis. The corresponding p-value is 0.0057 [in excel T.DIST(t-statistic value, degrees of freedom, tails) returns the p-value, T.DIST.2T(2.8927, 49) = 0.0057]. The values of t-statistic, t-critical value, rejection and retention regions are shown in Figure 6.11.

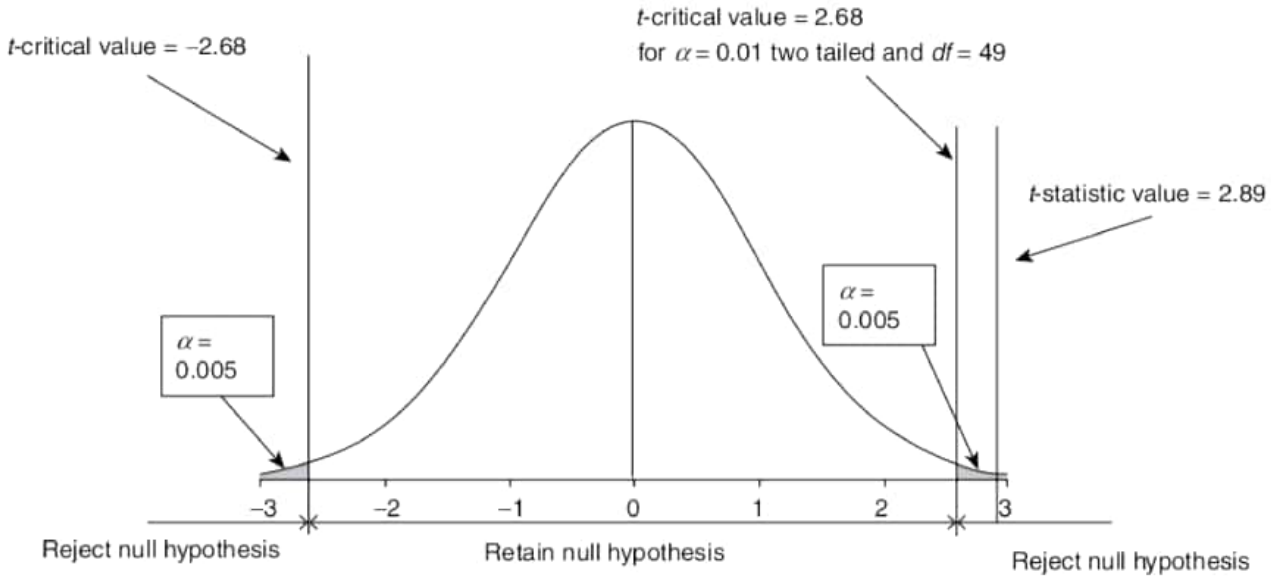


FIGURE 6.11 t-statistic, t-critical, rejection and acceptance regions for Example 6.6.