2.1 | Introduction to Descriptive Analytics

Descriptive analytics is the starting point of analytics-based solutioning. It helps understand data using simple descriptive statistics and visualisation, and provides directions for predictive and prescriptive analytics. Business Intelligence (BI), which generally involves creating reports and business dashboards that lead to actionable insights, is essentially a descriptive analytics exercise. Feature engineering is another important use of descriptive analytics.

Descriptive analytics is the science of describing past data and thus capturing 'what happened' in a business and social context. The primary objective of descriptive analytics is simple comprehension of data using summarization techniques such as pivot tables, basic statistical measures and visualization, and providing ideas for feature engineering and predictive analytics. Descriptive statistics such as measures of central tendency, measures of variation and measures of shape can provide useful insights. Visualization of data using plots such as histogram, bar chart, pie chart, box plot, scatter plot and tree diagram can provide insights into past data. Descriptive statistics and visualization can assist data scientists in generating new hypotheses and additional features as part of feature engineering.

Descriptive analytics is an important component of reporting across several industries, as it enables top management to take decisions after monitoring Key Performance Indicators (KPIs). Many organizations generate reports at regular intervals and design dashboards as part of Business Intelligence (BI) to monitor and communicate various aspects of the business to top management, stakeholders and the external world. Business reports include descriptive analytics in the form of tables, charts and innovative visuals such as bubble chart, heatmap, sunburst chart and tree map. With the advent of mobile technology, many real-time reports are generated which can be accessed by the top management on their mobile handsets, enabling them to take quick action, if necessary. For example, a retailer such as Big Bazaar or Reliance Retail in India might want to know the five top-selling brands in their stores by region, by city, by store, etc. Such information will help the management plan inventory, shelf space, pricing, etc. Similarly, the retailer may markdown the price of an item that is slow-moving. They can also monitor trends in revenue generated at regional, city and store levels over the past several periods. One of the primary applications of descriptive analytics is designing effective dashboards and scorecards to communicate the performance of the organization to various stakeholders.

Any analytics project should ideally start with descriptive analytics to gain insights before venturing into predictive analytics. Descriptive analytics can also assist in feature engineering, which is likely to improve the performance of predictive analytics algorithms.

2.2 | Data Types and Scales of Variable Measurement

Data and variables are classified into different categories based on data structure and the 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. In many cases, the dataset may include both structured and unstructured data. Structured data refers to data reported in a matrix form with labelled rows and columns. Rows represent different records in the data, whereas columns capture various independent (features) and dependent (outcome) variables. Any data not originally in matrix form with identifiable rows and columns is unstructured data. For example, e-mails, click streams, log files, textual data, images (photos and images generated by medical devices, satellite images) and videos fall under unstructured data. Machine-generated data such as images generated by satellite, magnetic resonance imaging (MRI), electrocardiogram (ECG), and thermography are a few examples of unstructured data. The generation of unstructured data has witnessed an upward trend due to social media platforms such as Facebook, Twitter, Instagram and YouTube, and the analysis of unstructured data is important for effective management. Internet of Things (IoT) is another source of unstructured data. IoT devices are used across such industries as aerospace, automobile and machine tools to capture data in real time.

The importance of unstructured data in decision-making has increased extensively in the recent past due to its applications across industries. For example, analysing social media data is important for companies to understand the sentiments expressed by customers about their products/services and take necessary remedial measures. A significant

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	2,70,000
2	M	21	76.33	ICSE	75.33	75.48	2,20,000
3	M	22	72	Others	78	66.63	2,40,000
4	M	22	60	CBSE	63	58	2,50,000
5	M	22	61	CBSE	55	54	1,80,000
6	M	23	55	ICSE	64	50	3,00,000
7	F	24	70	Others	54	65	2,40,000
8	M	22	68	ICSE	77	72.5	2,35,000
9	M	24	82.8	CBSE	70.6	69.3	4,25,000
10	F	23	59	CBSE	74	59	2,40,000

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/

proportion of social media data is natural language (text), along with images and videos. Emerging technologies such as driverless cars will involve analysing images in real time. Apart from social media, machine-generated data is usually unstructured (data generated from medical devices such as ECG, MRI, etc.). Voice-enabled systems such as Alexa and Siri require processing natural languages, and have become a primary area of research in analytics. A high percentage of Big Data problems involve the use of unstructured data. One of the main challenges in analysing unstructured data is the conversion of unstructured data to structured data, to enable analytics model development. Most analytics model-building involves matrix operations, and thus converting unstructured data to structured data innovatively is important. Examples of structured and unstructured data are shown in Tables 2.1 and 2.2.

The data in Table 2.2 is clickstream data (search behaviour of an internet user that captures the websites visited by the user). Clickstream data is useful to understand the behaviour of internet users. A few recommender systems use clickstream data to predict customer preferences. Based on surfing behaviour, individuals are targeted with advertisements for products and services. As we can see, unstructured data shown in Table 2.2 does not have a matrix structure such as in the case of structured data shown in Table 2.1. Before any analytics model can be built, unstructured data must be converted into structured data.

2.2.2 Cross-sectional, Time Series and Panel Data

Classification of data into the following three categories depends on the types of data collected.

1. **Cross-sectional data:** Data collected on many variables of interest at the same instance or time is called cross-sectional data. An example of cross-sectional data is data on movies released in 2019 pertaining to budget, box office collection, actors, directors, Facebook likes and genre. In this case, the time period (year 2019) is the same for all the records.
2. **Time-series data:** Data collected on a single variable such as demand for smartphones over several time intervals (weekly, monthly, etc.) is called 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). An example of panel data is the data collected on variables such as Gross Domestic Product (GDP), Gini index and unemployment rate for several countries over several years.

An appropriate analytical model and the diagnostics will depend on the type of data used for a study. For example, we have to check for autocorrelation of errors in the case of time-series data. In the case of panel data, the data scientist must decide whether to use Fixed Effect Model (FEM) or Random Effect Model (REM).

2.3 | Types of Variable Measurement Scales

Structured data may contain variables that are either numeric or alphanumeric, and may follow different scales of measurement (or measurement scale). It is important to understand the type of variable within the data with respect to the measurement scale since the model specification while building analytics models such as linear regression and logistic 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 variable), and are also known as categorical variables. For example, variables such as marital status (single, married, divorced) and industry type (manufacturing, healthcare, banking, insurance and finance) fall under the nominal scale. During data collection, the usual practice is to assign a numerical code to represent a nominal variable. For example, in the case of the categorical variable marital status, the data collector may have used the number 1 to represent single, 2 for married and 3 for divorced. The numbers 1, 2, and 3 are just codes and do not have any value attached to them. That is, basic mathematical operations are meaningless for a nominal scale (e.g., subtraction: married-unmarried or ratio: married/unmarried is meaningless). While developing statistical models, nominal scale variables are usually transformed as part of feature engineering before building the model. For example, when developing a regression model, categorical variables are converted using dummy variables before building the regression model (discussed in Chapter 10).

2.3.2 Ordinal Scale

Ordinal scale is a variable type in which the value of the variable is captured from a rank-ordered set. For example, in many survey data, the Likert scale is used. The Likert scale is finite (usually a five-point scale), and the data collector would have defined the order of preference (rank). For example, assume that feedback is collected on a training program using the five-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) or ratio 4/5 (Very Good/Excellent) are meaningless. Also, we cannot assume that a feedback score of 4 is twice as good as a feedback score of 2.

2.3.3 Interval Scale

An interval scale corresponds to a variable type in which the value of the variable is chosen from an interval set. Variables such as temperature measured in Centigrade (°C) and Intelligence Quotient (IQ) scores are examples of interval scales. In an 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 as smart as a person with an IQ score of 80. However, 40°C is 20°C more than 20°C, and an IQ score of 160 is 80 more than an IQ score of 80. Surprisingly, no human is assumed to have an IQ score of zero 😊. Grade Point Average (GPA) used by academic institutes is another example of interval scale. That is, in an interval scale, the difference between two values is meaningful, but not the ratio. In an interval scale, the reference point is fixed arbitrarily. For example, 0°C is fixed based on the freezing point of water. Variables such as dates (no zero), location in cartesian coordinate and time in 24-hour clock are other examples of interval scale.

2.3.4 Ratio Scale

Any variable for which ratios can be computed and are meaningful is called a ratio scale variable. Many continuous variables come under ratio scale – such as 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, 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 or records (often called cases, 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 2019, about 900 million people were eligible to vote in the Indian general elections (Source: Election Commission of India). Thus, the population size of the eligible voters in the 2019 Indian parliamentary elections was 900 million. During every election, media and other organisations collect data to predict the likely winner of the election through opinion polls (they rarely get it right due to complexities associated with collecting data). Records within a population in many cases may be unknown and dynamic (new records are added continuously, and old records are deleted). For example, if we define the population as diabetic patients in India, there is no existing database that has information related to all diabetic patients in India. It is very difficult (practically impossible) to collect data from all 900 million eligible voters about their choice of candidate, so opinion polls are based on opinions expressed by a subset of voters called **sample**.

A **sample** is a subset taken from a population. In many real-life problems, we make inferences about the population based on sample data or based on multiple samples for cross-validation. There are many challenges in sampling (process of selecting a record 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 those used to describe the data using a single value. **Mean, median** and **mode** are the three measures of central tendency frequently used to describe data and make comparisons between different datasets. Measures of central tendency help users summarize and comprehend the data.

2.5.1 Mean (or Average) Value

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

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

The 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 records from the entire population, then we have the population mean which is usually denoted by μ. Among all the measures of central tendency, mean is the most frequently used measure, since it uses values of all records (all X_i values) in the dataset (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 (or mean salary) is given by

$$\bar{X} = \frac{(270 + 220 + 240 + 250 + 180 + 300 + 240 + 235 + 425 + 240) \times 1000}{10} = 2{,}60{,}000$$

The average (or mean) salary is ₹260,000. Note that the average value need not be a part of the dataset, that is, none of the graduating students' salary is ₹260,000. In Microsoft Excel, function '*Average (array)*' can be used to calculate 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 all records from the mean is zero, that is

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

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 (an English statistician who developed concepts such as correlation and regression) attended a contest in Plymouth, UK, in which villagers were asked to guess the weight of an ox. The one who guessed the closest to the actual weight of the ox was given the ox as the prize for winning the contest. An estimated 800 villagers participated in the contest. Galton found that the average of all the weights entered by the participants in the contest was very close to the actual weight. In fact, the difference between the average and the actual weight was less than a pound. Also, the average turned to be better than the guess by the winner of the contest (Surowiecki, 2004).

Although mean is one of the most frequently used measures of central tendency, one should be careful about taking decisions based on the mean value of the data. There is a famous joke in statistics which goes, '*If someone's head is in a freezer and leg in an oven, the average body temperature would be fine, but the person may not be alive*'. Making decisions based solely on mean value is not advisable since variability in the data can be more critical than the mean. In capital asset procurement such as the 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) on its own is not a reliable measure to assess the reliability of the asset and therefore is not very useful while taking operational decisions. It must be used along with other measures such as standard deviation for better understanding of the data. Another issue with mean is that it is affected significantly by outliers. That is, the presence of an outlier can change the mean value significantly. If the data is captured in frequencies, then Eq. (2.2) can be used to calculate the average:

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

In Eq. (2.2) f_i is the frequency of the record X_i, that is, the total number of times the record X_i appears in the data. The frequency of the age of students in Table 2.1 is given as

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 and above the median will both be 50%. When a dataset containing n records is arranged in ascending order, the median is given by the value at position $(n+1)/2$ given n is odd. When n is even, the median is the average value of the $(n/2)$th and $\{(n/2)+1\}$th records.

Table 2.3 | Number of deposits in a bank

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

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.

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 observations below and above 260. In Microsoft Excel, the function '*Median (array)*' can be used to calculate the median of a dataset.

Consider the following data arranged in the increasing order of magnitude:

180, 220, 235, 240, 250, 260, 270, 300, 425, 500

There are 10 records in the data. The values of the fifth and sixth records are 250 and 260 respectively, and the average is 255. Thus, the median of the data is 255. Median is more stable than the mean value, as adding a new observation may not change the median significantly. However, the drawback of median is that it is not calculated using the entire dataset like in the case of mean. We are simply looking for the midpoint instead of using the actual values of the entire data.

2.5.3 Mode

Mode is the most frequently occurring value in the dataset. For example, in the data 'salary' in Table 2.1, the value 240,000 appears three times and is the mode since all other values are observed only once. In Microsoft Excel function '*Mode (array)*' can be used to calculate mode. It is the only measure of central tendency valid for qualitative (nominal) variables, since the measures mean and median for nominal scale variables are meaningless. For example, assume that customer data of a retailer has the marital status of customers under four categories, 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 a bar chart or histogram, mode is represented by the tallest column. It is possible that a dataset may not have any mode at all. For example, if each value in the dataset appears only once (or, an equal number of times), there is no mode in the dataset.

2.6 | Percentile, Decile and Quartile

Self study

Percentile, decile and quartile are frequently used to identify the position of an observation or record in the dataset. Percentile score is often used in the educational sector to identify the position of a potential student in the group. Another frequent application of percentile is percentile life used in asset management. Percentile, denoted as P_x, is the value of the data at which x percentage of the data lies below that value. For example, P_{10} denotes the value below which 10 percentage of the data lies. In the context of asset management and in the field of reliability, P_{10} life implies the time by which 10% of the products will fail. To find P_x, the data must be arranged in ascending order. The value of P_x will be given by the position in the data calculated using the approximate 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 a non-integer, in which case we can either round it off to the nearest integer or use an approximation technique as explained in Example 2.1. **Deciles** correspond to special values of percentile that divide the data into 10 equal parts. The first decile (P_{10}) contains the first 10% of the data (that is, 10% of the records will have a value less than or equal to P_{10} value), the second decile contains the first 20% of the data, and so on. Similarly, **Quartiles** divide the data into four equal parts. The first quartile (Q_1) contains the first 25% of the data; that is, 25% of the records in the data will have values less than the Q_1 value. Quartile 2 (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 the value of the 10th percentile.

Example 2.1

The time between failures (in hours) of a wire cut (wire used to cut the dough to different shapes of cookies) used in a cookie manufacturing oven is given in Table 2.4.

(a) Calculate the mean, median and mode of the time between failures of the 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} (or Q_1) and P_{75} (or Q_3).

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

Solution

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

(b) Note that the data in Table 2.4 is arranged in increasing order of 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 2.4 is 21 (10 percentage of observations in Table 2.4 have a value 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 off the value obtained from Eq. (2.3), we can use the following approximation:

$$P_{10} = 10 \times (51)/100 = 5.1$$

Value at 5th position is 21. Value at position 5.1 is approximated as

$$21 + 0.1 \times (\text{value at } 6^{\text{th}} \text{ position} - \text{value at } 5^{\text{th}} \text{ position}) = 21 + 0.1(1) = 21.1$$

The position of P_{90} is given by

$$P_{90} = 90 \times 51/100 = 45.9$$

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

$$90 + 0.9 \times (3) = 92.7$$

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

Note that the approximation suggested in this example is one of many approximations available in the literature. Tools such as Microsoft Excel may give slightly different values. However, the meaning will remain the same. In this example, any value between 21 and 22 is a P_{10} value.

(c) P_{25} (1st Quartile or Q_1) = 25 × 51/100 = 12.75
The value at the 12th position is 33, so
P_{25} = 33 + 0.75 (value at 13th position − value at 12th position) = 33 + 0.75 (1) = 33.75
P_{75} (3rd Quartile or Q_3) = 75 × 51/100 = 38.25
The value at the 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

self study

One of the primary objectives of analytics is to understand the variability in the data and, if possible, what causes such variability. Predictive analytics techniques such as regression attempt to explain the variation in the outcome variable (Y) using values of predictor variables or features (X). Measures of variability are useful in identifying how close the records are to the mean value and outliers in the data. Another important application of variability is in variable selection in analytics model-building. If a variable or feature has very low variability, it is unlikely to have a statistically significant relationship with an outcome variable. A trivial example is when a predictor variable (or feature) has the same value for all outcome variable values in the data. In such cases, the feature will not have any statistically significant association relationship with the outcome variable, and thus including the feature in model-building will be unnecessary. Such insights are very useful during feature extraction and feature selection stages of analytics project life-cycle. Variability in the data is measured using the following measures:

1. Range
2. Inter-Quartile Distance (IQD)
3. Mean Absolute Deviation (MAD)
4. Variance
5. Standard Deviation
6. Coefficient of Variation

We will discuss each of these measures in detail in the following sections.

2.7.1 Range

Range is the difference between the maximum and minimum values of the data in the sample. It captures the data spread. For the data provided in Table 2.4, the range = 102 − 2 = 100. If the range is large, then a data scientist may like to create buckets and use them in model-building. For example, the income of a loan applicant at a bank may range from a few hundred rupees to several thousand rupees per month. Instead of using the income as it is, the data scientist may use buckets such as:

Bucket 1: Less than 10,000
Bucket 2: Between 10,000 and 20,000
Bucket 3: Between 20,000 and 50,000
Bucket 4: Over 50,000

The conversion of a continuous number into buckets is a part of feature engineering and may improve the model.

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) in a dataset. 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. An 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 potential 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 many approaches used to identify outliers. Its use 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 and Cook's distance to identify outliers (discussed in Chapters 9 and 10).

2.7.3 Mean Absolute Deviation (MAD)

Mean Absolute Deviation (MAD) is the average value of absolute deviation of records from its mean in the data. MAD is mathematically given by

$$\text{MAD} = \sum_{i=1}^{n} \frac{|X_i - \mu|}{n} \tag{2.4}$$

MAD is always finite and defined, where measures such as variance can be infinite or undefined (for example, Cauchy distribution has no defined variance). MAD value for the data in Table 2.4 is 25.26.

2.7.4 Variance and Standard Deviation

Variance is a measure of variability in the data from the mean value of the data. Variance for population, σ^2, is calculated using the following equation (Eq. 2.5)

$$\text{Variance} = \sigma^2 = \sum_{i=1}^{n} \frac{(X_i - \mu)^2}{n} \tag{2.5}$$

where n is the number of records in the data. In Eq. (2.5), deviation from mean is squared since the sum of deviations from mean will always add up to zero, that is

$$\sum_{i=1}^{n} (X_i - \mu) = 0.$$

In many data analyses, variance is preferred over Mean Absolute Deviation. One drawback of MAD is that since it is an absolute value, it is not easy for mathematical operations such as differentiation. The variance for the data in Table 2.4 is 818.0304 [using Eq. (2.5)]. In case of a sample, the Sample Variance (S^2) is calculated using Eq. (2.6)

$$S^2 = \sum_{i=1}^{n} \frac{(X_i - \bar{X})^2}{n-1} \tag{2.6}$$

While calculating sample variance S^2, the sum of squared deviations

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

is divided by $(n - 1)$; this is known as Bessel's correction. For the data in Table 2.4, the sample variance is 834.7249. Microsoft Excel functions *Var.P(array)* and *Var.S(array)* are used to calculate population variance and sample variance respectively. Population standard deviation (σ) and sample standard deviation (S) are given by

$$\sigma = \sqrt{\sum_{i=1}^{n} \frac{(X_i - \mu)^2}{n}} \tag{2.7}$$

$$S = \sqrt{\sum_{i=1}^{n} \frac{(X_i - \bar{X})^2}{n-1}} \tag{2.8}$$

Table 2.5 | Sample of 10 observations from Table 2.4

2	3	5	9	21	93	99	99	99	102

For the data in Table 2.4, the standard deviation obtained using the Eq. (2.7) is 28.6012. The sample standard deviation S is given in Eq. (2.8).

Note that the standard deviation = $\sqrt{\text{variance}}$. For the data in Table 2.4, the standard deviation obtained using Eq. (2.8) is 28.8916. In Microsoft Excel, the functions *Stdev.P(array)* and *Stdev.S(array)* can be used to calculate 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. (2.6) and (2.8). One argument is that, when we take a sample and estimate the mean from the sample $\bar{X}$, we tend to underestimate the sum of squared deviations from the mean. For example, take a sample consisting of the first five (first column) and last five (last column) observations from Table 2.4. The sample is given in Table 2.5.

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

In Table 2.6, we can see that the numerator in Eq. (2.5) is underestimated (20,713.60) when we use sample average against population average (20,910.74). 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 (known as Bessel's correction).

Another argument of using Eq. (2.6) is through the concept of **degrees of freedom**. In simple terms, degrees of freedom is the number of values in a data that are free to vary. 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 $\bar{X}$ 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 the number of parameters estimated (Walker 1940, Toothaker and Miller, 1996). If there are n observations in the sample and k parameters are estimated

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	2,621.44	3,095.81
3	2,520.04	2,985.53
5	2,323.24	2,770.97
9	1,953.64	2,365.85
21	1,036.84	1,342.49
93	1,584.04	1,250.33
99	2,097.64	1,710.65
99	2,097.64	1,710.65
99	2,097.64	1,710.65
102	2,381.44	1,967.81
Sample Mean = 53.2	$\sum(X_i - \bar{X})^2 = 20{,}713.60$	$\sum(X_i - \mu)^2 = 20{,}910.74$

from the sample, then the number of degrees of freedom is $(n-k)$. While using Eq. (2.6) or Eq. (2.8), the value of $\bar{X}$ is estimated from a sample of size n. Thus, the degrees of freedom is $(n-1)$.

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

2.7.5 Coefficient of Variation (CV)

Coefficient of Variation (CV) is the ratio of standard deviation to the mean. CV is also a measure of inequality (such as inequality of income, treatment time of cancer patients, and so on). Mathematically, CV is given by

$$CV = \frac{\sigma}{\mu} \tag{2.9}$$

CV is variation in relation to the mean value. CV is independent of units of measurement, which makes it useful for comparison of variability of two different variables or features. CV is useful while comparing two populations in which there is a large difference in the absolute values. For example, consider the income in US dollars (USD) of citizens from India, Somalia and Sweden. There will be a huge difference in the absolute values (in USD) of the incomes of the citizens of these three countries. However, CV captures the variability within each population. In the manufacturing context, CV is useful to assess the precision associated with a manufacturing process. For the data in Table 2.4, the coefficient of variation is 0.501.

2.7.6 Chebyshev's Theorem

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

$$P(\mu - k\sigma \le X \le \mu + k\sigma) \ge 1 - \frac{1}{k^2} \tag{2.10}$$

Alternatively, Chebyshev's theorem can be written as

$$P(|X-\mu| \ge k\sigma) \le \frac{1}{k^2} \tag{2.11}$$

Eq. (2.10) is useful when the value of $k > 1$, otherwise it gives a trivial solution. Note that Chebyshev's inequality defined in Eq. (2.10) gives the lower bound of the probability of the random variable taking values between $\mu - k\sigma \le X \le \mu + k\sigma$. Chebyshev's inequality is useful for calculating the probability of observing a value within an interval written as a function of the mean and standard deviation of the data.

Example 2.2

The amount spent per month by a segment of credit card users of a bank has a mean value of 12,000 and a standard deviation of 2,000. Calculate the lower bound on the proportion of customers who spend between 8,000 and 16,000.

Solution

$$P(8{,}000 \le X \le 16000) = P(\mu - 2\sigma \le X \le \mu + 2\sigma) \ge 1 - \frac{1}{2^2} = 0.75$$

That is, the proportion of customers spending between 8,000 and 16,000 is at least 0.75 (or 75%)

2.8 | Measures of Shape – Skewness and Kurtosis

Self Study

Skewness is a measure of symmetry, or lack of it. A dataset is symmetrical when the proportion of data at equal distances (measured in terms of standard deviation) from the mean (or median) on either side is equal. That is, the proportion of data between μ and $\mu - k\sigma$ is the same as that between μ and $\mu + k\sigma$, where k is a positive constant. Measure of skewness can be used to identify whether the distribution is left-skewed (longer tail on the left side of the distribution) or right-skewed (longer tail on the right side of the distribution). Skewness plays an important role since many variables such as income, wealth, age, waiting time for a service, length of stay in a hospital, reliability of systems, return on investment and so on are usually skewed. In the context of finance and investment, Skewness is useful for measuring the extreme values of return on investment.

There are many different approaches to measuring skewness. **Pearson's moment coefficient of skewness** for a dataset with n observations is given by

$$g_1 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^3 / n}{\sigma^3} \tag{2.12}$$

The value of g_1 will be close to 0 when the data is symmetrical. The skewness of a normal distribution is zero. A positive value of g_1 indicates a positive skewness and a negative value indicates negative skewness. The formula in Eq. (2.12) 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):

$$G_1 = \frac{\sqrt{n(n-1)}}{n-2} g_1 \tag{2.13}$$

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 to calculate the value of skewness (G_1) calculated from a sample. In Figure 2.1, the positive-skewed, normal and negative-skewed distributions are shown.

Skewness is used in finance to understand risk and return. For example, negative skewness in data about returns on stocks would imply that the returns could be much lower than the mean, and it could be negative returns or loss. Positive skewed distribution of returns would imply the returns could be much higher than average.

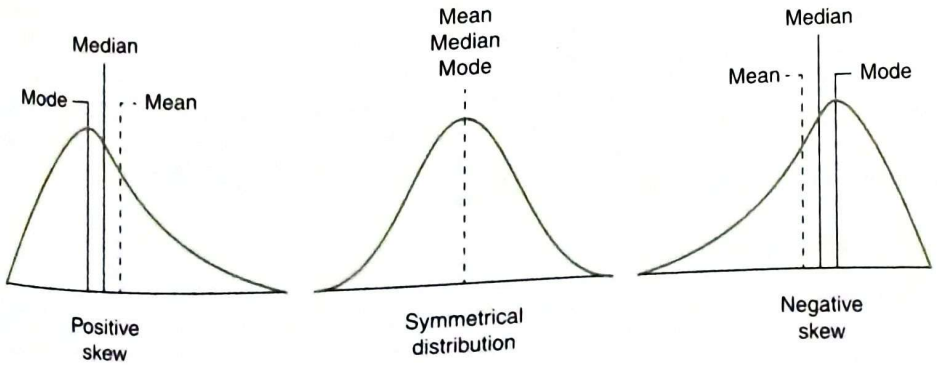

Figure 2.1 | Skewness.

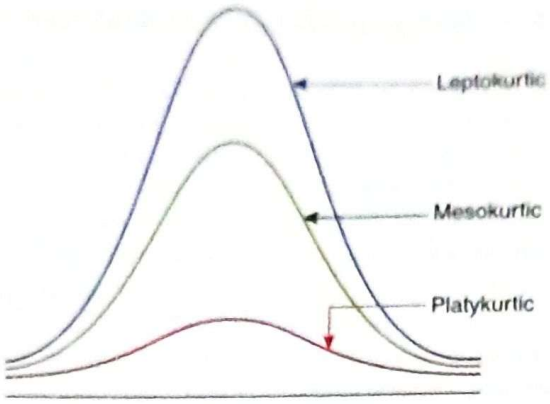

Figure 2.2 | Leptokurtic, mesokurtic, and platykurtic distributions.

Kurtosis is another measure of shape that goes by the shape of the tail – that is, whether the tail of the data distribution is heavy or light. Kurtosis measures the shape of the tails in comparison to the overall shape. Kurtosis is measured using the following equation:

$$\text{Kurtosis} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^4 / n}{\sigma^4} \tag{2.14}$$

A kurtosis value of less than 3 represents a **platykurtic distribution**, while one greater than 3 represents a **leptokurtic distribution**. A kurtosis value of 3 indicates standard normal distribution (also called **mesokurtic**). Excess kurtosis is a measure that captures deviation from kurtosis of a normal distribution and is given by

$$\text{Excess kurtosis} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^4 / n}{\sigma^4} - 3 \tag{2.15}$$

For the data in Table 2.4, excess kurtosis = −1.0968 (that is kurtosis is 1.9032). This implies that the tail is much lighter compared to a normal distribution. Figure 2.2 shows shapes of platykurtic, mesokurtic and leptokurtic distributions. In Microsoft Excel, *KURT(array)* can be used to calculate excess kurtosis. Excess kurtosis is used to measure deviation from normal distribution and is a useful metric in pathology.

2.9 | Data Visualization

Data visualization is an integral part of descriptive analytics; it assists decision-makers with useful insights and is useful for deriving new variables or features. There are many charts such as histogram, bar chart, pie chart and box plot which assist data scientists with visualization of the data. In the recent years, bubble charts, tree maps and sunburst maps, which can create insightful visuals of data, have become popular among analytics experts. Data is used in dashboards frequently used by organizations to continuously monitor key performance indicators. It is always advisable to start an analytics project with data visualization, since it can provide insights for developing predictive analytics models as well as for feature selection and feature engineering.

2.9.1 Histogram

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

1. Divide the data into a finite number of non-overlapping and consecutive bins (interval). The total number of bins to be used can be calculated using Eqs. (2.16) or (2.17).
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.

Histograms are very useful since they can be used to identify or assess the following:

1. The shape of the distribution and 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.
5. The presence of outliers.

One of the first steps in constructing a histogram is identifying the number of bins. Many different formulas are used in literature, one of the simplest being

$$\text{Number of bins } N = \frac{X_{\max} - X_{\min}}{W} \tag{2.16}$$

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

$$\text{Number of bins } N = \lfloor 1 + 3.3 \log_{10} n \rfloor \tag{2.17}$$

where n is the total number of observations in the dataset. $\lfloor N \rfloor$ is the closest integer less than or equal to N (known as floor function). Figures 2.3 and 2.4 show the histogram of movie budget in crores of rupees (1 crore = 10 million) and box office collections respectively, based on data of 149 Bollywood movies (Data file: Bollywood.Xls).

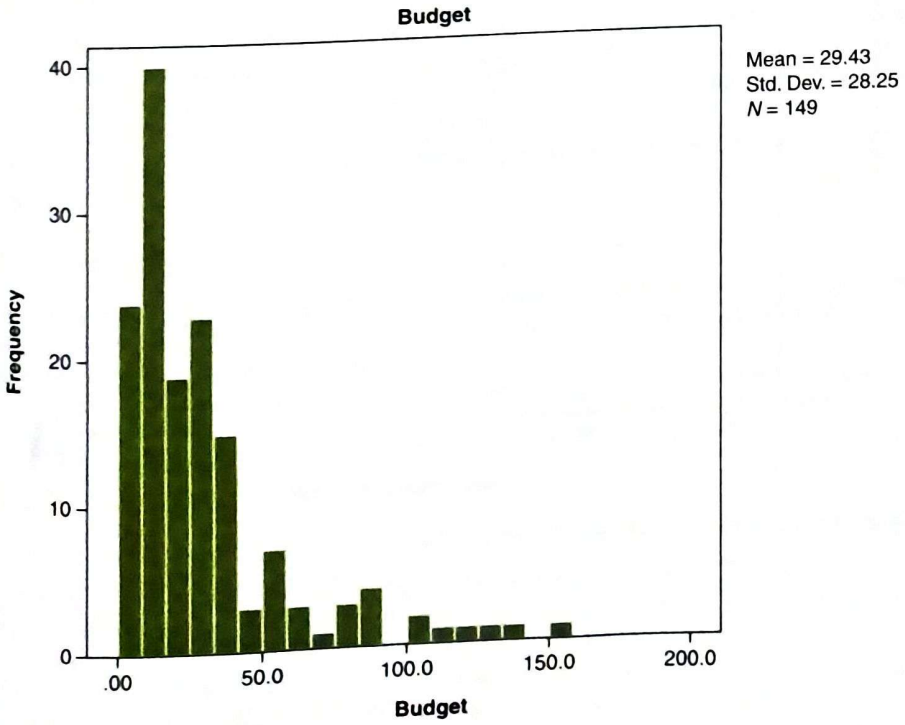

Figure 2.3 | Histogram of Bollywood movie budget.

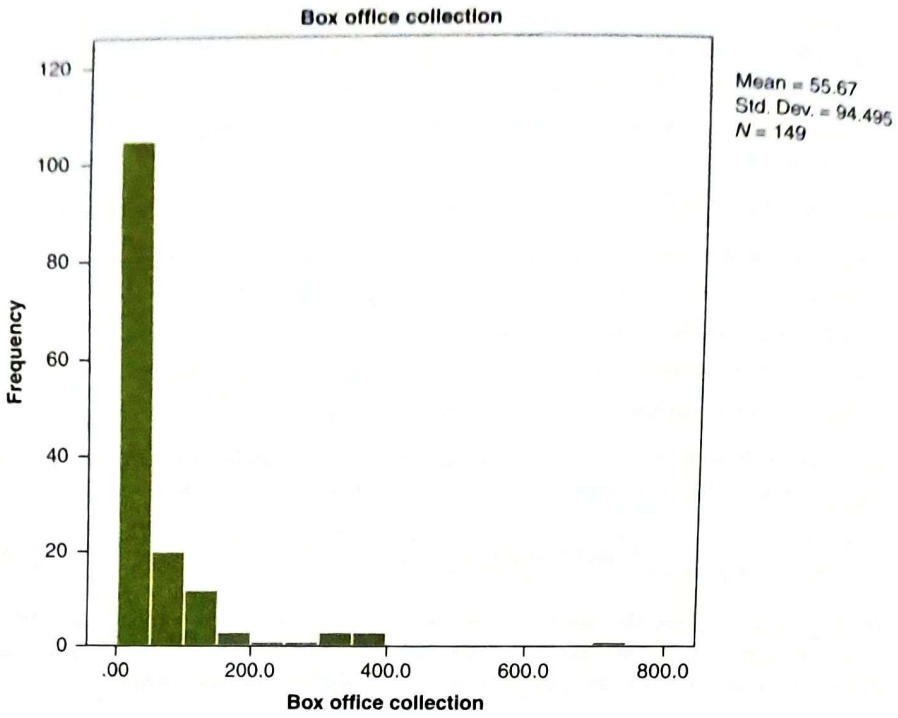

Figure 2.4 | Histogram of Bollywood movie box office collection.

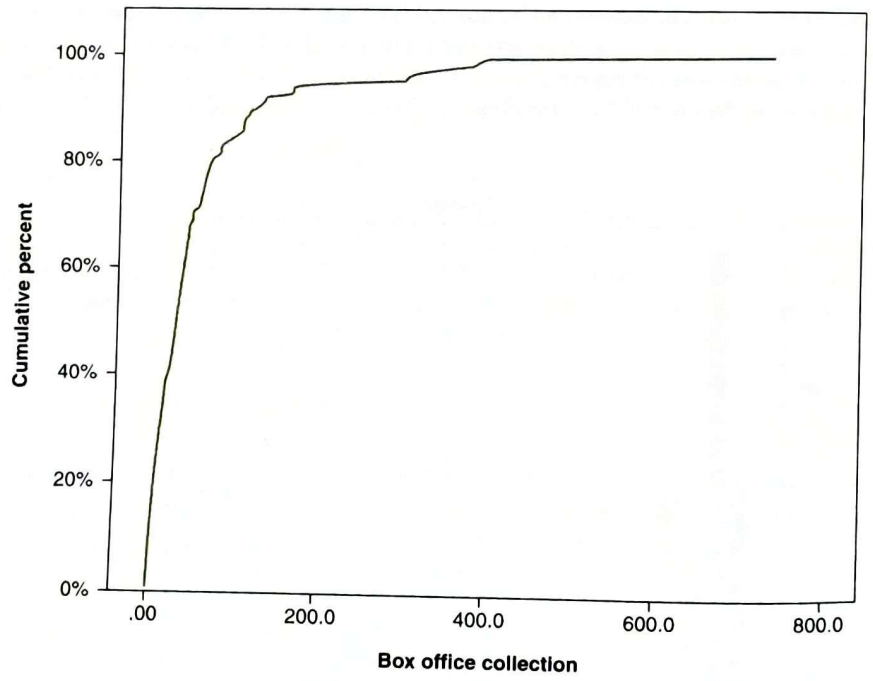

Figure 2.5 | Ogive curve for box office collection.

From Figure 2.3, we can infer that the budget for a large proportion of movies is less than ₹50 crore, and it is a right-skewed distribution (that is, a long tail on the right side). In Figure 2.4, we can also see an outlier where the box office collection is more than ₹700 crore (movie PK, featuring Amir Khan and directed by Rajkumar Hirani). The cumulative histograms are called **Ogive curves**. The Ogive curve for Bollywood box office collections is shown in Figure 2.5. The values on the Ogive curve provide the cumulative percentage of records on the vertical axis corresponding to a value of the variable on the horizontal axis.

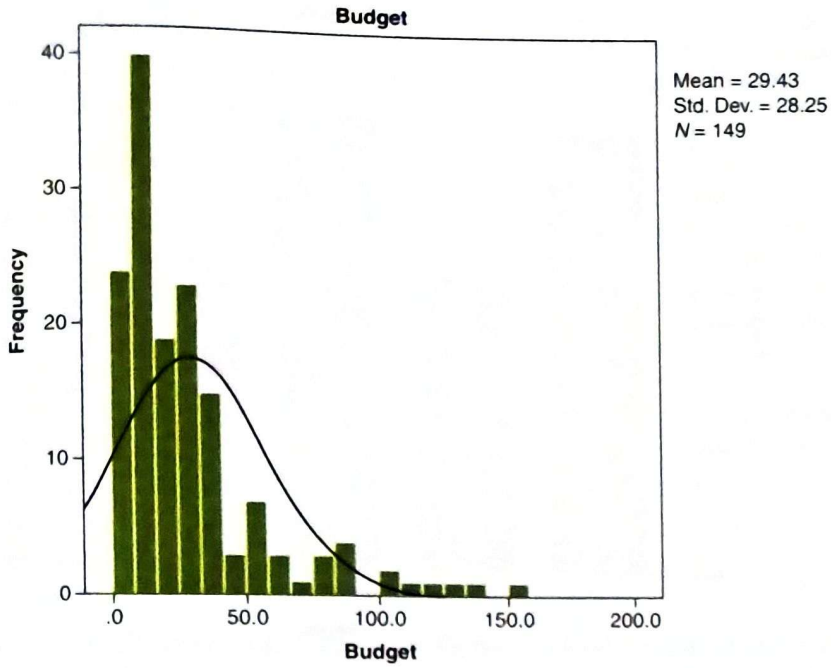

Figure 2.6 | Histogram of Bollywood movie budgets along with normal distribution frequency.

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 budgets superimposed with normal distribution. It is obvious that the frequency distribution of budgets is not normal. The distribution in Figure 2.6 shows a positive skewness.

2.9.2 Bar Chart

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

2.9.3 Pie Chart

Pie charts are circular charts that display the proportion of each category in the dataset and are mainly used for categorical data. The pie chart for movie genres based on the Bollywood movies dataset is shown in Figure 2.8.

Pie charts help visualize the proportion (percentage) of each category as a sector of a circle.

2.9.4 Scatter Plot

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

Figure 2.9 shows a linear relationship between budget and box office collection, as well as the existence of an outlier (the movie PK, for which the box office collection was more

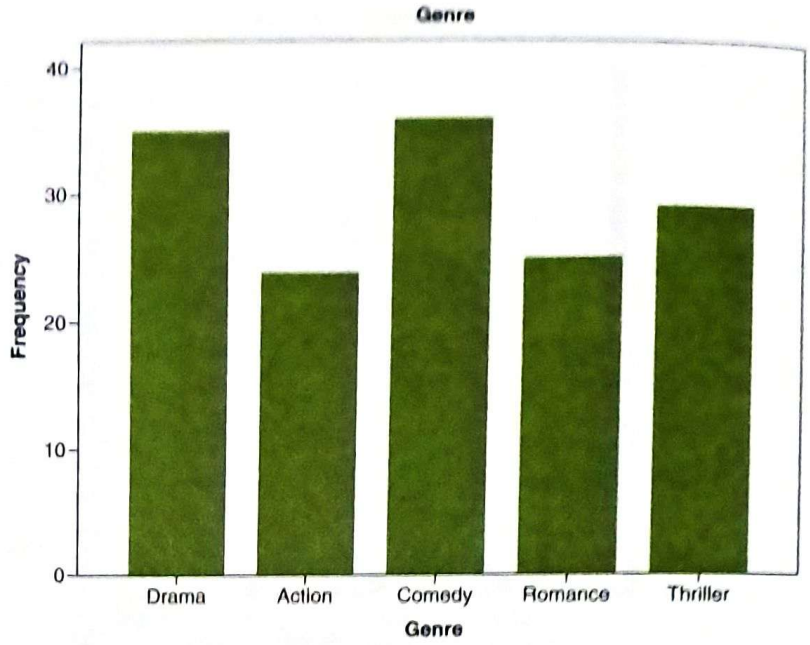

Figure 2.7 | Bar chart for movie genre.

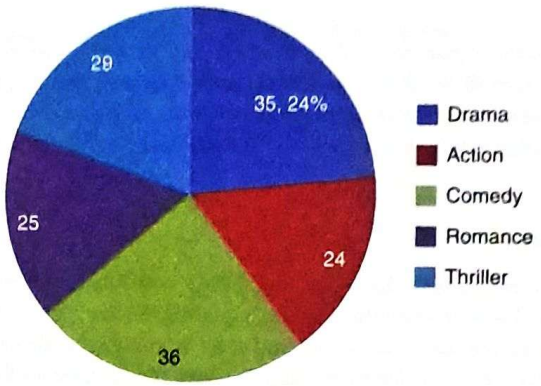

Figure 2.8 | Pie chart for movie genre.

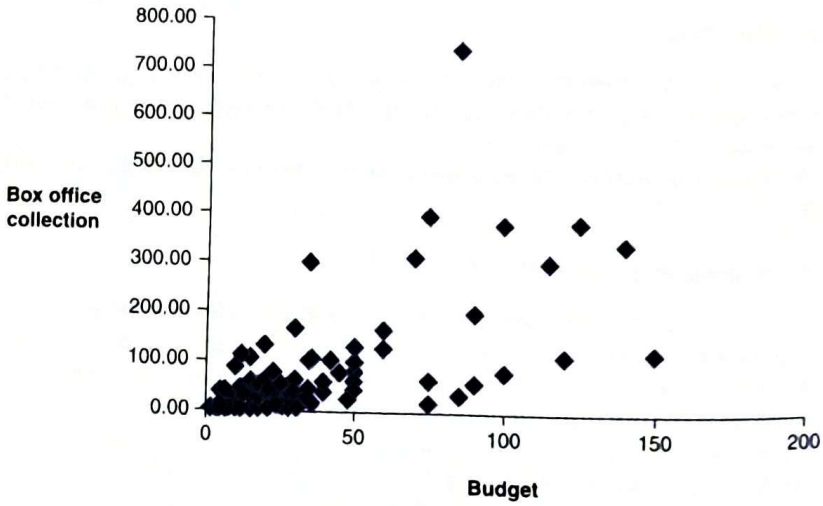

Figure 2.9 | Scatter plot between movie budget and box office collection.

than ₹700 crore – much higher than the average box office collection). Scatter plots are used during regression model-building to decide on the initial model. That is, scatter plot helps data scientists to decide the functional form of the relationship between the dependent (outcome) and independent (feature) variables.

2.9.5 Coxcomb Chart

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

Nightingale collected data from the Crimean War (between the British and the French on one side and the Russians on the other) on the 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 Nightingale), the largest area of the chart corresponds to the cause 'preventable diseases'.

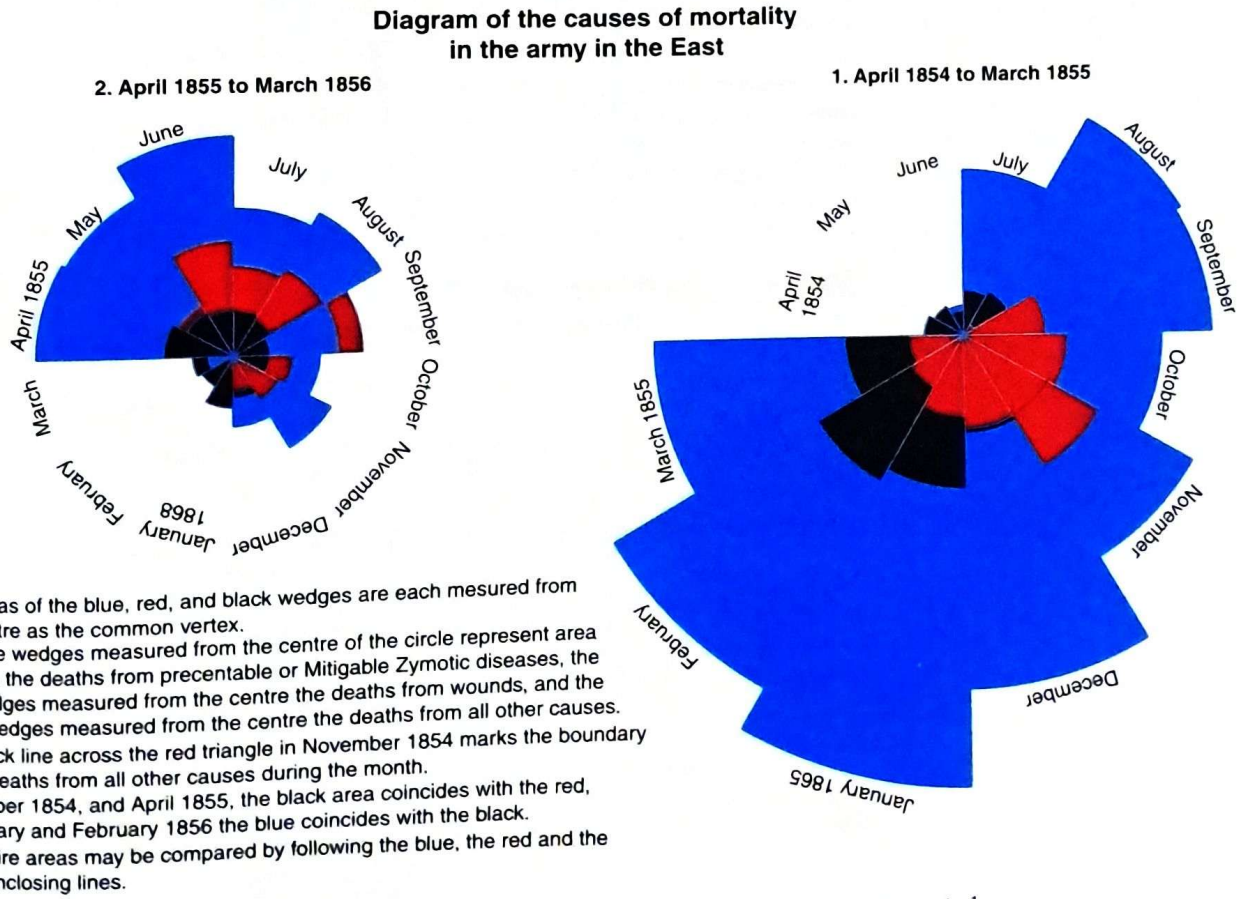

Figure 2.10 | Coxcomb chart on causes of mortality in the army, prepared by Florence Nightingale.[1]

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

2.9.6 Box Plot (or Box and Whisker Plot)

A 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. A 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 first Quartile (Q_1), Median, third Quartile (Q_3), IQR, minimum and maximum values. The box plot for the data in Table 2.4 is shown in Figure 2.11.

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 ($Q_1 - 1.5$ IQR and $Q_3 + 1.5$ IQR) are potential outliers. The box plot for the Bollywood movie budget (data file: Bollywood.xlsx) is shown in Figure 2.12.

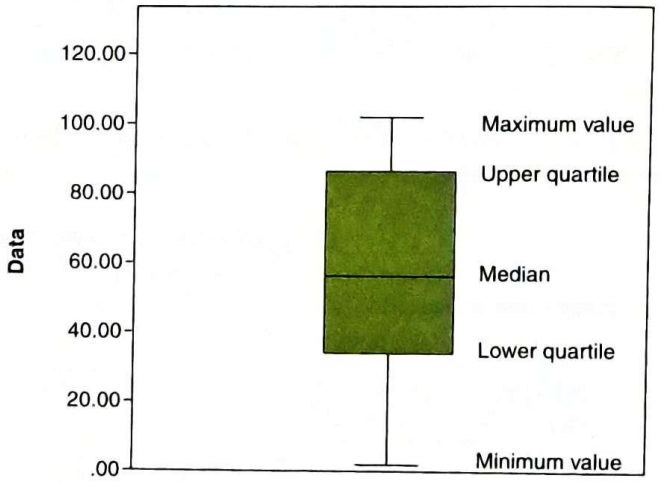

Figure 2.11 | Box plot of the data in Table 2.4.

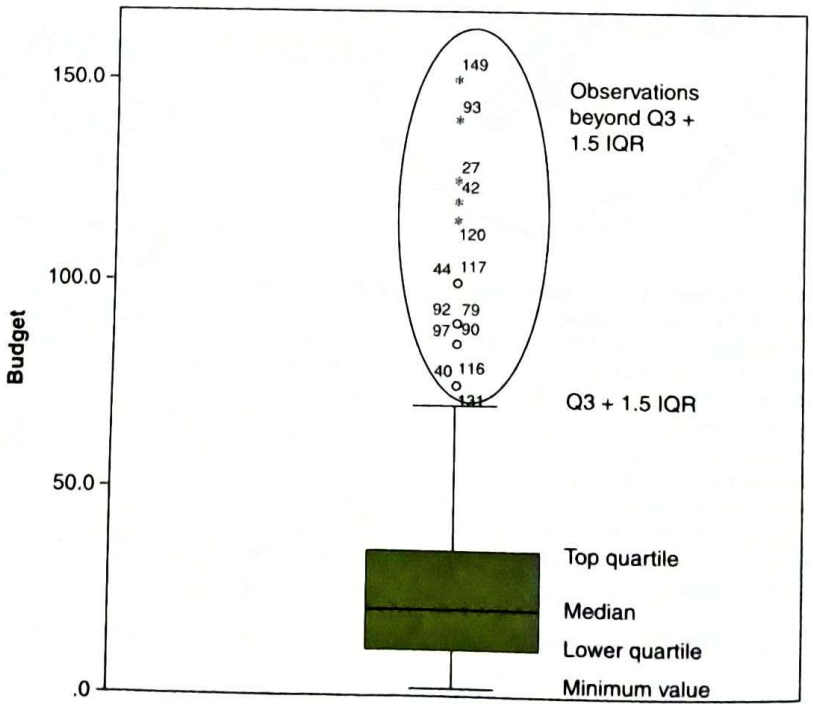

Figure 2.12 | Box plot for Bollywood movie budgets.

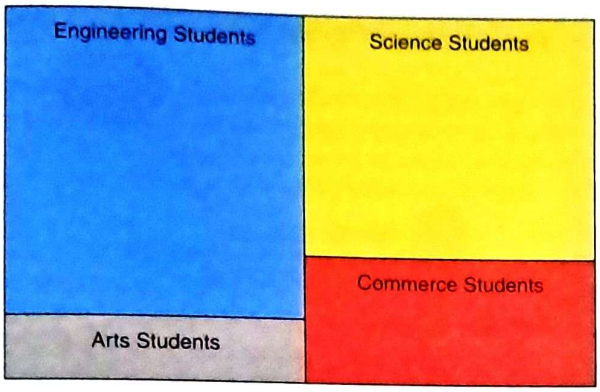

Figure 2.13 | Tree map of student disciplines at the undergraduate level.

The position of the lowest whisker in Figure 2.12 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 $Q_3 + 1.5\text{IQR} = 71$. All the observations beyond $Q_3 + 1.5\text{IQR}$ shown above the upper whisker are outliers.

2.9.7 Tree Map

A tree map is a hierarchical map made up of nested rectangles, frequently used as part of business intelligence reports, which helps organizations understand data hierarchically. To construct a tree map, the data should be hierarchical, containing several levels. The size and colour of each rectangle are used to describe/differentiate the characteristics of the data. A sample tree map is shown in Figure 2.13, in which the undergraduate level academic discipline of students admitted into an MBA programme is captured. The size of the rectangle captures the 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 an engineering background, whereas the area corresponding to arts students is the smallest, indicating that the least number of students in the MBA programme have arts background.

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

2.9.8 Bubble Chart

Bubble charts are usually three-dimensional charts which, along with the usual horizontal and vertical axes, use the size of the bubbles to represent a third dimension. Table 2.7 shows the data on number of people infected, recovered and killed due to a virus. The corresponding bubble chart is shown in Figure 2.14. The bubble chart in Figure 2.14 is created using Microsoft Excel. The option for bubble chart is available under scatter plot.

However, not all bubble charts have three dimensions. An example is bubble cloud, a variant of the bubble chart in which the bubbles are packed together, and the area of the bubble captures the data. In Figure 2.15, a bubble cloud is created using the margin of victory during the 2014 parliamentary elections in various constituencies in the state of Haryana (data file: Figure 2.15 Election Data.Xlsx).

Table 2.7 | Data on number of infected, deaths, and recovered due to virus

State	Infected Cases	Deaths	Recovered
Maharashtra	44,582	1,517	24,320
Tamil Nadu	14,753	98	4,300
Gujarat	13,268	802	9,020
Rajasthan	6,494	153	3,290

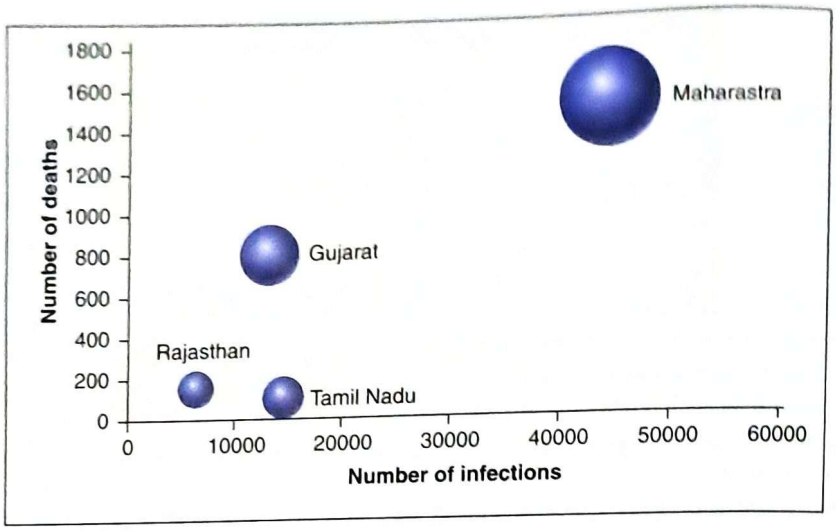

Figure 2.14 | Bubble chart.

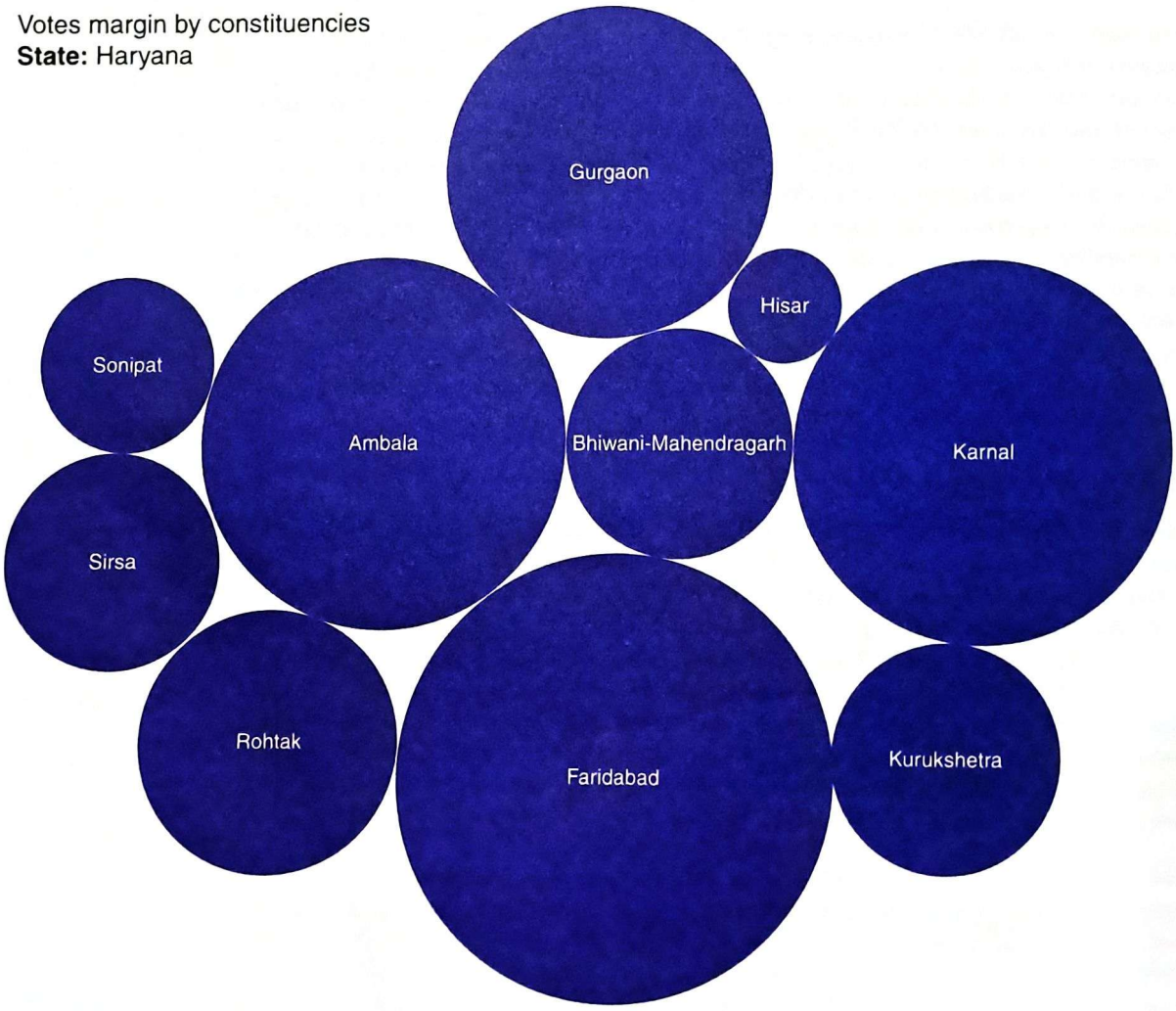

Figure 2.15 | Bubble cloud based on victory margin in 2014 parliamentary elections.

6

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

After reading this chapter, you will be able to:

- Understand hypothesis testing and its importance in analytics, especially in statistical learning algorithms.
- Learn to set up a hypothesis test, understand the concept of null and alternative hypotheses.
- Understand the concept of significance (α), probability value (p-value), Type-I and Type-II errors.
- Understand the association between central limit theorem and test statistic in one-sample Z-test and t-test.
- Understand simple one-sample hypothesis test for population mean when population variance is either known or unknown.
- Learn to conduct two-sample hypothesis tests and their applications in analytics.
- Understand the role of non-parametric tests such as chi-square test of independence.
- Learn goodness of fit tests and their application in identifying best probability distribution to describe a dataset.

Hypothesis Testing

Hypothesis testing is one of the most important concepts in analytics, but also a concept many students of statistics and analytics may find difficult to understand. A hypothesis, in statistics, is a claim made by a person/organization, usually about population parameters such as mean or proportion. We seek evidence against the claim from a sample (for example, a claim could be made that the average salary of analytics experts is at least USD 100,000. Another hypothesis could be that the mortality rate due to coronavirus amongst the female population is less than that of the male population). Hypothesis testing is a process used to either reject or retain a hypothesis. In statistical learning algorithms, feature selection is achieved using hypothesis testing. For example, in multiple linear regression, feature selection is carried out using partial F-test (discussed in Chapter 10) and in logistic regression, feature selection is achieved through likelihood ratio test (discussed in Chapter 11)

6.1 | Introduction to Hypothesis Testing

Blackout Babies

On 9 November 1965, a power failure resulted in a 12-hour-long blackout in New York and surrounding areas. Nine months later, in August 1966, The New York Times published a series of articles claiming that the birth rates in August 1966 were 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 about whether power failures result in procreation. Izenman and Zabell (1981) used time series data analysis to claim 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 have been made about the impact of power cuts on baby booms since then (Anon 2009, Fetzer *et al.*, 2013).

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 a statistically significant relationship between the dependent (outcome) and independent (feature) variables. However, in goodness of fit tests, used for checking whether the data follows a specific distribution or not, we would like to retain the null hypothesis.

A hypothesis is a claim or belief; **hypothesis testing** is the statistical process of either rejecting or retaining the claim. The concept of hypothesis testing was developed by Fisher (1925, 1935), and Neyman and Pearson (1933). Fisher used the term 'significance testing'; Neyman and Pearson created a formal framework for hypothesis testing. Hypothesis test consists of two complementary statements called *null hypothesis* and *alternative hypothesis*, only one of which can hold true (Neyman and Pearson, 1933). Hypothesis testing is one of the most important concepts in analytics, due to its role in inferential statistics and feature selection in statistical learning algorithms. It is used to establish evidence of an association relationship between an outcome variable and predictor variables.

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

1. Children who drink Complan (a health drink owned by the company Heinz in India) are likely to grow taller at a faster rate compared to children who do not drink Complan.
2. If you drink Horlicks, you can grow taller, stronger and sharper (3 in 1).
3. Using Fair and Lovely/Handsome cream can make one fair and lovely/handsome.
4. Wearing deodorant makes you attractive to the opposite gender (known as *Axe effect*).
5. Women take more selfies compared to men (Freier, 2016).
6. Beautiful people are likely to have female children (Miller and Kanazawa, 2007). This is one of my favourite hypotheses since I have a daughter and I can claim that I am good-looking 😊.
7. Married people are happier than people who are single (Anon, 2015), especially those who have married their best friend (many married people may not agree!).
8. Vegetarians miss fewer flights (Siegel, 2016).
9. Smokers are better salespeople.
10. Meditating leads to a longer lifespan (Schneider *et al.*, 2005)

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 claims are actually true? Hypothesis testing is used to check the validity of a claim using evidence found in 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 hypotheses or suggest hypotheses. In the case of confirmatory data analysis, the objective is to test the validity of a hypothesis (confirm whether it is true or not) using techniques such as hypothesis testing and regression. Hypothesis testing is classified into parametric hypothesis testing (in which the test is about a population parameter) and non-parametric hypothesis testing (in which

The Lady Tasting Tea (Fisher, 1935)

Dr Blanche Muriel Bristol, a friend of the famous statistician R A Fisher, claimed that she could tell by tasting tea whether the milk or the tea had been poured first. During a tea party hosted by Fisher, Dr Bristol refused the tea given to her, claiming that she preferred tea in which the milk had been poured first. Fisher rejected Dr Bristol's claim that the flavour of tea can be affected by the order in which tea and milk are added (Box, 1978). Fisher devised an experiment to prove that Dr Bristol must be just guessing. Dr Bristol was randomly given eight cups during this experiment – out of which four had been prepared by pouring tea first, and the remaining by pouring milk first. The claim that Dr Bristol was just guessing whether the tea or milk had been poured first was called the *null hypothesis*. The term 'null hypothesis' was first introduced in Fisher's book of design of experiments in 1935. It can be interpreted as there being no relationship between a characteristic (such as gender) and an outcome (mortality rate due to coronavirus). In the 'lady tasting tea' experiment, the outcome is the taste of the tea and the characteristic is milk or tea being poured first.

In this test of tasting tea with eight cups, the probability of correctly identifying all four cups in which tea had been poured first is 1/70 (since four out of eight cups served had tea poured first, and there are ${}^{8}C_{4}$ (=70) ways of selecting four cups out of eight). Dr Bristol was able to correctly identify all four cups in which tea was poured first – the probability of a person identifying all four cups correctly by chance alone is 0.014. Fisher introduced this as 'the lady tasting tea' experiment in his book on the design of experiments (Fisher, 1935). This formed the basis for the development of hypothesis testing.

the test is about other characteristics such as the distribution of the data). The following steps are used in hypothesis testing:

1. Describe the hypothesis in words. In parametric test, a hypothesis is described using a population parameter (such as mean, standard deviation, proportion etc.) about which a claim (hypothesis) has been made. A few examples of hypothesis are:
   (a) The average time spent by women using social media is greater than that spent by men.
   (b) On average, women upload more photos on social media than men.
   (c) Customers of mobile phone service providers with more than one mobile handset are more likely to churn.
   (d) The average mortality rate due to coronavirus is more for male compared to female.
2. Based on the claim made in Step 1, define the null and alternative hypotheses. Initially, we believe that the null hypothesis is true. In general, null hypothesis means 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 on social media'). Null and alternative hypotheses are defined using a population parameter.
3. Identify the test statistic to be used for collecting evidence against the null hypothesis. Test statistic is the standardized difference between the estimated value and the hypothesis value. Test statistic will enable us to calculate the evidence against the null hypothesis using probability value (p-value). The test statistic will depend on the probability distribution of the sampling distribution; for example, if the test is for population mean value and the mean is calculated from a large sample with a known population standard deviation, then the sampling distribution will be a normal distribution and the test statistic will be a Z-statistic (that is, it follows a standard normal distribution).
4. Decide the criteria for rejection and retention of null hypothesis. This is called *significance value*, traditionally denoted by the symbol α. The value of α will depend on the context; usually 0.1, 0.05, and 0.01 are used. The significance value α corresponds to the maximum allowable value of Type I error (discussed in Section 6.4).
5. Calculate the p-value (probability value), which is the conditional probability of observing a test statistic value as extreme as the one observed in the sample, given the

null hypothesis is true. In simple terms p-value is the evidence against the null hypothesis. A small p-value (for example, less than 0.05) indicates significant evidence against the null hypothesis. In other words, the evidence is strongly against the null hypothesis. P-value was proposed by Fisher (1950) as an index that can be used to decide the strength of evidence against null hypothesis (Dahiru, 2008).

> P-value is an often misused and misunderstood concept in statistical inference. We would like to provide the following definition for p-value.
>
> According to Fisher (1925):
>
> - 'P-value represents the probability of obtaining an effect size equal to or more extreme than the one observed in the data considering the null hypothesis is true.'
> - P-value thus provides quantitative strength of evidence against the null hypothesis.
>
> According to Biau *et al.* (2009):
>
> - 'The effect can be a measurement of difference between two groups or any measure of association between two variables.'

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 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, as in the examples listed below:

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

6.2.2 Null and Alternative Hypothesis

Hypothesis testing 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 the alternative hypothesis, she/he may like to reject the null hypothesis by collecting evidence against null hypothesis.

Null hypothesis, usually denoted as H_0 (H zero and H naught), refers to the claim that there is no relationship or 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.

> According to R A Fisher (1935):
>
> "Null hypothesis is never proved or established, but is possibly disproved, in the course of experimentation. Every experiment may be said to exist only in order to give the facts a chance of disproving the null hypothesis".

Alternative hypothesis, usually denoted as H_A (or H_1), is the complement of the null hypothesis. Alternative hypothesis is what the researcher believes to be true. The term alternative hypothesis was coined by Neyman and Pearson (1933). 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	The average annual salary of machine learning experts is different for males and females.	$H_0: \mu_m = \mu_f$ $H_A: \mu_m \neq \mu_f$
	(In this case, the null hypothesis is that there is no difference in the salaries of male and female machine learning experts)	μ_m and μ_f are average annual salary of male and female machine learning experts respectively.
2	On average, people with a Ph.D. in analytics earn more than those with a 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.

6.2.3 Test Statistic

A test statistic is a standardized value that measures the distance (in terms of number of standard deviations) between the value of the parameter estimated from the sample(s) and the value of the parameter in null hypothesis. The test statistic is used for calculating the ***p*-value**.

The *p*-value is the conditional probability of observing a statistic value as extreme as the one observed in the data when the null hypothesis is true. For example, consider the following research hypothesis: The average annual salary of machine learning experts is at least \$100,000. The corresponding null hypothesis is H_0: $\mu_m \leq$ 1,00,000. Assume that the estimated value of the salary from a sample is 110,000 (that is $\bar{X} = 1{,}10{,}000$) and that the population standard deviation is known and that the standard error of the sampling distribution is 5,000 (that is, $\sigma/\sqrt{n} = 5{,}000$, where n is the sample size using which $\bar{X} = 1{,}10{,}000$ was calculated). The standardized distance between the estimated salary and the hypothesized salary is (1,10,000 – 100,000)/5,000 = 2. That is, the standardized distance between the estimated value and the hypothesis value is 2. We can find the probability of observing such statistic value (that is $Z \geq 2$) 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 $\frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$. Standard normal distribution and the *p*-value corresponding to $Z \geq 2$ are shown in Figure 6.1.

Note that the ***p*-value** is a conditional probability. It is the conditional probability of observing the statistic value as extreme as the one observed in test statistics given that the null hypothesis is true. *p*-value is the evidence against the null hypothesis.

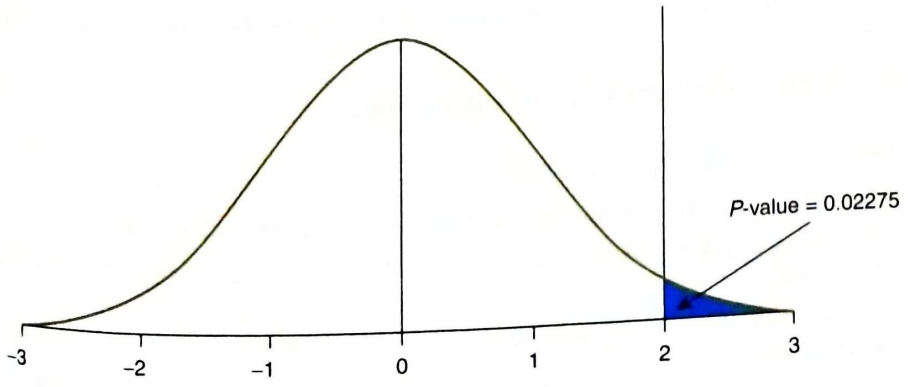

Figure 6.1 | Standard normal distribution and *p*-value.

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

The probability of observing a value of 2 or higher from a standard normal distribution is 0.02275. That is, if the population mean is 100,000 and the standard error of the sampling distribution is 5,000, then the probability of observing a sample mean greater than or equal to 110,000 is 0.02275. The value 0.02275 is the p-value, which is the evidence against the null hypothesis. That is, there is only 2.275% chance that the null hypothesis is true.

$$p\text{-value} = P(\text{Observing test statistic value as extreme as one observed in the data} \mid \text{null hypothesis is true}) \quad (6.1)$$

Alternatively, for the example discussed in the above paragraph:

$$p\text{-value} = P(\text{observing test statistic value} \geq 2 \mid \mu_m \leq 1{,}00{,}000)$$

6.2.4 Decision Criteria – Significance Value

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.

The primary task in hypothesis testing is to take the decision to either reject the null hypothesis or fail to reject (retain) it; we therefore need some criteria to take this 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 the p-value. The decision to reject or retain will depend on whether or not the calculated p-value crosses the threshold value α. The decision criteria are 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 the reason that we start the process of hypothesis testing with the assumption that the null hypothesis is true. Unless there is strong evidence against this assumption, we will not reject the null hypothesis. The value of the statistic for which the probability under the sampling distribution equals 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), we reject the null hypothesis, whereas if the statistic value is less than the critical value, we retain the null hypothesis. In case of the left-tailed test, if the calculated statistic value is less than the critical value (p-value will be less than α-value), we reject the null hypothesis, whereas if the statistic value is greater than the critical value, we retain the null hypothesis. The areas beyond the critical values are known as *rejection region*.

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

6.3 | One-Tailed and Two-Tailed Test

Consider the following three hypotheses:

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

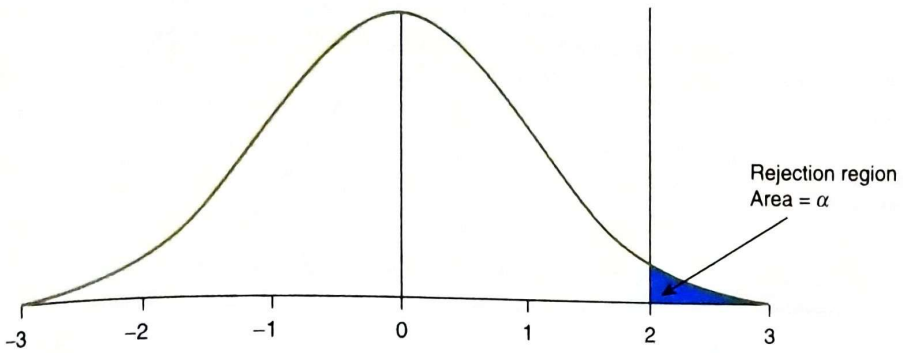

Figure 6.2 | Right-tailed hypothesis test's rejection region.

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

$$H_0: \mu_m \leq 1{,}00{,}000$$

$$H_A: \mu_m > 1{,}00{,}000$$

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, the 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 $\mu_m > 1{,}00{,}000$, this is called a *right-tailed test*.

Statement 2 – The Average waiting time at the 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 the Heathrow security check. In this case, the rejection region will be on the left side (known as *left-tailed test*) of the distribution as shown in Figure 6.3.

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

$$H_0: \mu_m = \mu_f$$

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

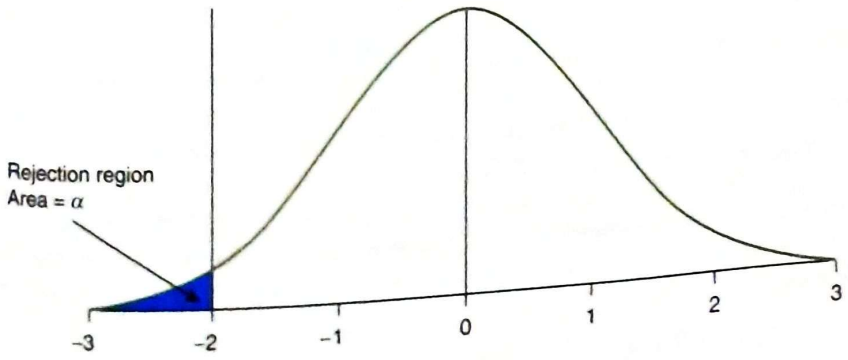

Figure 6.3 | Rejection region in case of left-sided test.

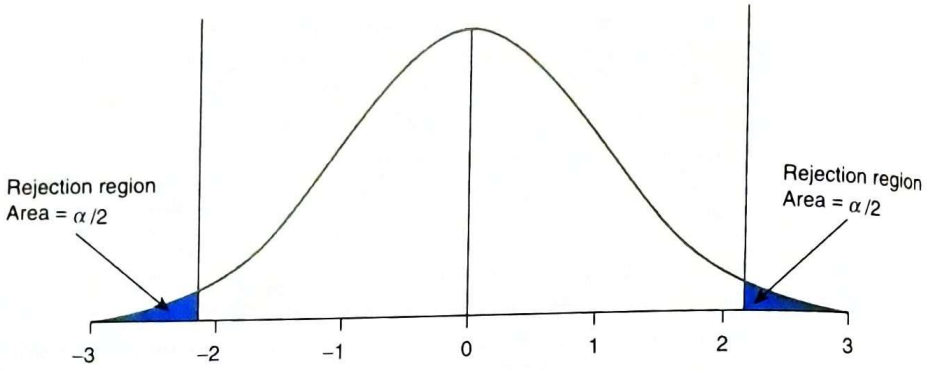

Figure 6.4 | Rejection region in case of two-tailed test.

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.

> An easy way of identifying the direction of the test is by checking the sign of the parameter in the alternative hypothesis. If the alternative hypothesis has > sign ($\mu_m > 1{,}00{,}000$) then it is a right-tailed test. If the alternative hypothesis has < sign ($\mu_w < 30$) then it is a left-tailed test. If the alternative hypothesis has ≠ sign ($\mu_m \neq \mu_f$) then it is a two-tailed test.

6.4 | Type I Error, Type II Error, and Power of the Hypothesis Test

In hypothesis testing, we are faced with the following options:

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 the 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 maximum 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 against the null hypothesis, since a p-value of 0.03 implies that there is only a 3% chance that the null hypothesis is true. The 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 the probability of observing the data given the null hypothesis is true, whereas the significance level α refers to incorrect rejection of null hypothesis when it is true under *repeated trials*.

2. **Type II Error:** The 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:

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

Table 6.3 | Description of Type I error, Type II error, and the power of test

	Decision made about null hypothesis based on the hypothesis test	
Actual value of H_0	Reject H_0	Retain H_0
H_0 is true	Type I error	Correct decision
	$P(\text{Reject } H_0 \mid H_0 = \text{true}) = \alpha$	$P(\text{Retain } H_0 \mid H_0 = \text{true}) = (1 - \alpha)$
H_0 is false	Correct decision (power of test)	Type II error
	$P(\text{Reject } H_0 \mid H_0 = \text{false}) = 1 - \beta$	$P(\text{Retain } H_0 \mid H_0 = \text{false}) = \beta$

The value $(1 - \beta)$ is known as the *power of the 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 I error, Type II error, and the power of hypothesis test is shown in Table 6.3.

6.5 | Hypothesis Testing for Population Mean when Population Variance is Known: One-Sample *Z*-Test

A **one-sample *Z*-test** is used when a claim (hypothesis) is made about a population parameter 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), we know that the sampling distribution of mean for a large sample from an independent and identically distributed population with mean μ and standard deviation σ follows a normal distribution with mean μ and standard deviation $\sigma/\sqrt{n}$. The standardised value $\dfrac{\bar{X} - \mu}{\sigma/\sqrt{n}}$ follows a standard normal distribution. *Z*-test exploits CLT to conduct a hypothesis test for population mean when the population variance is known; the test statistic 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 to find the critical *Z*-value for a left-tailed test. NORMSINV($1 - \alpha$) [and NORM.S.INV($1 - \alpha$)] will give the critical *Z*-value for a right-tailed test. NORMSINV($\alpha/2$) and NORM.S.INV($1 - \alpha/2$) will give critical *Z*-values for a two-tailed test. The decision criteria for rejection or retention of the null hypothesis is described in Table 6.5.

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 when the sample size $n > 30$.

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-statistic| > |Critical value|	Reject H_0
	|Z-statistic| ≤ |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 the city is greater than ₹4,200, with a standard deviation of ₹3,200. From a random sample of 40,000 families, the average disposable income was estimated as ₹4,250. Assume that the population standard deviation is known to be ₹3,200. Conduct an appropriate hypothesis test at 95% confidence level ($\alpha = 0.05$) to check the validity of the claim by the agency.

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

Solution

Claim: Average disposable income is more than ₹4,200.

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

$$H_0: \mu \leq 4{,}200$$

$$H_A: \mu > 4{,}200$$

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

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} = \frac{4{,}250 - 4{,}200}{3{,}200/\sqrt{40{,}000}} = 3.125$$

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.

This is a right-tailed test (sign in alternative hypothesis is greater than). The corresponding Z-critical value at $\alpha = 0.05$ for right-tailed test is approximately 1.64 [in Excel NOMRSINV($1 - \alpha$) 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.

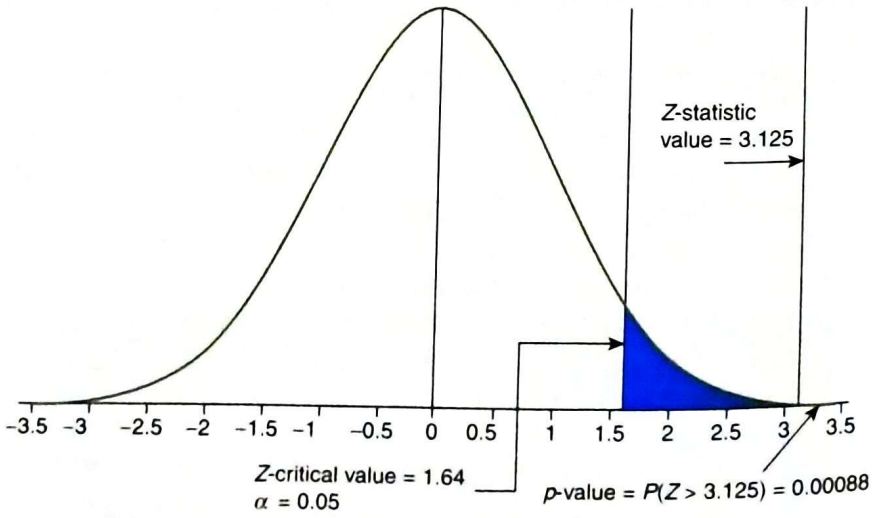

Figure 6.5 | Critical value, Z-statistic value, and corresponding p-value.

7 Analysis of Variance

Analysis of variance is not a mathematical theorem,
but rather a convenient method of arranging the arithmetic.

—**Ronald Fisher**

LEARNING OBJECTIVES

After reading this chapter, you will be able to:

- Understand the need for analysis of variance (ANOVA).
- Understand the difference between two-sample *t*-test for mean and ANOVA.
- Understand one-way ANOVA and calculation of *F*-statistic.
- Understand computation within the group variation, between the group variation, and *F*-statistic
- Learn to conduct a two-way ANOVA, and the computations involved in conducting a two-way ANOVA.

Analysis of Variance (ANOVA)

In many situations, we may have to conduct a hypothesis test to compare mean values simultaneously for more than two groups (comparing more than two population means) created using a factor (or factors). For example, a marketer might want to understand the impact of three different discount values (such as 0%, 10% and 20%) on the average quantity of sales. When we have to compare the impact of a factor on the mean in multiple groups (created by different levels of the factor) simultaneously, hypothesis tests such as two-sample *t*-tests discussed in Chapter 6 are not an ideal approach since they can result in an incorrect estimation of Type I and Type II errors. We use the ANOVA to understand the differences in population means among more than two populations.

7.1 | Introduction to ANOVA

Consider a retail store which would like to study the impact of different levels of price discounts (factor) on the sales (outcome variable) of a specific product or brand. Price discount can range from 0% to 100% (theoretically). For ease of understanding, assume that the levels of discounts are 0%, 10% and 20%. The marketing manager would like to understand whether the variable 'price discount' has any significant impact on the average sales quantity. Such studies are called *single-factor experimental design* (R A Fisher, 1934, 1935). Different discount rates correspond to different levels of the factor and different levels (such

The objective of ANOVA is to check simultaneously whether population mean from more than two populations are different.

as 0%, 10% and 20%) are assigned randomly to different units. In the case of price discounts, units refer to different days chosen randomly, since the quantity of sales may also depend on the day of the week. It is possible that the weekend sales quantity may be higher than the sales quantity on weekdays. In many cases, we may deal with observational studies in which we observe the impact of a factor on a variable. For example, the impact of specialization in MBA such as analytics, finance, marketing, etc. on the income of the graduates upon graduation. Here the specialization in MBA chosen by the students is unlikely to be under the control of the researcher. To understand whether the factor (different levels of a factor) has any statistical significance on the population parameter, we compare two models as described below:

A non-zero τ_i value in Equation (7.2) implies that the factor has influence on the value of the outcome variable Y_{ij}.

In ANOVA, our objective is to verify whether the variation due to treatment is different from the variation due to randomness.

1. **Means Model:** It is given by

$$Y_{ij} = \mu + \varepsilon_{ij} \tag{7.1}$$

where Y_{ij} is the value of the outcome variable of j^{th} observation for i^{th} factor level, μ is the overall mean value of all observations, ε_{ij} is the error assumed to be a normal distribution with mean 0 and standard deviation σ. The model defined in Equation (7.1) is often called the *reduced model*, in which the mean μ is common for all levels of the factor.

Since we assume that the error ε_{ij} is normally distributed with mean 0 and standard deviation σ, the outcome variable Y_{ij} is normally distributed with mean μ and standard deviation σ.

2. **Factor Effect Model:** It is given by

$$Y_{ij} = \mu + \tau_i + \varepsilon_{ij} \tag{7.2}$$

In Equation (7.2), μ is the overall mean and τ_i is the effect of factor i (or factor effect). τ_i is the difference between the overall mean and the factor level mean (or deviation from the overall mean). Our interest in this case would be to check whether the values of τ_i are different from zero. The model in Equation (7.2) is called a *full model*; the reduced model in Equation (7.1) is a special case of the model defined in Equation (7.2) in which τ_i is zero for all i.

7.2 | Multiple *t*-Tests for Comparing Several Means

Continuing with the example from Section 7.1, if we had only two values for 'price discount', then we could have used the two-sample *t*-test to check whether there is a statistically significant relationship between price discount and average sales quantity. When we have more than two levels of discounts, one option is to compare the population parameters two at a time (two discount values). For example, we can compare each of the following three cases using a two-sample *t*-test:

1. Test between 0% and 10%
2. Test between 0% and 20%
3. Test between 10% and 20%

The Type I error will be estimated incorrectly if we conduct the three different tests listed above. For example, assume that the mean sale (population mean) at 0%, 10% and 20% discount is μ_0, μ_{10} and μ_{20} respectively. Consider three two-sample *t*-tests shown in Table 7.1.

Table 7.1 | Three different two-sample *t*-tests

Test	Null Hypothesis	Alternative Hypothesis	Significance (α)
A	$H_0: \mu_0 = \mu_{10}$	$H_A: \mu_0 \neq \mu_{10}$	$\alpha = 0.05$
B	$H_0: \mu_0 = \mu_{20}$	$H_A: \mu_0 \neq \mu_{20}$	$\alpha = 0.05$
C	$H_0: \mu_{10} = \mu_{20}$	$H_A: \mu_{10} \neq \mu_{20}$	$\alpha = 0.05$

Let

$$P(A) = P(\text{Retain } H_0 \text{ in test A} | H_0 \text{ in test A is true})$$

$$P(B) = P(\text{Retain } H_0 \text{ in test B} | H_0 \text{ in test B is true})$$

$$P(C) = P(\text{Retain } H_0 \text{ in test C} | H_0 \text{ in test C is true})$$

Note that values of $P(A) = P(B) = P(C) = 1 - \alpha = 1 - 0.05 = 0.95$.

The conditional probability of simultaneously retaining all three null hypotheses when they are true is $P(A \cap B \cap C) = 0.8573$. Now, consider the following null hypothesis:

$$H_0: \mu_0 = \mu_{10} = \mu_{20} \tag{7.3}$$

If we retain the null hypothesis based on the three individual t-tests, then the significance or Type I error is not equal to the α-value, but much higher than α (Lunney, 1969; Siegel, 1990). For the case discussed above, if we retain the null hypothesis based on three individual tests, then the Type I error is $1 - 0.8573 = 0.1426$. That is, when more than two groups are involved, checking the population parameter values simultaneously using t-tests is inappropriate since the Type I error will be estimated incorrectly. For n simultaneous comparisons, the probability of Type I error is $1 - (1 - \alpha)^n$; for five simultaneous comparisons, the Type I error will be approximately 0.22 (Kao and Green, 2008). For this reason, we use ANOVA whenever we need to compare three or more groups for population parameter values simultaneously.

7.3 One-Way ANOVA

One-way ANOVA is appropriate under the following conditions:

1. We would like to study the impact of a single treatment (also known as *factor*) at different levels (thus forming different groups) on a continuous response variable (or outcome variable). For the example discussed in Section 7.1, the variable 'price discount' is the treatment (or factor) and 0%, 10% and 20% price discounts are the different levels (three levels, in this case); different levels of discount are likely to have varying impacts on the quantity of sales of the product. The term 'treatment' is used since one of the initial applications of ANOVA was to find the impact of different fertilizer treatments on agricultural yield as studied by British statistician R A Fisher (1934, 1935).
2. In each group, the population response variable follows a normal distribution and the sample subjects are chosen using random sampling.
3. The population variances for different groups are assumed to be same. That is, variability in the response variable values within different groups is the same.

Although conditions 2 and 3 are necessary for one-way ANOVA, the model is robust and minor violations of the assumptions may not result in an incorrect decision about the null hypothesis. However, we need to check whether conditions 2 and 3 are met to ensure the validity of ANOVA, as a best practice. Normality assumption can be checked either using P–P plot (probability–probability plot) or using goodness of fit tests such as chi-square goodness of fit test. The equality of variance can be checked through the hypothesis test for equal variance discussed in Chapter 6.

7.3.1 Setting up an ANOVA

Assume that we would like to study the impact of a factor (such as discount) with k levels on a continuous variable (such as sales quantity). Then, the null and alternative hypotheses for one-way ANOVA are given by

$$H_0: \mu_1 = \mu_2 = \mu_3 = \ldots = \mu_k$$

H_A: Not all μ values are equal

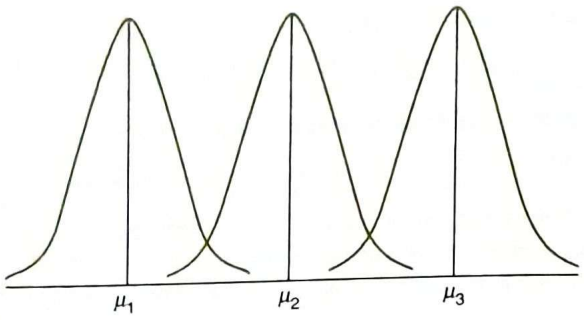

Figure 7.1 | Comparing three means (μ_1, μ_2, and μ_3).

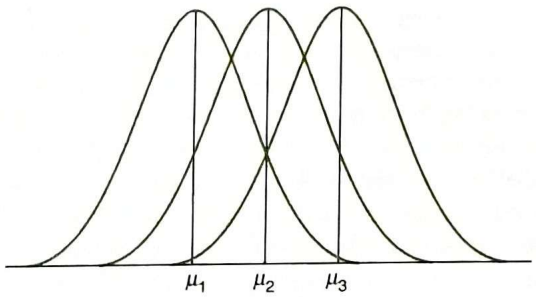

Figure 7.2 | Comparing three means (μ_1, μ_2, and μ_3).

Note that the alternative hypothesis, 'not all μ values are equal', implies that some of them could be equal. The null hypothesis is equivalent to stating that the factor effects $\tau_1, \tau_2, \ldots, \tau_k$ defined in Equation (7.2) are zero. The hypothesis test can be visualised as shown in Figure 7.1. Different values of mean (μ_1, μ_2, and μ_3) imply statistically significant impact of factor levels on the response variable. We expect the group means (μ_1, μ_2, and μ_3) to be closer to one another if the factor levels do not have any impact (Figure 7.2).

If the mean values of different groups are not equal, then the variation of cases within the group will be much smaller compared to variations between groups. Assume that we are interested in analyzing single factor effect with k levels, thus we will have k groups.

Let

k = Number of groups (or samples drawn from different populations)

n_i = Number of observations in group i ($i = 1, 2, \ldots, k$)

n = Total number of observations $\left(= \sum_{i=1}^{k} n_i\right)$

Y_{ij} = Observation j in group i

$$\mu_i = \text{Mean of group } i = \frac{1}{n_i} \sum_{j=1}^{n_i} Y_{ij}$$

$$\mu = \text{Overall mean} = \frac{1}{n} \sum_{i=1}^{k} \sum_{j=1}^{n_i} Y_{ij}$$

To arrive at the statistic, we calculate the following measures, which are variations within a group and between groups:

1. **Sum of Squares of Total Variation (SST):** Total variation is the sum of squared variation of all values of response variable (Y_{ij}) from the overall mean (μ) and is given by

$$\text{SST} = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (Y_{ij} - \mu)^2 \tag{7.4}$$

The degrees of freedom for SST is $(n-1)$ since only the value of μ is estimated from n observations, and thus only one degree of freedom is lost. Mean Square Total (MST) variation is given by

$$\text{MST} = \frac{\text{SST}}{n-1} \tag{7.5}$$

2. **Sum of Squares of between (SSB) Group Variation:** Sum of squares of between variation is the sum of squared variation between the group mean (μ_i) and the overall mean (μ) of the data, and is given by

$$\text{SSB} = \sum_{i=1}^{k} n_i \times (\mu_i - \mu)^2 \tag{7.6}$$

The degrees of freedom for SSB is $(k-1)$. Since the overall mean μ is estimated from the data, one degree of freedom is lost. Mean Square Between variation (MSB) is given by

$$\text{MSB} = \frac{\text{SSB}}{k-1} \tag{7.7}$$

3. **Sum of Squares of Within (SSW) Group Variation:** Sum of squares of within the group variation is the sum of squared variation of all observations (Y_{ij}) from that group mean (μ_i) and is given by

$$\text{SSW} = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (Y_{ij} - \mu_i)^2 \tag{7.8}$$

The degrees of freedom for SSW is $(n-k)$. Here k degrees of freedom are lost since we estimate k group means (μ_i). The mean square of variation within the group is

$$\text{MSW} = \frac{\text{SSW}}{n-k} \tag{7.9}$$

We can prove algebraically

$$\sum_{i=1}^{k} \sum_{j=1}^{n_i} (Y_{ij} - \mu)^2 = \sum_{i=1}^{k} n_i \times (\mu_i - \mu)^2 + \sum_{i=1}^{k} \sum_{j=1}^{n_i} (Y_{ij} - \mu_i)^2 \tag{7.10}$$

That is

$$\text{SST} = \text{SSB} + \text{SSW} \tag{7.11}$$

Visualization of variability between groups (SSB) and variability within groups (SSW) is shown in Figure 7.3.

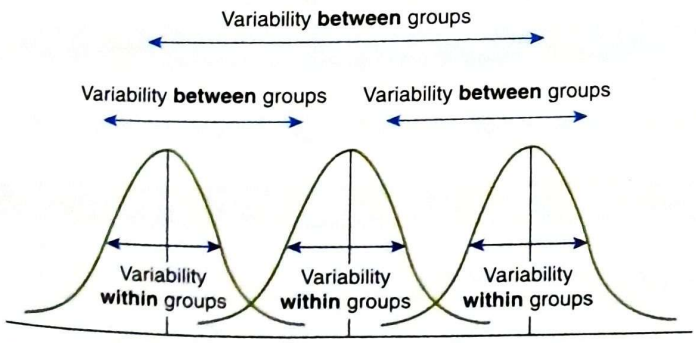

Figure 7.3 | Visualization of variability within groups and between groups.

7.3.2 Cochran's Theorem

According to Cochran's Theorem (Kutner *et al.*, 2013, page 70):

> 'If $Y_1, Y_2, \ldots, Y_n$ are drawn from a normal distribution with mean μ and standard deviation σ and sum of squares of total variation [Eq. (7.11)] is decomposed into k sum of squares (SS_r) with degrees of freedom df_r, then the ratio (SSr/σ^2) are independent χ^2 variables with df_r degrees of freedom if $\sum_{r=1}^{k} df_r = n-1$.'

Note that in Equation (7.11), the SST is decomposed into two sums of squares (SSB and SSW) and thus, SSB/σ^2 and SSW/σ^2 are chi-square variables.

7.3.3 The *F*-test

If the null hypothesis is true, then there will be no difference in the mean values, which will result in no difference between MSB and MSW. Alternatively, if the means are different, then MSB will be larger than MSW. That is, the ratio MSB/MSW will be close to 1 if there is no difference between the mean values and will be larger than 1 if the means are different. Following Cochran's Theorem (Kirk, 1995) MSB/MSW is a ratio of two chi-square variate which is an *F*-distribution. Thus, the statistic for testing the null hypothesis is

$$F = \frac{SSB/(k-1)}{SSW/(n-k)} = \frac{MSB}{MSW} \tag{7.12}$$

Note that the test statistic is a one-tailed test (right-tailed) since we are interested in finding whether the variation between the groups is greater than variation within the groups. Although we are checking whether the means are equal in the null hypothesis, the actual testing is carried out by checking whether the variation between the groups is higher than within the groups; thus it is a one-tailed (right-tailed) test. It is important to note that rejecting the null hypothesis will not tell us exactly which means differ from each other, it will only indicate that there is a difference in at least one of the group means. We may have to conduct two-sample *t*-tests to find out which mean values are different. Alternatively, one may use Tukey's HSD (Honestly Significant Difference) test for multiple comparison of population means.

Example 7.1

Ms Rachael Khanna, the brand manager of ENZO detergent powder at the 'one stop' retail, was interested in understanding whether price discounts have any impact on the sales quantity of ENZO. Towards that end, discounts of 0% (no discount), 10% and 20% were given on randomly selected days. The quantity (in kilograms) of ENZO sold in a day under different discount levels is shown in Table 7.2. Conduct a one-way ANOVA to check whether discount had any significant impact on the average sales quantity at $\alpha = 0.05$.

Table 7.2 | Sales of ENZO at different price discounts

No discount (0% discount)									
39	32	25	25	37	28	26	26	40	29
37	34	28	36	38	38	34	31	39	36
34	25	33	26	33	26	26	27	32	40
10% Discount									
34	41	45	39	38	33	35	41	47	34
47	44	46	38	42	33	37	45	38	44
38	35	34	34	37	39	34	34	36	41

(Continued)

Table 7.2 | *(Continued)*

20% Discount									
42	43	44	46	41	52	43	42	50	41
41	47	55	55	47	48	41	42	45	48
40	50	52	43	47	55	49	46	55	42

Solution

In this case, the number of groups $k = 3$; $n_1 = n_2 = n_3 = 30$; $\mu_1 = 32$, $\mu_2 = 38.77$, $\mu_3 = 46.4$; and $\mu = 39.05$.

The sum of squares of between groups variation (SSB) is given by

$$SSB = \sum_{i=1}^{k} n_i \times (\mu_i - \mu)^2 = 30 \times [(32 - 39.05)^2 + (38.77 - 39.05)^2 + (46.4 - 39.05)^2]$$

$$= 3114.156$$

So

$$MSB = \frac{SSB}{k-1} = \frac{3114.156}{2} = 1557.078$$

The sum of squares of within the group variation is given by

$$SSW = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(Y_{ij} - \mu_i)^2 = \sum_{j=1}^{30}(Y_{1j} - 32)^2 + \sum_{j=1}^{30}(Y_{2j} - 38.77)^2 + \sum_{j=1}^{30}(Y_{3j} - 46.4)^2$$

$$= 2056.567$$

$$MSW = \frac{SSW}{n-k} = \frac{2056.567}{90-3} = 23.63$$

The F-statistic value is

$$F_{2,87} = \frac{MSB}{MSW} = \frac{1557.078}{23.6387} = 65.86$$

The critical F-value with degrees of freedom (2, 87) for $\alpha = 0.05$ is 3.101 [Excel function FINV (0.05, 2, 87) or F.INV.RT(0.05, 2, 87)]. The p-value for $F_{2,87} = 65.86$ is 3.82×10^{-18} [using Excel function FDIST (65.86, 2, 87) or F.DIST.RT(65.86, 2, 87)]. Since the calculated F-statistic is much higher than the critical F-value, we reject the null hypothesis and conclude that the mean sales quantity values under different discounts are different. The Excel output of ANOVA is shown in Table 7.3.

Table 7.3 | **One-way ANOVA Excel output for Example 7.1**

ANOVA: Single factor						
SUMMARY						
Groups	*Count*	*Sum*	*Average*	*Variance*		
No discount	30	960	32	27.17241		
10% Discount	30	1,163	38.76667	20.46092		
20% Discount	30	1,392	46.4	23.28276		
ANOVA						
Source of variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Between groups	3,114.15556	2	1,557.078	65.86986	3.82E-18	3.101296
Within groups	2,056.56667	87	23.6387			
Total	5,170.72222	89				

Example 7.1 is an experimental study in which the marketer was trying to study the impact of discounts on sales. Example 7.2 is an observational study in which we understand the impact of different sectors on stock returns.

Example 7.2

Share Raja Khan (SRK) is a top stockbroker and believes the average annual stock return depends on the industrial sector. To validate his belief, SRK collected annual return of stocks from three different industrial sectors – consumer goods, services and industrial goods. The annual return of shares in 2015–2016 for different sectors is shown in Table 7.4.

Table 7.4 | Annual return of stocks under different industrial sector

Annual return on 30 consumer goods stocks									
6.32%	14.73%	11.95%	12.36%	10.28%	3.81%	10.15%	11.06%	6.29%	5.15%
8.44%	14.28%	8.89%	5.98%	6.96%	11.62%	5.22%	5.34%	5.93%	7.10%
10.91%	8.20%	10.19%	9.04%	8.61%	9.39%	2.63%	2.77%	4.76%	9.60%
Annual return on 30 services stocks									
13.70%	3.58%	1.36%	17.41%	10.01%	10.88%	15.63%	−0.04%	10.32%	7.40%
11.48%	9.71%	11.19%	8.21%	1.64%	1.45%	10.12%	13.85%	−10.27%	5.26%
12.05%	4.47%	8.71%	5.59%	10.02%	7.65%	10.03%	7.87%	6.59%	13.60%
Annual return on 30 industrial goods stocks									
6.74%	7.11%	5.69%	2.48%	5.42%	8.00%	2.55%	8.34%	4.99%	3.39%
8.73%	13.85%	5.29%	9.06%	2.84%	5.82%	7.66%	4.12%	9.10%	8.76%
10.77%	1.48%	4.71%	10.66%	0.44%	2.94%	6.55%	2.84%	3.90%	7.28%

Solution

In this case, the number of cases $k = 3$; $n_1 = n_2 = n_3 = 30$; $\mu_1 = 0.082$, $\mu_2 = 0.079$, $\mu_3 = 0.0605$; and $\mu = 0.0743$.

The sum of squares of between groups (SSB) variation is given by

$$SSB = \sum_{i=1}^{k} n_i \times (\mu_i - \mu)^2$$

$$= 30 \times [(0.082 - 0.0743)^2 + (0.079 - 0.0743)^2 + (0.0605 - 0.0743)^2] = 0.0087$$

Therefore,

$$MSB = \frac{SSB}{k-1} = \frac{0.0087}{2} = 0.0043$$

The sum of squares of within the group variation is given by

$$SSW = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (Y_{ij} - \mu_i)^2$$

$$= \sum_{j=1}^{30} (Y_{1j} - 0.082)^2 + \sum_{j=1}^{30} (Y_{2j} - 0.079)^2 + \sum_{j=1}^{30} (Y_{3j} - 0.0605)^2 = 0.1463$$

So $$MSW = \frac{SSW}{n-k} = \frac{0.1463}{90-3} = 0.0016$$

The F-statistic value is

$$F_{2,87} = \frac{MSB}{MSW} = \frac{0.0043}{0.0016} = 2.592$$

The critical F-value with degrees of freedom (2, 87) for $\alpha = 0.05$ is 3.101 [Excel function FINV(0.05, 2, 87) or F.INV.RT(0.05, 2, 87)]. The p-value for $F_{2,87} = 2.592$ is 0.0805 [using Excel function FDIST(2.592, 2, 87) or F.DIST.RT(2.592,2,87)]. Since the calculated F-statistic is less than the critical F-value, we retain the null hypothesis and conclude that the average annual returns under industrial sectors consumer goods, services and industrial goods are not different (Figure 7.4 shows the F-critical value and F statistic value for a F distribution with degrees of freedom 2 and 87 for numerator and denominator respectively). The Excel output of ANOVA is shown in Table 7.5.

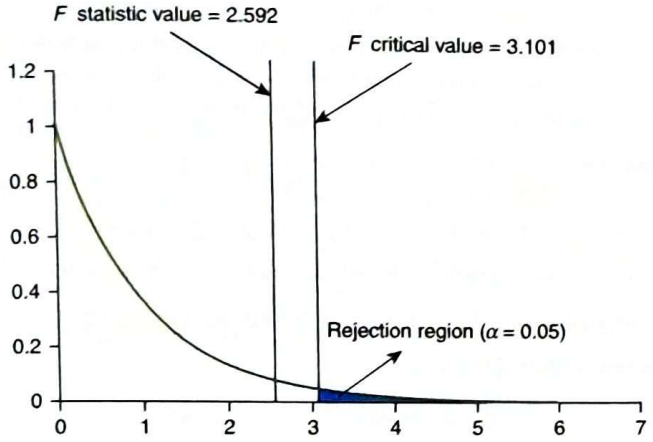

Figure 7.4 | F-distribution with critical value for the example 7.2.

Table 7.5 | Microsoft Excel ANOVA Table for Example 7.2

ANOVA: Single factor						
SUMMARY						
Groups	*Count*	*Sum*	*Average*	*Variance*		
Consumer goods	30	2.4796	0.082653	0.00101		
Services	30	2.3947	0.079823	0.003073		
Industrial goods	30	1.8151	0.060503	0.000963		
ANOVA						
Source of variation	*SS*	*df*	*MS*	*F*	*P-value*	*F critical*
Between Groups	0.008722	2	0.004361	2.59294	0.080572	3.101296
Within groups	0.146317	87	0.001682			
Total	0.155039	89				

7.4 | Two-Way ANOVA

The values of the response variable may be influenced by several factors. For example, in addition to price discounts, location of stores may also play an important role in the sales quantity. The discount may not have much impact if the store is located near an affluent community, compared to stores located near non-affluent communities. We would like to understand the impact of both factors (price discount and location) simultaneously on sales by trying to answer the following questions:

1. Are there differences in the average sales quantity with different levels of price discounts?
2. Are there differences in the average sales quantity with respect to different locations?
3. Are there interactions between price discounts and location with respect to average sales quantity?

The two-way ANOVA model can be expressed as

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \alpha_i\beta_j + \varepsilon_{ijk} \tag{7.13}$$

where

Y_{ijk} = Value of the k^{th} observation ($k = 1, 2, \ldots, c$) of the response variable at level i($i = 1, 2, \ldots, a$) of factor A and level j ($j = 1, 2, \ldots, b$) of factor B.

μ = Overall mean value of the response variable Y_{ijk}

α_i = Level (effect) of factor A ($i = 1, 2, \ldots, a$)

β_j = Level (effect) of factor B ($j = 1, 2, \ldots, b$)

$\alpha_i\beta_j$ = Interaction of i^{th} level of factor A and j^{th} level of factor B

ε_{ijk} = Error associated with k^{th} of observation at level i of factor A and level j of factor B.

The hypothesis tests associated with two-way ANOVA are as follows:

1. **Test of Factor A Main Effects:**

H_0: $\alpha_i = 0$ for all i ($i = 1, 2, \ldots, a$)

H_A: Not all α_i are equal to zero

2. **Test of Factor B Main Effects:**

H_0: $\beta_j = 0$ for all j ($j = 1, 2, \ldots, b$)

H_A: Not all β_j are equal to zero

3. **Test of Interaction Effects:**

H_0: $\alpha_i\beta_j = 0$ for all i ($i = 1, 2, \ldots, a$) and j ($j = 1, 2, \ldots, b$)

H_A: Not all $\alpha_i\beta_j$ are equal to zero

The sum of squares in the case of two-way ANOVA with equal sample sizes is given by (Fisher, 1934)

$$\text{SST} = \text{SSA} + \text{SSB} + \text{SSAB} + \text{SSW} \tag{7.14}$$

Various components in Equation (7.14) are provided as follows:

1. **Sum of squared of total deviation (SST):**

$$\text{SST} = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{c}(Y_{ijk} - \mu)^2 \tag{7.15}$$

where c is the number of observations in each group and μ is the overall mean.

2. **Sum of squares of deviation due to factor A (SSA):**

$$\text{SSA} = b \times c \times \sum_{i=1}^{a}(\mu_i - \mu)^2 \tag{7.16}$$

where μ_i is the mean of all observations in level i of factor A, and c is the number of observations in each group (assumed to be same for all groups).

3. **Sum of squares of deviation due to factor B (SSB):**

$$\text{SSB} = a \times c \times \sum_{j=1}^{b}(\mu_j - \mu)^2 \tag{7.17}$$

Here, μ_j is the mean of all observations in level j of factor B.

4. **Sum of squares of deviation due to interaction of factors A and B (SSAB)**

$$\text{SSAB} = c \times \sum_{i=1}^{a}\sum_{j=1}^{b}(\mu_{ij} - \mu_i - \mu_j + \mu)^2 \tag{7.18}$$

where μ_{ij} is the average of i^{th} level of factor A and j^{th} level of factor B.

5. **Sum of squares of deviation within a group (SSW):**

$$\text{SSW} = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{c}(Y_{ijk} - \mu_{ij})^2 \tag{7.19}$$

Different factors, degrees of freedom and F-statistic for two-way ANOVA with equal number of samples are given in Table 7.6.

Table 7.6 | Sum of squares of deviation for various effects and the corresponding F-statistic in a two-way ANOVA with equal sample size

Sum of Squared Variation	Degrees of Freedom	Mean Squared Variation	F-Statistics
SSA	$a-1$	$\text{MSA} = \text{SSA}/(a-1)$	$F = \text{MSA}/\text{MSW}$
SSB	$b-1$	$\text{MSB} = \text{SSB}/(b-1)$	$F = \text{MSB}/\text{MSW}$
SSAB	$(a-1)(b-1)$	$\text{MSAB} = \text{SSAB}/(a-1)(b-1)$	$F = \text{MSAB}/\text{MSW}$
SSW	$ab(c-1)$	$\text{MSW} = \text{SSW}/ab(c-1)$	

Example 7.3

Table 7.7 shows the sales quantity of detergents at different discount values and different locations collected over 20 days. Conduct a two-way ANOVA at $\alpha = 0.05$ to test the effects of discounts and location on sales.

Table 7.7 | Sales quantity at different locations under different discount rates

Location 1			Location 2		
Discount			Discount		
0%	10%	20%	0%	10%	20%
20	28	32	20	19	20
16	23	29	21	27	31
24	25	28	23	23	35
20	31	27	19	30	25
19	25	30	25	25	31
10	24	26	22	21	31
24	28	37	25	33	31
16	23	33	21	26	23
25	26	27	26	22	22
16	25	31	22	28	32
18	22	37	25	24	22
20	24	28	23	23	29
17	26	25	23	26	25
26	28	23	24	16	34
16	21	26	20	30	30
21	27	33	23	22	25
24	25	28	18	16	39
19	20	30	19	25	32
19	26	30	19	34	29
21	26	26	30	23	22

The two-way ANOVA with replication (since the data in Table 7.7 is repeated for locations) output from Microsoft Excel is shown in Table 7.8.

Table 7.8 | Two-way ANOVA with replication Excel output

ANOVA						
Source of Variation	*SS*	*Df*	*MS*	*F*	*P-value*	*F crit*
Sample (location)	7.008333	1	7.008333	0.443898	0.506593	3.92433
Columns (discount)	1,240.317	2	620.1583	39.27997	1.06E-13	3.075853
Interaction	84.81667	2	42.40833	2.686085	0.07246	3.075853
Within	1,799.85	114	15.78816			
Total	3,131.992	119				

In Table 7.8, the sample stands for the row factor (which in this case is location), column stands for the column factor (discount in this case), and interaction stands for interaction effect (location × discount). The p-value for locations (data in rows) is 0.5065; thus, it is not statistically significant (we retain the null hypothesis that locations have no statistically significant influence on sales), whereas for discount rates (data in column), the p-value is 1.06×10^{-13}, so we reject the null hypothesis (that is, discount rate has influence on sales). The p-value for the interaction effect is 0.0724 and is not significant. That is, only the factor discount is statistically significant at $\alpha = 0.05$.

Summary

1. Analysis of Variance (ANOVA) is a hypothesis testing procedure used for comparing means from several groups simultaneously.
2. In a one-way ANOVA, we test whether the mean values of an outcome variable for different levels of a factor are different. Using multiple two-sample t-tests to simultaneously test group means will result in incorrect estimation of Type I error; ANOVA overcomes this problem.
3. ANOVA plays an important role in multiple linear regression model diagnostics. The overall significance of the model is tested using ANOVA.
4. In a two-way ANOVA, we check the impact of more than one factor simultaneously on several groups.

Multiple Choice Questions (Questions may have more than one correct answer)

1. For a one-way ANOVA, which of the following assumptions should be satisfied?
   - **(a)** The samples are drawn from a normal population.
   - **(b)** The response variable should be a continuous variable.
   - **(c)** The standard deviation of different groups should be equal.
   - **(d)** All of above
2. For an experiment with a single factor with k levels with n observations, the degrees of freedom for sum of squares of variation within the group is
   - **(a)** $n - 1$
   - **(b)** $k - 1$
   - **(c)** $n - k$
   - **(d)** $n - k + 1$
3. For a one-way ANOVA, the hypothesis test is a
   - **(a)** Right-tailed test
   - **(b)** Left-tailed test
   - **(c)** Two-tailed test
   - **(d)** Depends on null hypothesis