Abstract  In the domain of data science and machine learning, statistics plays a huge role. When it comes to gaining insights and building quality features out of the data to train any model, statistical tools and techniques along with the concepts of exploratory data analysis assist in doing the same. A data scientist or data analyst is incomplete without the knowledge of statistics because this is the building block of a machine learning or deep learning model which has learned or needs to learn trends and patterns from the features which were built by analysing the data end-to-end, be it in any tabular form or in picture format or video format. Also, as it covers a lot many concepts under statistics like variables, sampling, correlation, outlier treatment and much more, this chapter solely aims to take the reader to a tour of applicational statistics and how it can be combined with exploratory data analysis to easily work on data science and machine learning. Also, data analysis and machine learning are domains that are experiment heavy and need correct statistical methods for correct inferencing. Hence, for these experiments, the different statistical methods in place are discussed here in detail. There are different languages like Python, MATLAB, R and much more which have libraries for statistical mathematics and make simple API calls to do the required experiments within any dataset.

Keywords  Statistics · EDA · Data science · Machine learning · Data analysis · Python

1  Introduction

Statistics and exploratory data analysis are interlinked where both hand-in-hand assist in seeking and deriving insights by doing an end-to-end analysis on top of the data.

I. Das (✉)
HighRadius Technologies, Bhubaneswar, India
e-mail: indrashisdas98@gmail.com

A. Mishra
Dell Technologies, Bhubaneswar, India
e-mail: anoushka1900@gmail.com

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2021
S. Rautaray et al. (eds.), Trends of Data Science and Applications,
Studies in Computational Intelligence 954,
https://doi.org/10.1007/978-981-33-6815-6_4
But very often, budding data scientists and analysts fall in a dilemma to understand which tool or technique to use. Therefore, all those tools and techniques are discussed here such that it can help terminate all such dilemmas that turn up in the mind. For example, many times just by seeing a variable, a person might not be able to identify the variable type for either being a categorical variable or numerical variable. Hence, to vanish such confusions, several scenarios are explained using examples.

1.1 Statistics and Exploratory Data Analysis

Talking about statistics, it is the methodological science for collecting data, defining data, analysing data and inferring insights from data. Now, when we talk of collecting and defining data, the concepts of population and sample along with the concept of variable types pop up. It is very true that before proceeding with analysing the data, there should always be a sanity check so that the data scientist or analyst gets aware of the discrepancies or anomalies in the data before moving for the analysis phase.

The next phase in the process turns out to be analysing data and inferring insights from it. This particular phase is a handshake between exploratory data analysis and statistical tools and techniques which demands experimenting on the data to understand the hidden trends and patterns. With this phase invoked, concepts like measures of central tendency, outlier treatment, probability distributions and many other such concepts show up. Hence, this is a guide that can help you out with all these concepts under applicational statistics.

1.2 Statistical Tools and Techniques

The concepts are organized in an order which a data scientist or data analyst would follow to do the general exploratory data analysis. Further comes up the concept of variables and types of variables. It is important to analyse the data and define the variables before doing modelling. Once we understand the variables, the phase of hardcore data analysis begins. Hardcore data analysis includes describing the required variables for analysis by its mean, median and mode known as the measures of central tendency. Apart from measures of central tendency, other statistical techniques included are various distributions in statistics, outlier treatment, correlation analysis, variance analysis, chi-square analysis, ANOVA, Z-score, bias and variance, and hypothesis testing. Accordingly, Fig. 1 shows how the concepts of probability and statistics fused together along with exploratory data analysis can help in finding meaningful insights from the data. It is very important to find insights from the data as these act as features for training a machine learning model.
2 Sampling Techniques

2.1 Population Versus Sample

For dealing with any problem, it is very important to understand the data related to it. All data that satisfies the objective being studied is called a population. Let us say we need to understand the job possibility of students in a particular university. This would not just include students on-campus, but also those involved in foreign exchange programmes and distance education as well. All these students constitute the population of the problem statement. This leads to the accumulation of a large volume of data that subjects our system. To overcome this, a portion of the population referred to as a sample is considered. In the above case, analysing only the on-campus students could give us the required insights about such students only.

Sampling is a statistical technique that focuses on selecting a portion of the population to obtain cost model training. Sampling of a population is not as simple as it seems. While sampling, each section of the population should be considered. Let us consider a COVID-19 dataset representing the probabilities of people from different age groups being affected. Let us say, we train our model with a sample that contains people above 50 years of age. This model may not capture the required pattern since it represents only a specific age group of the dataset. Hence, it is very important to choose a sample that depicts the characteristics of the entire data and satisfies the problem statement. The following are some sampling techniques that help us achieve these objectives.
2.2 Sampling Methods

As depicted in Fig. 2, there are two broad categories of sampling which include probability sampling and non-probability sampling. Below are the techniques used under each one of them.

**Simple Random Sampling**

It is the technique of randomly selecting a sample from a population where each member has the same possibility of being preferred.

For example, 5 students have been chosen from a class of 100 for a group discussion. In this case, the population is 100 and the sample size is 5. Each student has an equal possibility of being selected. It has been proven that the storage required is only $O(\sqrt{k})$, where $k$ is the sample size as discussed by Meng [1].

**Systematic Sampling**

This sampling method is applicable when the entire population from which the sample is drawn is available. Under this method, elements of the sample are picked at regular intervals say $k$, where $k$ can be derived by taking the ratio of the population and the proposed size of the sample. For example, if 50 elements are to be selected from a population of 400 elements, then every 8th (400/50) element is selected, i.e. 8, 16, 24, 32, 40… 384, 392, 400.

**Stratified Sampling**

This sampling method is applicable when there is a population of varied nature; i.e. subdivisions are possible. The population is divided into groups, and independently from each group, elements are selected to form the sample. For example, if you have 4 groups with 200, 400, 600 and 800 respective sizes and you decide to choose 1/4 as a sampling fraction, then you would have to select 50, 100, 150 and 200 members from the respective stratum.

![Fig. 2 Types of sampling](image_url)
Cluster Sampling
This method proposes the division of the population into clusters in such a way that every cluster is homogeneous and has a representation of all the groups in a population. Any one of the clusters is selected through random or systematic sampling. For example, in a survey of students from a city, we first select a sample of schools, then we select a sample of classrooms within the selected schools, and finally, we select a sample of students within the selected classes.

Convenience Sampling
In convenience sampling, selection is made based on convenience and proximity and no emphasis is given on the representation of the entire population. For example, Tata Steel is planning to have some vendors who would provide them raw iron. It selects the top 5 cities based on the accessibility.

Voluntary Sampling
This is an important non-probability sampling. In voluntary sampling people willing choose to be a part of the survey. Generally, such people have a strong opinion about the topic and such samples are always subjected to bias.

Purposive Sampling
Purposive sampling is a sampling technique where the researcher trusts his/her judgement while choosing members of the population to participate in the study. The selection of people is based on a particular profile.

Snowball Sampling
This is a non-probability sampling technique based on chain referral. After selecting the initial subject, these subjects enrol future subjects from their contacts. Snowball sampling is of different types like linear, exponential discriminative, exponential and non-discriminative snowball sampling (Table 1).

3 Types of Variables
There are several types of variables, and it is highly appreciated to define the variables and the types so that it becomes easier to analyse the data. Different variables consist of different values and might be of different data types; i.e. it might be an integer-type or float-type variable or may be a date-type variable as well. Hence, Fig. 3 explains the different types of variables that one can find while analysing the data. The variables which are shown in the diagram above are explained below starting with random variables and then going forward towards categorical, numerical, qualitative and quantitative variables. Each variable is cited with an example so that it gets easier for the reader to understand what the variable is all about and what it tries to explain.
| Sampling technique | Description | Application |
|--------------------|-------------|-------------|
| Simple random sampling | Technique of probabilistic sampling where each sample has equal probability of being picked | Technique is mostly used with homogeneous data. Example sample of employee to understand work efficiency |
| Systematic sampling | Entire population is available to pick samples where samples are picked after a certain interval of \( k \) | Example sample of employee to understand work efficiency |
| Stratified sampling | Equally proportionate samples are picked from each class or stratum | Sample of employee collected from each level of employee to understand their satisfaction with the job |
| Cluster sampling | Entire population is divided into homogeneous groups or clusters where simple random or systematic sampling is applied | Sampling of people of different strata of life to understand their lifestyle |
| Convenience sampling | Emphasis on representing the entire population is not taken care of. Rather, based on proximity samples are chosen | Pilots use convenience sampling to understand the weather condition around then to make sure flight is safe or not |
| Voluntary sampling | People are chosen for surveys. These people have domain knowledge. Such techniques are subjected to bias | Sampling of workers by the union leader to take feedback from workers and understand if the work environment is healthy or not |
| Purposive sampling | This is based on a researcher’s judgements or understanding. The samples are selected based on the judgement of the researcher | A researcher sending his work to other scientist for feedback |
| Snowball sampling | This is a chain mechanism of selecting samples. Once one sample is picked, the other ones are picked by the previous samples | A survey conduction where a form is circulated and people who have filled further circulate the form to others; in this way data is collected |

**Fig. 3** Types of variables
3.1 Random Variable

A random variable is one that takes values from a whole set of predefined values. Random variables are important to understand probability distribution in statistics. There are 2 types of random variables—discrete and continuous random variables. If a random variable considers integer values in a particular range such as 1, 2, 3, 4..., then it is a discrete random variable else is a continuous random variable. Random variables are different from the normal variables.

For example, \( x + 5 = 10 \). Here, the value of \( x = 5 \). This is a normal variable. A random variable on the other hand is defined as \( X = \{0, 1, 2, 3\} \) where \( X \) can take values 0, 1, 2 or 3 randomly where each one has the same or different probabilities.

3.2 Categorical Data

Categorical data helps to group the information collected based on various classes in the variable. Categorical variables are of two types:

**Nominal Variable**
A nominal variable is one which does not follow an order while naming a variable. These are sometimes known as labelled or name data. For example, gender has two categories (male and female), season having four categories (summer, winter, monsoon and spring).

**Ordinal Variable**
An ordinal variable has a clear ordering among its values, for example, the size of T-shirts or shoes.

3.3 Numerical Data

Numerical data is represented in numbers such as temperature, weight, height and length. Numerical data can be further divided into two types:

**Discrete Variable**
A discrete variable represents countable items. The values can be grouped into lists that can be finite or infinite. The number of buckets of water present in a tank is finite countable, and the number of buckets of water present in an ocean is infinite countable.

**Continuous Variable**
In continuous variable, the values can take any number from the number line, instead of counting numbers. For example, the heights of people in a class can take values from 0 to 10 where the value can be in decimals as well.
3.4 Qualitative Data

Qualitative data is descriptive in nature. They are also known as categorical data. Types of such data include:

**Nominal Variable**
It is derived from the Latin word “Nomen” (meaning name), name’s variables. Nominal data does not have any numeric characteristics. Such data cannot be manipulated using available mathematical operations.

**Ordinal Variable**
This is a category of qualitative data where variables are represented as ordered categories and there is some relation among those categories. For example, a product having 4 categories is excellent, good, bad and worst.

3.5 Quantitative Data

This data is numerical, and they can be counted, measured and manipulated.

**Interval**
This is a continuous data and is not ordered on a scale (e.g. ratings of 1–10). Each value has equal spacing from the value after and before it. For example, the separation between 19 and 20 km is equivalent to that between 6 and 7 km.

**Ratio**
Such data has an absolute zero value, and the data is continuous. Ratio data and interval data are similar, for example, temperature, which can take to zero degrees (Table 2).

4 Visualizing Data

This section explains the visualizations available for viewing categorical and numerical data. There are separate visualizations for each of these broad categories. The different visualizations are shown below.
### Table 2  Summarizing types of variables

| Type of variable   | Description                                                                 | Application                                                                                     |
|-------------------|-----------------------------------------------------------------------------|-------------------------------------------------------------------------------------------------|
| Random variable   | Takes values from a whole set of defined values. Can be further divided into discrete and continuous variables | Rating of a product in a review analysis dataset                                                |
| Categorical variable | Refers to the stringed categorical variables where data can be grouped based on classes. Can be further divided as nominal and ordinal variables | Nominal: a column in dataset containing the gender of the people Ordinal: a column having T-shirt size |
| Numerical variable | Refers to the variables with data in number format. This can also be divided as discrete and continuous variables | Count of boxes in a dataset comes under discrete data A dataset containing temperature of different places as a column is a continuous column |
| Qualitative variable | These variables are descriptive in nature. Can be divided as nominal and ordinal variables | Label encoding of weather will give nominal qualitative variable If satisfaction of a product (good, average, bad) is label encoded, it comes under ordinal label encoding |
| Quantitative variable | Numerical data that can be counted, measured and manipulated. Can be further divided as interval and ratio | Cost of a place height of a person comes under quantitative variables |

### 4.1 Categorical Data

**Pie chart**

Pie charts are helpful when we represent qualitative data, where an attribute or a feature is not numerical. Each part of pie illustrates a different category, and each slice represents a particular trait. The below example in Fig. 4 shows T-shirt size.

![Pie chart](image)

**Fig. 4** Pie chart
Fig. 5  Bar chart

Bar Plot
This kind of plot helps to visualize qualitative data. Each bar represents the frequency of the individual category, and it helps to compare the frequency count of two categories. As shown in the below diagram, i.e. Fig. 5, this is a typical bar chart that is used for exploratory data analysis on categorical variables.

4.2 Numerical Data

Histograms
These are used to represent quantitative data. The X-axis represents the range of values, i.e. classes, and the classes with maximum frequency have taller bars. Histograms are different from bar plots in the level of measurement of data. Figure 6 justifies the above statement.

Fig. 6  Histogram
**Scatter Plot**
As shown in Fig. 7, a scatter plot helps to show trends or correlation among variables. It also helps us to determine the shape and strength of the correlation. It also helps us to determine outliers.

**Box Plot**
It helps us to understand the distribution of the data as shown in Fig. 8. It helps us to determine the maximum, minimum and interquartile range of the values. It also helps us to find outliers in our data.

**P-P Plots**
Probability–probability plot is a distribution plot that determines how comparable the two data are. It helps to compare a sample distribution with a theoretical distribution. Figure 9 justifies the above statement (Table 3).
Table 3  Summarizing types of visualizations

| Visualization name | Description                                                                 | Application                                                                                           |
|--------------------|-----------------------------------------------------------------------------|--------------------------------------------------------------------------------------------------------|
| **Categorical variable** |                                                                             |                                                                                                        |
| Pie chart           | A circular chart with sections as percentage division of classes            | Analysis of percentage of population from different states in a conference                           |
| Bar chart           | A chart that shows frequency of different classes in a variable             | Count analysis of favourite music in a movie                                                           |
| **Numerical variable** |                                                                             |                                                                                                        |
| Histograms          | Show frequency of continuous variables by binning data on the X-axis        | Histograms can be used to visualize data of students’ performance by providing different buckets       |
| Scatter plot        | Shows data points as simple dots or scatters. Mostly used to see data concentration and outliers | Plotting distance travelled with petrol consumption under their relation                                |
| Box plot            | Used to view outliers in the data. Can also be used to view IQR and other quartiles | Price hike of a place can be determined using a box plot                                               |
| P-P plot            | Shows how comparable two variables are                                      | Comparison of output generated by model with the already existing output column                      |

5 Measures of Central Tendency

There are three measures of central tendency which basically include the mean, median and mode. Apart from this, standard deviation and variance are also included in this section.
5.1 Mean

Mean is the average of all the values in a set. There are different ways of calculating mean, and this includes:

**Arithmetic Mean**

Arithmetic mean is the most used method. If \( X_1, X_2, X_3, \ldots, X_n \) are the values, then mean of these values is denoted by \( \mu \)

\[
\mu = (X_1 + X_2 + X_3 \ldots X_n)/n \tag{1}
\]

**Geometric Mean**

Geometric mean is the nth root of the n positive values \( X_1, X_2, X_3, X_4 \ldots X_n \).

\[
G = \sqrt[n]{(X_1 X_2 X_3 \ldots X_n)} \tag{2}
\]

**Harmonic Mean**

For a set of \( n \) values \( X_1, X_2, X_3, X_4 \ldots X_n \), harmonic mean is the reciprocal of the arithmetic mean of the reciprocals of the values

\[
H = 1/\left(\sum (1/X_i)\right) \tag{3}
\]

5.2 Median

The value in the exact centre, when the data is arranged in ascending order, is known as the median. In case when the number of observations is odd, median is the \((n + 1)/2\)th observation and when there are even numbers of observations median is the mean of \(n/2\)th and \(((n/2) + 1)\)th observation. For example, consider this set \{1, 13, 2, 34, 11, 57, 27, 47, 9\} and the arranged set is \{1, 2, 9, 11, 13, 27, 34, 47, 57\}. The median of this set is 13.

5.3 Mode

As mentioned by Manikandan [2], mode is the value with the maximum frequency. For example, there is a set \{1, 3, 3, 6, 7, 7, 7, 9, 10\}. In this set since 7 has the maximum frequency as it appears 4 times in the set, the mode is 7. However, there are cases when there is no mode such a situation arises when all the values in the set have a single occurrence.
5.4 Variance

It is the expected value of the squared variation of a variable from its mean value. It estimates how far a set of numbers are from their mean value. Let us consider a sample of \( n \) numbers, i.e. \( x_1, x_2, x_3, x_4 \ldots x_n \). Hence, the variance would be

For a population,

\[
S^2 = \frac{\sum (X_i - X_{\text{mean}})^2}{n} \quad (4)
\]

For a sample,

\[
\sigma^2 = \frac{\sum (X_i - X_{\text{mean}})^2}{n - 1} \quad (5)
\]

5.5 Standard Deviation

The square root of variance is called standard deviation. The symbols \( S \) and \( \sigma \) represent standard deviation for a sample and population.

For a population,

\[
S = \sqrt{\frac{\sum (X_i - X_{\text{mean}})^2}{n}} \quad (6)
\]

For a sample,

\[
\sigma = \sqrt{\frac{\sum (X_i - X_{\text{mean}})^2}{n - 1}} \quad (7)
\]

To understand variance and standard deviation in more detail, you can refer to Ahn et al. [3]

6 Distributions in Statistics

A very eminent and influential part of statistics includes distributions which is very intuitive in understanding various variables and fields while building features in the process of machine learning. Basically, distributions speak about how values are distributed for a particular variable, be it more towards larger values, or smaller values, or maybe towards a higher probability for moderate values in case if the distribution is Gaussian or normal.
6.1 Probability Distributions

As discussed by Hernandez [4] in his paper, probability distributions generally depend on two factors of a random variable which includes probability distribution of each random element in the random variable along with particular selection probability of each element. Therefore, probability distributions are functions that basically depict the probability or chance of occurrence of each element in a random variable.

For an example, if we find a probability distribution of scores for a set of students in any examination or test, then the probability distribution formed will always be a Gaussian or normal curve which is the perfect example of randomness, because in a class, there can be higher number of average students scoring average marks because of which probability for average scores would always be high and less number of below average and above average students scoring either too less or too high in any examination leading to lesser probability for such scores.

Further, there are some common probability distributions which are discussed below along with some terminologies associated or linked with the concept of probability distributions.

6.2 PMF Versus PDF

Probability Mass Function
As discussed by Hossain et al. [5] in the paper, to visualize probability distribution for discrete random variables as shown in Fig. 10, probability mass functions are used. Probability mass function helps to define discrete probability distributions. Bernoulli distribution, binomial distribution, Poisson distribution and geometric distribution fall under probability mass function. Also, in case of multivariate probability mass function, a joint probability distribution is used generally.
In the case of probability density function as shown in Fig. 11, the distribution is used to visualize probability distribution for continuous random variables. Normal or Gaussian distribution falls under probability density function. In the probability density function, for any sample in the sample space, the function returns a chance of occurrence or probability of occurrence for that particular sample.

### 6.3 Common Probability Distributions

**Uniform Distribution**

Uniform probability distributions also known as rectangular probability distribution have a constant density over the interval, and hence for all elements in a random variable, there is an equal probability of occurrence for each random element in the random variable. For example as shown in Fig. 12, while we roll a die, there are 6 values that can turn up, i.e. \{1, 2, 3, 4, 5, 6\}. Now, if we do not add any bias while rolling a die, then there is an equal chance of any value turning up; i.e. all values have a probability of 1/6, be it 1 or 2 or any number from the set. Hence, this will form a uniform probability distribution. Formula for uniform distribution is

\[
P(x) = \frac{1}{b - a}, \quad a \leq x \leq b
\]

**(8)**

**Gaussian Distribution**

Dytso et al. [6] described in his paper that Gaussian distribution has been very much accepted by the engineering community because of its flexible parametric form. Gaussian distribution, also known as normal distribution, is the best example of
randomness. Gaussian distributions are symmetric about the mean having a bell-shaped curve, with mean, median and mode on the same point in the $X$-axis. For example, one can see in Fig. 13 that as discussed in the above sections, scores in a test for a class of 80 students are Gaussian in nature. Formula for Gaussian distribution is

$$P(x) = \left(\frac{e^{-(x-\mu)^2/2\sigma^2}}{\sqrt{2\pi\sigma^2}}\right), \quad -\infty \leq x \leq \infty$$  \hspace{1cm} (9)$$

where $\mu$ is the mean, $\sigma$ is the standard deviation and $\pi$ is the universal constant pi.

**Bernoulli Distribution**

One can see in Fig. 14 that the Bernoulli distribution is a discrete distribution, which has binary outcomes that include 0 and 1. Here, 1 means success that occurs with probability $p$, while 0 means failure that occurs with probability $(1-p)$. For example, the distribution of heads and tails in tossing a coin is a Bernoulli distribution. Formula for Bernoulli distribution is

Fig. 12  Uniform distribution

Fig. 13  Gaussian distribution
Bernoulli distribution

\[ P(x) = p^x (1-p)^{1-x}, \quad x = 0 \text{ or } x = 1 \quad (10) \]

where \( x \) is the outcome which is either 0 or 1.

**Binomial Distribution**

As discussed in his paper by Palmisano [7], binomial distribution is the method of modelling likelihood of only two certain outcomes, i.e. by either a success or failure. The conditions for a binomial distribution are that the number of trials is fixed, there is either a success or failure to the trials, there is an equal probability to success in all the trials and the trials are independent as shown in Fig. 15. For example, binomial distribution can be applied to flipping a coin \( n \)-number of times. The trials are fixed, i.e. either a head or tail may come, there are two cases possible for getting a head, either it is a success or failure, also, the trials are independent and the probability of success is constant, i.e. probability of getting a head is 0.5. Formula for binomial distribution is

\[ P(x) = \frac{n!p^x(1-p)^{n-x}}{x!(n-x)!}, \quad 0 \leq x \leq \infty \quad (11) \]

**Fig. 14** Bernoulli distribution

**Fig. 15** Binomial distribution
Table 4  Summarizing the formulas for different distributions

| Distribution type | Distribution formula |
|-------------------|----------------------|
| Uniform           | $P(x) = 1/(b - a)$, $a \leq x \leq b$ |
| Gaussian/normal   | $(x) = (e^{-(x-\mu)^2/2\sigma^2})/\sqrt{2\pi\sigma^2}$, $-\infty \leq x \leq \infty$ |
| Bernoulli         | $P(x) = p^x(1-x)$, $x = 0$ or $x = 1$ |
| Binomial          | $P(x) = (n!p^x(1-p)^{n-x})/x!(n-x)!$, $0 \leq x \leq \infty$ |
| Poisson           | $P(x) = (\lambda^x e^{-\lambda})/x!$, $0 \leq x \leq \infty$ |

where $n$ is number of trials, $x$ is the specified number of successes and $p$ is the probability of success.

**Poisson Distribution**

Unlike binomial distribution, in Poisson distribution, it does not model frequencies of success and failures. Rather, in a fixed unit of time and space, it provides an expected number of events that can occur. Poisson distribution helps to formulate if any event was generated in randomness. The random variable is a positive integer that exists from 0 to infinity due to no cut-off. Also, Poisson distribution approximates binomial distribution when $n$ is large and $p$ is small. Formula for Poisson distribution is

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

where $\lambda$ is the number of events occurring per time per space and $x$ is the general number of events (Table 4; Fig. 16).

Fig. 16  Poisson distribution
6.4 Kurtosis

Introduction to Kurtosis and Interquartile Range

Kurtosis can be defined as the measure that defines how well are the tails of a distribution spread with respect to the normal distribution. It defines if tails of a distribution contain extreme values or not.

Usually, kurtosis is a quality statistical measure for various domains. For example, a higher kurtosis refers to financial risk as it signifies lesser returns; hence, financers look forward to a higher kurtosis for moderate-level risk.

Usually, kurtosis is one of the finest methods or measures to detect outliers in any variable. Many times, extremely smaller or larger values might signify some meaningful or logical information in the dataset for which such values should not be removed as it might show some important pattern. This is the reason why outlier treatment should be done carefully. In the below sections, these concepts will be discussed in detail. Formula for kurtosis is

\[
\text{Kurtosis} = \left( \frac{\sum_{i=1}^{n} (x_i - \mu)^4 / n}{\left( \sum_{i=1}^{n} (x_i - \mu)^2 / n \right)^2} \right)
\]

(13)

where \(x_i\) is an element in the random variable, \(\mu\) is the mean and \(n\) is the number of samples in the random variable.

As it can be seen from Fig. 17, there are certain terminologies which need to be explained and hence are explained in detail below:

25th Quartile—This can be defined as the lower quartile starting which 50% of the values from the random variable lie in the box plot.

![Box plot and normal distribution](image.png)
50th Quartile—This can be defined as the middle quartile which is midway to 50% of the values from the random variable in the box plot. This value also signifies the median of the variable.

75th Quartile—This can be defined as the upper quartile till which 50% of the values in the random variable exist.

Interquartile Range—The range of scores from lower quartile to the upper quartile of the box in box plot is known as the interquartile range.

Formula to calculate interquartile range is

\[
IQR = 75\text{th Quartile} - 25\text{th Quartile}
\]  
(14)

Lower Whisker—This is the lower boundary beyond which any value is considered as an outlier.

Formula to calculate lower whisker is

\[
\text{Lower Whisker} = 25\text{th Quartile} - 1.5 \times IQR
\]  
(15)

Upper Whisker—This is the upper boundary beyond which any value is considered as an outlier.

Formula to calculate upper whisker is

\[
\text{Upper Whisker} = 75\text{th Quartile} + 1.5 \times IQR
\]  
(16)

Types of Kurtosis

Kurtosis is mainly divided into three types as shown in Fig. 18.

Fig. 18 Types of kurtosis
Mesokurtic—Such distributions are in sync with the normal distribution; i.e. they have a moderate height, and the tails are spread lesser. They have a kurtosis of zero or near to zero.

Leptokurtic—These distributions have a higher peak with larger number of outliers because of which the tails are spread wide. Also, these distributions have a positive excess kurtosis.

Platykurtic—These distributions have a flatter peak with lesser number of outliers because of which the tails are also flat. Also, these distributions have a negative excess kurtosis.

6.5 Skewness in Distributions

If one of the tails of a distribution is stretched towards a particular direction be it towards left or towards right, then the distribution is known as a skewed distribution. The prime reason for skewed distributions is due to higher probability of very higher values and lower probability of extreme lower values or higher probability of lower values and lower probability of extremely higher values. The different types of skewed distributions are discussed below.

Left Skewed Distribution
These are the distributions which have its tail towards the negative side as shown in Fig. 19 for which it is also known as negative skewed distribution. Also, higher ranged values being more in the random variable lead to higher probability for such values while lower ranged values having lesser probability due to lesser occurrence in the random variable. Value on the X-axis for which the probability is highest shows the mode of the distribution which is followed by the median and the mean from left to right.

Right Skewed Distribution
These are the distributions which have its tail towards the positive side as shown in Fig. 20 for which it is also known as positive skewed distribution. Also, lower ranged

Fig. 19  Probability distribution, P-P plot and box plot for a left skewed distribution
values being more in the random variable lead to higher probability for such values while higher ranged values having lesser probability due to lesser occurrence in the random variable. Value on the $x$-axis for which the probability is highest shows the mode of the distribution which is followed by the median and the mean from right to left.

**Non-skewed Distribution**

This is a perfect normal distribution or bell-shaped curve that has the highest peak at the centre of the range of values on the $X$-axis. The mean, median and mode lie on the same point in the $X$-axis.

### 6.6 Scaling and Transformations

Generally, while working on datasets it happens that for some continuous random variable, the distribution is skewed, i.e. either it is left skewed or right skewed. Hence in order to make the random variable Gaussian, transformations are required as models understand Gaussianity well. Also, many times, values in random variables are so large that they need to be scaled for which scaling techniques are applied. The requirement for scaling comes as larger values take time to be learned by the model. The purpose of scaling values is to reduce time complexity while training models (Fig. 21).

**Standard Scaling**

This is a technique that is used to scale extremely larger values to smaller ranges by keeping the distribution or Gaussianity intact. The transformed values fall in a positive and negative range, i.e. to $\pm 1$ or $\pm 2$ range. Values lesser than the mean fall in the negative range, while values higher than the mean fall in the positive range. Formula for standard scaling is

$$x' = (x_i - \mu)/\sigma$$

(17)
where \( x_i \) is an element in the random variable, \( \mu \) is the mean, \( \sigma \) is the standard deviation and \( x' \) is the transformed value.

This is also known as the z-score method which is discussed in detail in the upcoming sections (Fig. 22).

**Min–Max Scaling**

In this method, the minimum and maximum values in a random variable are used to scale values from large ranges to a lower range keeping the distribution or Gaussianity same in the random variable. Also, in this method of scaling, the transformed values fall in a range between 0 and 1. Formula for min–max scaling is

\[
x' = \frac{(x_i - x_{\text{min}})}{(x_{\text{max}} - x_{\text{min}})}
\]  

(18)

where \( x_i \) is an element in the random variable, \( x_{\text{min}} \) is the minimum value in the random variable, \( x_{\text{max}} \) is the maximum value in the random variable and \( x' \) is the transformed value (Fig. 23).

**Logarithmic Transformation**

According to simple mathematics, logarithm is a power to which if a base is raised, then we can get the exact number for which we are trying to find a log for. Hence, as the above statement is intuitive enough, logarithmic transformations are used to make left skewed distributions more Gaussian as the larger values are decreased.
Fig. 23  Probability distribution, P-P plot and box plot for min–max scaled data

Fig. 24  Probability distribution, P-P plot and box plot for logarithmic transformation

to lower scale by taking a log. For example, log of 1000 with base 10 is 3 which reduces the corresponding value from 1000 to 3. But logarithmic transformations have a disadvantage which is that it does not work for random variables containing negative values and the number zero as well (Fig. 24).

**Exponential Transformation**

In comparison with logarithmic transformation, exponential transformation can not only be used to Gaussianize left skewed distributions, but be used to Gaussianize right skewed distributions as well. Hence, it is a dual-purpose transformation. Also, another advantage is that it can work for negative values as well. For example, in case of right skewed data, higher degree power can be used to transform values to a larger scale while in case of left skewed data, lower degree power can be used to transform values to lower scale to make the distribution Gaussian (Fig. 25).

**Reciprocal Transformation**

This is another process of transforming left skewed distributions to a normal distribution. Similar to logarithmic transformations, it can be used to scale a left skewed random variable. But in comparison with logarithmic transformation, this can be used to downscale large negative numbers as well. For example, if a random element is valued 1000 in a random variable, then its reciprocal would be 0.001 which is a higher downscaling of value to increase the Gaussianity of the random variable (Fig. 26).
Root-Based Transformation

Similarly, this is also a mechanism to downscale larger values for left skewed random variables. This is exactly similar to logarithmic transformation as in this case as well negative values cannot be downscaled due to generation of complex numbers. Hence, random variables with only positive values can be downscaled. For example, if a random variable has a random element to be valued as 1,048,576, then, on square rooting this number, the transformed value would be 20 which is a very small value in comparison with 1,048,576. Also similarly, if we square root $-1,048,576$, then it will return $20 \sqrt{i}$ which cannot be understood by a machine learning model (Fig. 27).
Fig. 28  Probability distribution, P-P plot and box plot for Box–Cox transformation

**Box–Cox Transformation**

Another effective transformation technique is the Box–Cox transformation. As discussed in his paper by Amir et al. [8], Box–Cox transformation is a transformation that is used to make data work in accordance with assumptions of linear regression and ANOVA. Box–Cox transformation is usually applied to non-normal-dependent random variables. Box–Cox transformation comes from a family of power transformations which is applied to skewed data to convert it to a normal distribution. Also, it cannot be applied to negative random elements in random variables. However, there is another transformation named Yeo–Johnson transformation that can be applied to zero and negative values to random elements in a random variable as well. Formula for Box–Cox transformation is (Fig. 28)

\[ y(\lambda) = \begin{cases} 
  \frac{y^\lambda - 1}{\lambda}, & \text{if } \lambda \neq 0 \\
  \log(y), & \text{if } \lambda = 0 
\end{cases} \]  

(19)

7 **Outlier Treatment**

Outliers play a significant role in screwing up the data used for modelling. Such entities should be removed from the dataset. But outlier treatment should be done extremely carefully because many times, it might show some important patterns or intuitive information, which when excluded or removed might be lossful for the model in the process of training. For example, a very high value in any continuous column might indicate null values, which when removed from the dataset might be a loss in capturing pattern. Hence, this section specifically talks about outliers which can be either disastrous for machine learning models or a blessing in disguise for the same.
7.1 Understanding Outliers

To define outliers, as described by Yang et al. [9] in his paper, it is nothing but deviations in typical data. So a value that cannot exist typically in a variable, but is present somehow is considered as an outlier.

For example, usually test score data of a class full of students cannot be negative, but suddenly you find a negative value in the variable. Now, there are certain cases possible for the same. Firstly, it can be an error from the teacher’s end who entered the data in the portal. In such a case, the data scientist who wants to predict future score range for a student will have to clarify the same from the teacher and take action on it. Now, if the value is not supposed to be negative, then either it can be imputed by some other value, or the record might be dropped or ignored according to the requirement. Secondly, if there was some negative scoring on wrong answers, then in such a case the records should exist without any imputation in the same. Hence, as you saw in the above example, this is how the decision of the data scientist would change depending on the case.

7.2 Detecting Outliers

Below are some techniques for detecting outliers:

**Standard Deviation**

As explained by Yang et al. [9], earlier the method to detect outliers included mean $\pm 3\sigma$, i.e. three times the standard deviation following the standards of normal distribution. But there the scientists decided to keep an unknown control parameter which any user could define according to its requirements. The formula for calculating outlier score using standard deviation is as follows

$$
\text{Threshold}_{\text{Min}} = \mu - a \times \sigma \\
\text{Threshold}_{\text{Max}} = \mu + a \times \sigma
$$

where $\mu$ is the mean, $\sigma$ is the standard deviation and $a$ is the control parameter that can be defined by the user.

Now, smaller the control parameter, more would be the outliers in the data, while larger the control parameter, lesser would be the number of outliers.

**Median and Median Absolute Deviation Method**

Similar to the above process as described by Yang et al. [9], in this method, mean is replaced by median and standard deviation is replaced by median absolute deviation. Median is considered in comparison with mean because median is robust to outliers and it does not get affected. The formula for calculating outlier score using median and median absolute deviation is as follows

$$
\text{Threshold}_{\text{Min}} = \text{median}(x) - a \times \text{MAD} \\
\text{Threshold}_{\text{Max}}
$$
\[ \text{median}(x) + a \times \text{MAD} \]  
(21)

\[ \text{MAD} = b \times \text{median}(|x - \text{median}(x)|) \]  
(22)

where \( \text{median}(x) \) is the median of the random variable, MAD is the median absolute deviation, \( a \) is the control parameter and \( b \) is a constant value 1.4826.

MAD or median absolute deviation contains a term \( \text{median}(|x - \text{median}(x)|) \) which means that for every variable in a random variable, first subtract the median from each value and then find the absolute value of each median deviation. Further, from the set of median deviations, find the median value.

**Box Plots and Interquartile Range**

Box plots are a significant help to visualize and detect outliers in the dataset. It is considered that any value below the lower whisker and any value above the upper whisker is an outlier. This helps understand the range in which the maximum data exists. As discussed in the above sections, IQR is the range of values in the box plot, i.e. the difference between 75th quartile and 25th quartile.

**Scatter Plots**

Scatter plots are also beneficial in detecting a lesser number of outliers from the data. When it can be observed that there are lesser outliers in a particular dataset, then scatter plots can be used to detect and make a call for action on the same. Extreme values can be detected via this method, but close outliers would be hard to detect using the same. This method can just be an indication to outliers if any exist in the dataset.

**Probability Distribution Technique**

Similarly, probability distribution technique is only intuitive enough to know if outliers exist in a variable or not. If there exist any outliers in the dataset, then the distribution is supposed to be skewed, either left or right in accordance with lower value outliers or higher value outliers. But, this method cannot provide exact values to the outliers, which is a demerit in this case (Table 5).

**8 Correlation Analysis**

Two or more variables are correlated means that if one variable is altered, the other would possibly get altered. Correlation is a statistical technique used for analysing the behaviour of two or more variables. Correlation provides us the measure of the direction and degree of sympathetic movement in two or more variables (Table 6).

The coefficient of correlation which ranges between \(-1\) and \(+1\) gives us a measure of the degree of correlation. Also, the below diagram, i.e. Fig. 29, shows correlation matrix using visualization tools like heatmaps.
### Table 5 Summarizing outlier detection techniques

| Outlier detection technique                  | Description                                                                 | Application                                                                 |
|---------------------------------------------|-----------------------------------------------------------------------------|------------------------------------------------------------------------------|
| Standard deviation                          | Detect outliers by defining a range based on the standard deviation, mean and control parameter | Outliers in height for a particulate set students                             |
| Median and median absolute deviation        | Detect outliers by defining a range based on the median absolute deviation, median and control parameter | Outliers in price of a place can be determined using this technique and removed so that prediction of prices of the houses |
| Box plots and interquartile range           | Find outliers by visualizing values beyond the lower and upper whisker in box plots | Price hike of a place can be determined using a box plot and outliers can be determined |
| Scatter plots                                | Find extreme outliers by visualizing using scatter plots. Close outliers cannot be detected using this method | Scatter plot can be used to determine outliers in areas, for example ground living area versus sales, and remove them |
| Probability distribution                    | Find extreme outliers by visualizing if the distribution of a variable is skewed or not. Here also, it is hard to find the exact values of the outliers, but a range can be defined such that any value outside it would be an outlier | Probability distribution can also be used to determine outliers in case of income versus people of a particular place are outliers can be removed for better results |

### Table 6 Bucketing correlation values

| Degree                           | Direction |
|----------------------------------|-----------|
|                                  | Positive  | Negative |
| Perfect                          | +1        | −1        |
| Significant (very high)          | +0.75 to +1 | −0.75 to −1 |
| High                             | +0.5 to +0.75 | −0.5 to −0.75 |
| Low                              | +0.25 to +0.5 | −0.25 to −0.5 |
| Insignificant (very low)         | 0 to +0.25 | 0 to −0.25 |
| Absent                           | 0         | 0         |

### 8.1 Steps for Correlation Analysis

Let us consider two separate data series to determine their correlation.

The various steps involved are as follows

1. Calculate mean of two series, $X$, $Y$
Fig. 29 Example of a heatmap showing correlation between various variables

\[ \mu_x = \left( \frac{1}{n} \sum_{i=1}^{n} X_i \right)/n, \quad \mu_y = \left( \frac{1}{n} \sum_{i=1}^{n} Y_i \right)/n \]  

(23)

2. Calculate variable and standard deviation for each variable
   Sample variance,
   \[ \sigma^2 = \frac{(X_i - \mu_x)^2}{(n - 1)} \quad \sigma^2 = \frac{(Y_i - \mu_y)^2}{(n - 1)} \]  

(24)
   Standard deviation,
   \[ \sigma_{\text{index 1}} = \sqrt{\frac{(X_i - \mu_x)^2}{(n - 1)}} \quad \sigma_{\text{index 2}} = \sqrt{\frac{(Y_i - \mu_y)^2}{(n - 1)}} \]  

(25)

3. Covariance between two series is determined.
   \[ \text{COV}(X, Y) = \sum_{i=1}^{n} (X_i - \mu_x) (Y_i - \mu_y)/(n - 1) \]  

(26)

4. Calculate the correlation coefficient.
   \[ \text{Correlation} = \frac{\text{COV}(X, Y)}{\sigma_{\text{index 1}} \sigma_{\text{index 2}}} \]  

(27)
8.2 Autocorrelation Versus Partial Correlation

Autocorrelation is a representation of the amount of similarity between a given time series and a delayed version of itself. For example, let’s say we have a time series data of daily temperature for a week. We start by shifting the time series one day. Now we find the correlation between the shifted time series and the original one. Now we shift the time series by two days and find the correlation between the original time series and the shifted time series. Similarly, we keep increasing the days and find the correlation. We observe that there is not much difference in the temperature in the consecutive days. The correlation between the original time series and the time series shifted by a day is still very high. As we shift the time series by more number of days, the correlation starts decreasing.

The mean is

\[
\mu_y = \frac{1}{n} \sum_{i=1}^{n} Y_i
\]  \hspace{1cm} (28)

The autocovariance function at lag \( k \) for \( k > 0 \) is

\[
\sigma_k = \frac{1}{n} \sum_{i=1}^{n} (Y_i - \mu_y)(Y_{i-k} - \mu_y)
\]  \hspace{1cm} (29)

Autocorrelation function at lag \( k \) for \( k > 0 \) is

\[
r_k = \frac{\sigma_k}{\sigma_0}
\]  \hspace{1cm} (30)

When there are more than one independent variable controlling the dependent variable in correlation analysis, the correlation between the independent variable and dependent variable is called partial correlation or net correlation, the influence of other correlations is excluded in a partial correlation analysis.

For example, weight loss (dependent variable) is dependent on both dieting (independent variable) and exercise (independent variable). If we consider the effect of dieting on weight loss without considering exercise or vice versa, this measure is known as partial correlation.

\( r_{xy} \) is the correlation between \( x \) and \( y \).
\( r_{xy,z} \) is the partial correlation coefficient for \( x \) and \( y \) controlling \( z \).

\[
r_{xy,z} = \left( r_{yx} - (r_{yz})(r_{xz}) \right) / \left( \sqrt{1 - r_{yz}^2} \sqrt{1 - r_{xz}^2} \right)
\]  \hspace{1cm} (31)
9 Variance and Covariance Analysis

While working with various datasets and doing multivariate analysis on the same, people often get confused between ANOVA, ANCOVA, MANOVA and MANCOVA. But it is necessary to understand how different the above terms are. This section completely focuses to explain and differentiate between these terminologies such that the purpose of each analysis is understood well.

9.1 Analysis of Variance (ANOVA)

Analysis of variance or ANOVA is a method of analysing continuous dependent variables for mean differences under three or more groups. As explained by Akbay et al. [10], there are two types of ANOVA, one-way ANOVA and two-way ANOVA. In case of one-way ANOVA, groups of an independent variable are compared with respect to the mean difference obtained for each group from the continuous dependent variable, while in the case of two-way ANOVA, groups of two independent variables are compared with respect to mean difference obtained for each group from the continuous dependent variable. For example, by level of education if we analyse the mean of test score, then it is a one-way ANOVA while if we introduce country of education to differentiate out level of education and then analyse mean of test scores, it is called a two-way ANOVA.

9.2 Analysis of Covariance (ANCOVA)

As discussed by Rasch et al. [11] in the chapter, ANCOVA combines ANOVA with components of regression analysis. Now, ANCOVA can be understood as multiple linear regression analysis such that there is one categorical independent variable and one continuous independent variable. In this, a dependent variable is compared using both, i.e. the categorical and continuous independent variables. For example, if test score is a continuous dependent variable while level of education is categorical independent variable and hours spent in studying is continuous independent variable, then an analysis of mean test score is considered as an experiment of ANCOVA. The formula for covariance is as follows

\[
\text{Cov}(x, y) = \frac{1}{n} \sum_{i=1}^{n} (x_i - \mu_x)(y_i - \mu_y)
\]

(32)

where \(x\) and \(y\) are two independent variables, \(\mu_x\) is the mean of independent variable \(x\) and \(\mu_y\) is the mean of independent variable \(y\).
9.3 Multiple Analysis of Variance (MANOVA)

The difference between ANOVA and MANOVA stands that in MANOVA there are multiple dependent variables to analyse while in ANOVA there is only one dependent variable to analyse. Similarly, this also has one-way MANOVA and two-way MANOVA. From the example in ANOVA, if we analyse test scores along with annual income and introduce the same to both one-way and two-way ANOVA, then it becomes a MANOVA experiment.

9.4 Multiple Analysis of Covariance (MANCOVA)

Similarly, MANCOVA is nearly similar to ANCOVA. While it is just that for MANCOVA, there are two independent variables to be analysed. With respect to the example in ANCOVA, if we introduce annual income as a dependent continuous variable to be analysed with test scores, then it can be considered as a MANCOVA experiment.

10 Chi-Square Analysis

As per MacFarland et al. [12], the above analysis method helps to tell how well two columns are related or what is the difference in proportions of categories in two or more variables. The formula used is stated below.

\[
\chi^2 = \sum_{i=1}^{n} \frac{(x_o^i - x_e^i)^2}{x_e^i}
\]

(33)

where \(x_o^i\) is the observed value and \(x_e^i\) is the expected value.

Let us assume that we have education level and annual income slab as two categorical columns (Table 7).

| Education level | $10,000–$30,000 | $300,001–$100,000 | >$100,000 | Total |
|-----------------|-----------------|------------------|-----------|-------|
| High school     | 30              | 15               | 5         | 50    |
| B.Tech.         | 20              | 40               | 10        | 70    |
| M.Tech.         | 10              | 40               | 30        | 80    |
| Ph.D.           | 5               | 45               | 50        | 100   |
| Total           | 65              | 140              | 95        | 300   |
After seeing this table, we can calculate the table with expected values as shown below using the formula (Table 8)

Expected Value = (Column Total * Row Total) / Total # of Observations  \hspace{1cm} (34)

Now, calculate the chi-square for each observed and expected value (Table 9). With this, it can be seen that the degree of freedom = (# of column - 1) * (# of rows - 1) = 6 in this case.

For the degree of freedom to be 6 and significance level of 0.05, the critical value is 12.59, and as the total chi-square value is 81.8 which is very higher than the critical value, there exists a high relation between education level and annual income slab which is true in real life as well.

**Table 8**  Preprocessed dataset for chi-square analysis with expected values from observed values

| Education level | $10,000–$30,000 | $300,001–$100,000 | >$100,000 | Total |
|-----------------|-----------------|-----------------|-----------|------|
| High school     | 10.83           | 23.33           | 15.83     | 50   |
| B.Tech.         | 15.17           | 32.67           | 22.17     | 70   |
| M.Tech.         | 17.33           | 37.33           | 25.33     | 80   |
| Ph.D.           | 21.67           | 46.67           | 31.67     | 100  |
| Total           | 65              | 140             | 95        | 300  |

**Table 9**  Intermediate calculations to perform chi-square analysis

| Observed value | Expected value | \((O - E)^2\) | \((O - E)^2/E\) |
|----------------|----------------|--------------|-----------------|
| 30             | 10.83          | 367.49       | 33.93           |
| 20             | 15.17          | 23.33        | 1.53            |
| 10             | 17.33          | 53.72        | 3.1             |
| 5              | 21.67          | 277.88       | 12.82           |
| 15             | 23.33          | 69.38        | 2.97            |
| 40             | 32.67          | 53.72        | 1.64            |
| 40             | 37.33          | 7.12         | 0.19            |
| 45             | 46.67          | 2.79         | 0.06            |
| 5              | 15.83          | 117.29       | 7.41            |
| 10             | 22.17          | 148.11       | 6.68            |
| 30             | 25.33          | 21.80        | 0.86            |
| 50             | 31.67          | 335.98       | 10.61           |
| Total          |                |              | 81.8            |
Table 10  Example showing Z-score analysis

| Score ($x_i$) | Mean | Score mean | $Z$-score ($x_i - \mu$)/$\sigma$ |
|--------------|------|------------|----------------------------------|
| 85           | 80   | +5         | +0.54                            |
| 80           | 80   | 0          | 0                                |
| 68           | 80   | −12        | −1.30                            |
| 75           | 80   | −5         | −0.54                            |
| 92           | 80   | +12        | +1.30                            |

Table 11  Example showing Z-score visualization

|          | 52.4 | 61.6 | 70.8 | 80   | 89.2 | 98.4 | 107.6 |
|----------|------|------|------|------|------|------|-------|
| $\mu - 3\sigma$ |       |      |      |      |      |      |       |
| $\mu - 2\sigma$ |       |      |      |      |      |      |       |
| $\mu - 1\sigma$ |       |      |      |      |      |      |       |
| $\mu$      |       |      |      |      |      |      |       |
| $\mu + 1\sigma$ |      |      |      |      |      |      |       |
| $\mu + 2\sigma$ |      |      |      |      |      |      |       |
| $\mu + 3\sigma$ |      |      |      |      |      |      |       |
| $Z = 0$   | −3   | −2   | −1   | Z = 0 | +1   | +2   | +3    |

11  Z-Score

Z-score is used to calculate how far off a point is from the mean in terms of standard deviation. The formula used is stated below

\[ Z = \frac{(x_i - \mu)}{\sigma} \]  \hspace{1cm} (35)

where $\sigma$ is the standard deviation, $\mu$ is the mean and $x_i$ is the value from a field.

Let us assume that we have the score of students where $\sigma$ is 9.2 here (Tables 10 and 11).

When we add 1 standard deviation to our mean, the $Z$-score is +1 similarly. And, when we add 2 standard deviations to mean it becomes +2 and when we subtract 1 standard deviation it becomes −1 and so on.

This is a real-world application of Z-score explained by Bandyopadhyay [13].

12  Bias Versus Variance

Bias is termed as the error that occurs on approximating a problem which is extensively complicated, by a simpler model. If we consider a linear regression model, it is assumed that there is a linear relation between the input and the output variables which is unlikely to occur in a real scenario.

For example, consider height versus weight plot where weight is the output column. If we try to fit a linear model, it will undoubtedly result in some error in the estimate since a short-heighted obese person would weigh more. The true function is nonlinear so irrespective of the size of the training data, our linear model will not predict the output accurately. This sum of the vertical distances between the true value and the linear function line where the output column is on the y-axis
is called bias. A high bias will cause the model to miss a dominant pattern of the
variable.

When the bias is too high, it is assumed that the model does not favour the
complexity of data.

When a machine learning model is asked to predict unseen data, and if the
predicted values differ from the true values largely, then the model is said to have
variance. In a model with high variance, the predicted values are too far from the
true values. On the other hand in a model with low variance, the predicted values
are away close to the true values which directly affects performance of the model. A
medium variance model is accepted which shows a generic medium fit.

12.1 Bias–Variance Trade-Off

Bias–variance trade-off is the fundamental topic for understanding the model’s
performance. It is not always the case that if a model performs well on the training
set then it would perform well on the unseen/testing data as well. If this happens that
means the model has a low bias (the model has fitted very well to the training data
or it is overfitted) and high variance, the predicted values are far off from the true
values while when we test on the training data, it gives a quality performance, due
to either output leakage or overfitting.

Low Bias and Low Variance
It means that the error with the training data and the error with test data both are less.
This is an ideal case, and we would want your final model to be as close to this as
possible.

Low Bias and High Variance
It means although the model has less error with the training error while it had a
greater error with the test data, i.e. the predictions were deviated with respect to the
true values.

High Bias and Low Variance
It means that the error with the training data is very high and that with the test data
is low. In this case even if the output error is low, it takes a toll on the reliability of
the model since it has high training error which means the model has not captured
the true pattern of the data which makes it unreliable for future predictions.

High Bias and High Variance
It means that there is high error with the training data and high error with the testing
data.

When we try to decrease the bias by adding complexity to the function, it so
happens that the model starts overfitting the training data and which increases the
variance on the testing data.
On the other hand in order to decrease the variance on the testing data, if we try to decrease the complexity of the training data, i.e. if we increase the bias, this may cause the model to underfit. We need to find an optimum ratio between the bias and the variance, in order to obtain a model having optimum bias and variance. This trade-off between the bias and variance in order to obtain the optimum value is called bias–variance trade-off.

One can also refer to Briscoe et al. [14] to get a better understanding of bias–variance trade-off and its complexity.

### 12.2 Overfitting and Underfitting

#### Overfitting
The bias is very less in the model, while variance is very high. It has captured the unnecessary patterns in the training data. Equation: \( \theta_0 + \theta_1x + \theta_2x^2 + \theta_3x^3 + \theta_4x^4 \).

#### Underfitting
The bias is high and fails to capture any pattern in the data having a low variance. Equation: \( \theta_0 + \theta_1x \).

#### Medium Fitting
This is the perfect fit which is needed when neither bias nor variance is very high. Equation: \( \theta_0 + \theta_1x + \theta_2x^2 \) (Fig. 30).

| Fitting     | Description                                      | Application                                                                 |
|-------------|--------------------------------------------------|-----------------------------------------------------------------------------|
| Underfitting| When the model fails to capture necessary data    | When less number of features are provided to train the model example if we do not provide price per-square fit in housing cost prediction model |

(continued)

![Fig. 30](image)  
**Fig. 30**  Example of underfitting, medium fitting and overfitting
| Fitting       | Description                                                                 | Application                                                                 |
|--------------|------------------------------------------------------------------------------|-----------------------------------------------------------------------------|
| Medium fitting | When the model is able to capture all necessary pattern in the training data | When exact features are provided to train example if we provide price per-square, number of bedroom and other important features |
| Overfitting  | When the model capture unnecessary patterns of a training data                | When more features are provided to train then what is necessary              |

## 13 Hypothesis Testing

An assumption about a feature of the population is called hypothesis, and the test means validating something. Hence, hypothesis testing is the process of validating an assumption about a population feature using statistics.

The steps involved in hypothesis testing are referred below:

Step-1: Specify the null hypothesis. Null hypothesis \((H_0)\) is a statement of no relationship, effect or the difference between two or more groups or factors.

Example: \(\mu \geq 0.15\)

Step-2: An alternate hypothesis is specified. Alternate hypothesis is the effect or the difference. The hypothesis that is to be proven is the alternative hypothesis.

Example: \(\mu < 0.15\)

Step-3: The significance level \((\alpha)\) is set. The value of the significance level is generally 0.05; this means that there is a 0.05 chance that even when the null hypothesis is true, the alternative hypothesis will be selected. A smaller value of \(\alpha\) makes it difficult to prove the null hypothesis.

Step-4: The corresponding \(P\)-value is calculated from the test statistics. So now we consider a sample data assuming the mean of the sample to be \(x\) where \(\sigma\) is the size of the standard deviation of the sample, \(\mu\) is the hypothesis mean and \(n\) is the sample size. Now we calculate the \(Z\) statistics. This \(Z\) is a representative of what the sample data is indicating,

\[
Z = \frac{(x - \mu)}{(\sigma / \sqrt{n})}
\] (36)

One thing that we need to keep in mind is that if the sample size is less than 30 then the \(Z\) statistics is called \(T\) statistics.
*P*-value also called probability value is the probability of determining a sample “more extreme” than the one observed in your data, assuming that the null hypothesis is true. The value of *P* can be calculated from the normal distribution table using the *Z* statistics value.

Step-5: Construct the acceptance and rejection region. From Fig. 31, the *Z* value is evident. The area left and right of *Z* statistics is the probability value. While considering the *T* statistics, only one side is considered because in smaller samples the distribution cannot be normal as shown in Fig. 32.

Step-6: Drawing a conclusion. *P*-value less than or equal to significance level (*α*) is used to reject the null hypothesis; i.e. the alternative hypothesis is preferred. If *P*-value is greater than significance level (*α*), we fail to reject the null hypothesis. This outcome is statistically significant.

**Fig. 31** Graph showing *P*-value for *Z* statistics

**Fig. 32** Graph showing *P*-value for *T* statistics
Table 12  Confusion matrix for hypothesis testing

| Decision | In reality | $H_0$ is true | $H_0$ is false |
|----------|------------|---------------|----------------|
| Accept $H_0$ | Ok | Type II error $\beta = \text{Probability of type II error}$ |
| Reject $H_0$ | Type I error $\alpha = \text{Probability of type I error}$ | Ok |

### 13.1 Errors in Hypothesis Testing

Given below is a contingency table (Table 12).

**Type I Error**
When the $P$-value is less than the significance level, the hypothesis is rejected. However, there is a possibility of type I error. There is usually no warning when it occurs and the sample error could have overestimated the chance. Even though we do not know the significance of this error, we know the rate of occurrence of this error, i.e. $\alpha$. If we reduce $\alpha$, the probability of false positive reduces.

**Type II Error**
When we perform hypothesis test and the $P$-value is greater than the significance level, it is not statistically significant. However, there are chances that the effects are present in the population. Hence, it might be a type II error and $\beta$ is called the probability of type II error. This type II error could have occurred because of small effect size, small sample size or high data variability. $1 - \beta$ is called statistical power analysis, which is calculated, and then $\beta$ is derived from it. Low variability and large effect size reduce the type II error which increases the statistical power.

Figure 33 shows the graphical representation of type I and type II errors.

### 14 Conclusion

This chapter is focused for researchers seeking for all concepts related to statistics in an integrated form. As discussed in the above work, one can find concepts beginning with different types of sampling techniques, concepts of different variables and how to visualize them using standard visualization techniques. Further, one can get to know about the measures of central tendency, different distributions, scalings, transformations, outlier detection techniques and much more. Also, the above work discusses different statistical tests like correlation analysis, covariance analysis, $Z$-score testing, chi-square testing, bias–variance concepts and hypothesis testing.
All the concepts and statistical tests are discussed in detail. They follow the standard approaches with available figures and comparative tables for easy understanding and conceptualization.

References

1. Meng, X. 2013. Scalable simple random sampling and stratified sampling. In 30th international conference on machine learning, ICML 2013.
2. Manikandan, S. 2011. Measures of central tendency median and mode. Journal of Pharmacology and Pharmacotherapeutics 2 (3): 214.
3. Ahn, S., and J. Fessler. 2003. Standard errors of mean, variance, and standard deviation estimators.
4. Hernandez, H. 2020. Probability distributions in groups of random elements.
5. Hossain, M.F., and A. Joarder. 2007. Identities based on probability mass functions.
6. Dytso, A., R. Bustin, H.V. Poor, and S. Shamai. 2018. Analytical properties of generalized Gaussian distributions. Journal of Statistical Distributions and Applications 5 (1): 1–40.
7. Palmisano, A. 2018. Poisson and binomial distribution.
8. Amir, M., W.M.A. Ahmad, N.N. Naing, and N. Halim. 2008. An application of box-cox transformation to biostatistics experiment data.
9. Yang, J., and S. Rahardja. 2019. Outlier detection: How to threshold outlier scores?.
10. Akbay, L., Akbay, T., Erol, O., and M. Kiliç. 2019. Inadvertent use of ANOVA in educational research: ANOVA is not a surrogate for MANOVA. Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi.
11. Rasch, D., Verdooren, R., and J. Pilz. 2019. Analysis of covariance (ANCOVA).
12. MacFarland, T., and J.M. Yates. 2016. Chi-square.
13. Bandyopadhyay, A. 2006. Predicting probability of default of Indian corporate bonds: logistic and Z-score model approaches. Journal of Risk Finance 7 (3): 255–272.
14. Briscoe, E., and F. Jacob. 2011. Conceptual complexity and the bias/variance tradeoff. Cognition 118 (1): 2–16. ISSN 0010-0277.