**CSE-Data Science/Data Science, IV semester**

**CD404 INTRODUCTION TO DATA SCIENCE**

**UNIT 3**

**Exploratory Data Analysis**

**NOTES**

**SYLLABUS**

Unit – I: Introduction

Introduction to Data Science – Evolution of Data Science – Data Science Roles – Stages in a Data Science Project – Applications of Data Science in various fields – Data Security Issues.

Unit – II: Data Collection and Data Pre-Processing Data Collection Strategies – Data Pre-Processing Overview – Data Cleaning – Data Integration and Transformation – Data Reduction – Data Discretization.

Unit – III: Exploratory Data Analytics Descriptive Statistics – Mean, Standard Deviation, Skewness and Kurtosis – Box Plots – Pivot Table – Heat Map – Correlation Statistics – ANOVA.

Unit – IV: Model Development Simple and Multiple Regression – Model Evaluation using Visualization – Residual Plot – Distribution Plot – Polynomial Regression and Pipelines – Measures for In-sample Evaluation – Prediction and Decision Making.

Unit – V: Model Evaluation Generalization Error – Out-of-Sample Evaluation Metrics – Cross Validation – Overfitting – Under Fitting and Model Selection – Prediction by using Ridge Regression – Testing Multiple Parameters by using Grid Search.

**3.1 Exploratory Data Analysis**

Exploratory Data Analysis is a data analytics process to understand the data in depth and learn the different data characteristics, often with visual means. This allows us to get a better feel of our data and find useful patterns in it. Exploratory data analysis (EDA) is used by data scientists to analyse and investigate data sets and summarize their main characteristics, often employing data visualization methods. EDA helps to determine how best to manipulate data sources to get the answers we need, making it easier for data scientists to discover patterns, spot anomalies, test a hypothesis, or check assumptions. EDA is primarily used to see what data can reveal beyond the formal modeling or hypothesis testing task and provides a better understanding of data set variables and the relationships between them. Can also help determine if the statistical techniques we are considering for data analysis are appropriate or not? Originally developed by American mathematician John Tukey in the 1970s, EDA techniques continue to be a widely used method in the data discovery process today.

**3.2 Importance**

It is crucial to understand the data in depth before we perform data analysis and run our data through an algorithm. We need to know the patterns in our data and determine which variables are important and which do not play a significant role in the output. Further, some variables may have correlations with other variables. We also need to recognize errors in our data. All of this can be done with Exploratory Data Analysis. It helps us to gather insights and make better sense of the data, and removes irregularities and unnecessary values from data.

* + - Helps us to prepare our dataset for analysis.
    - Allows a machine learning model to predict our dataset better.
    - Gives us more accurate results.
    - It also helps us to choose a better machine learning model.

The main purpose of EDA is to help look at data before making any assumptions. It can help identify obvious errors, as well as better understand patterns within the data, detect outliers or anomalous events, find interesting relations among the variables.

Data scientists can use exploratory analysis to ensure the results they produce are valid and applicable to any desired business outcomes and goals. EDA also helps stakeholders by confirming they are asking the right questions. EDA can help answer questions about standard deviations, categorical variables, and confidence intervals. Once EDA is complete and insights are drawn, its features can then be used for more sophisticated data analysis or modeling, including machine learning.

**3.3 Types of EDA**

There are four primary types of EDA:

* + - Univariate non-graphical.
    - Univariate graphical.
    - Multivariate non-graphical.
    - Multivariate graphical.

**Univariate non-graphical.**

This is simplest form of data analysis, where the data being analyzed consists of just one variable. Since it’s a single variable, it doesn’t deal with causes or relationships. The main purpose of univariate analysis is to describe the data and find patterns that exist within it.

The standard goal of univariate non-graphical EDA is to know the underlying sample distribution or data and make observations about the population. Outlier detection is additionally part of the analysis.

The characteristics of population distribution include:

* Central tendency
* Spread
* Skewness and kurtosis

**Univariate graphical.**

Non-graphical methods don’t provide a full picture of the data. Non-graphical methods are quantitative and objective, they are not able to give the complete picture of the data; therefore, graphical methods are used more as they involve a degree of subjective analysis, also are required.

**Common types of univariate graphics include:**

* Stem-and-leaf plots
* Histograms
* Box plots

Stem-and-leaf plots, which show all data values and the shape of the distribution.

Histograms, a bar plot in which each bar represents the frequency (count) or proportion (count/total count) of cases for a range of values.

Box plots, which graphically depict the five-number summary of minimum, first quartile, median, third quartile, and maximum

**Multivariate non graphical:**

Multivariate data arises from more than one variable. Multivariate non-graphical EDA techniques generally show the relationship between two or more variables of the data through cross-tabulation or statistics.

**Multivariate graphical:**

Multivariate data uses graphics to display relationships between two or more sets of data.

The most used graphic is a grouped bar plot or bar chart with each group representing one level of one of the variables and each bar within a group representing the levels of the other variable.

**Other common types of multivariate graphics include:**

Scatter plot, which is used to plot data points on a horizontal and a vertical axis to show how much one variable is affected by another. Multivariate chart, which is a graphical representation of the relationships between factors and a response. Run chart, which is a line graph of data plotted over time. Bubble chart, which is a data visualization that displays multiple circles (bubbles) in a two-dimensional plot. Heat map, which is a graphical representation of data where values are depicted by colour.

**3.4 TOOLS REQUIRED FOR EXPLORATORY DATA ANALYSIS:**

* R
* Python

**3.5 Steps involved in EDA**

* Data Collection
* Finding all Variables and Understanding Them
* Cleaning the Dataset
* Identify Correlated Variables
* Choosing the Right Statistical Methods
* Visualizing and Analyzing Results

**3.6 Advantages of Using EDA**

* Gain Insights Into Underlying Trends and Patterns
* Improved Understanding of Variables
* Better Pre-process Data to Save Time
* Make Data-driven Decisions

**3.7 Some key takeaways:**

EDA is subjective as it summarizes the features and characteristics of a dataset. So, depending on the project, data scientists can choose from the various plots discussed to explore the data before applying machine learning algorithms.  Since the nature of EDA depends on the data, we can say that it is an approach instead of a defined process. EDA presents hidden insights from data through visualizations such as graphs and plots. Graphical and non-graphical statistical methods can be used to perform EDA.  Univariate analysis is simpler than multivariate analysis. The success of any EDA will depend on the quality and quantity of data, the choice of tools and visualization, and its proper interpretation by a data scientist. EDA is crucial in AI-driven businesses such as retail, e-commerce, banking and finance, agriculture, healthcare, and so on.

**3.8 Statistics and Mathematics Essentials in Data Science:**

Data science revolves around these two fields and draws their methods and models to operate on data.

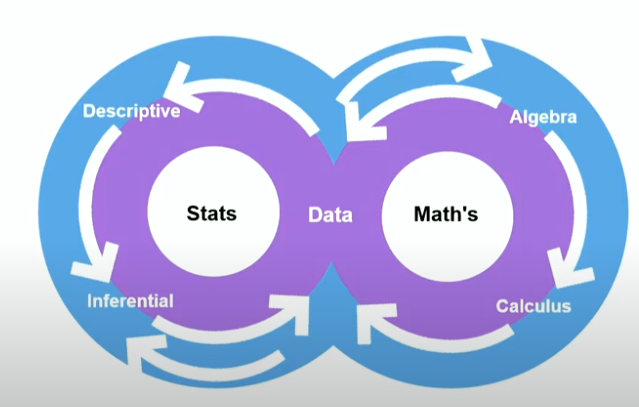


Fig.3.1 Relationship between maths and statistics

**Statistics Methods in Data Science**

* Descriptive statists work on facts about past data.

Inferential statistic works on samples collected from the population to draw conclusions about the population.

**Mathematical methods in Data Science**

Two main methods of mathematics contribute to data science:

* Linear algebra is widely used in image recognition,text analysis and also dimensionality reduction.
* Calculus is used in optimization techniques for machine learning.

**3.9 Statistical Inferences**

Mathematical Discipline

Including methods of

* Data Collection
* Data Organization
* Data Analysis
* Data Interpretation
* Data Presentation

**Example of Statistical Data**

Height,Weight,blood Presure,current,voltage

Number of cars , number of children's,Distance,speed, time, pressure

**Data Collection**

* Surveys
* Census
* Sales
* Bank Transactions
* Online Shopping

**Data Organization**

Data identification, gathering and structuring fro analysis:-

* Tumour registry perform
* Motor vehicle registration
* Voters registration list
* List of tax payers

**Data analysis**

* Number of patients with a particular cancer type
* Number of cars of a particular brand

Number of registered female voters in a constituency

**Data Interpretation**

* Demographic of cancer patients
* Relation between number of tax paid and number of cars owned.
* Female voters turnout in a particular districts.
* Urban and rural divide between tax payers.

**Data presentation**

* Reports
* Charts
* Dashboards

**3.10 Types of Statistics**

**In general statistics analysis falls into two category**

* Descriptive Statistics
* Inferential Statistics

**Descriptive statistic**

Precisely describes collected data

Report data characteristics

* + - Measure of central tendency
    - Data spread
    - Relative standing

**Characteristics of descriptive analysis**

* Describes summary of the data
* Presents data in meaningful way
* Describes already known data
* Data tables, charts and dashboards are used
* Measures of central tendency, spread of data are the tools of descriptive statistics.

**Descriptive data examples**

* Maths score for GCSE students
* Collect data for last 3 years
* One can quickly compare the results
* Number of students with A grade
* Number of students with B grade
* Number of students with C grade
* Number of students Failed

**Measures of central tendency**

* Mean:- In mathematics and statistics, the arithmetic mean or simply the mean or the average, is the sum of collection of numbers divided by the count of number in the collection.
* The collection is often a set of results of an experiments or an observational study, or frequently a set of results from a survey
* Median: - In statistics and probability theory, the median is the value separating the higher half from the lower half of a data sample, a population or a probability distribution. For a data set , it may be thought of as the “MID” value
* Mode:- The mode is the value that appears most often in a set of data values.
* If X is a discrete random variable, the mode is the value x at which the probability mass function takes its maximum value.
* In other words it is the value that is most likely to be sampled

**Data Spread**

* Variance:- In probability theory and statistics, variance is the expectation of the squared deviation of a random variable from its population mean or sample mean. Variance is a measure of dispersion, meaning it is a measure of how far a set of numbers is spread out from their average value
* Standard deviation:- In statistics the standard deviation is the measure of the amount of variation or dispersion of a set of values. A low standard deviation indicates that the values tend to be close to the mean of a set, while a high standard deviation indicates that the values are spread out over a wider range
* Correlation:- In statistics, correlation or dependence is any statistical relationship, weather casual or not, between two random variables or bivariate data. In the broadest sense correlation is any statistical association, through it actually refers to the degree to which a pair of variables are linearly related
* Covariance:- In probability theory and statistics, covariance is a measure of the joint variability of two random variables. If the greater values of one variable mainly correspond with a greater values of the other variable, and the same holds for the lesser values, the covariance is positive
* Range:- In statistics the range of a set of data is the difference between the largest and smallest value. Difference here is specific, the range of a set of data is the result of subtracting the ample maximum and minimum.However in descriptive statistics, this concept of range has a more complex meaning

**Relative Standing**

* Percentile:- In statistics , a k-th percentile, is a score below which a given percentage k of scores in its frequency distribution falls or a score at or below which a given percentage fall.
* For examples, the 50th percentile is the score below which or at or below which 50% of the score in the distribution may found
* Quartile:- A quartile is a statistical term that describes a division of observations into four defined intervals based on the values of the data and how they compare to the entire set of observation
* Inter-Quartile rang:- The interquartile range defines the difference between the third and the first quartile. Quartiles are the partitioned values that divide the whole series into 4 equal parts. So, there are 3 quartiles.
* First Quartile is denoted by Q1 known as the lower quartile, the second Quartile is denoted by Q2 and the third Quartile is denoted by Q3 known as the upper quartile.
* Therefore, the interquartile range is equal to the upper quartile minus lower quartile

**Inferential Statistics**

Draw conclusions about population based on sample data

* Generalize from sample to population
* Hypothesis testing
* Making Predictions
* Analysis of random sample of data taken to describe and make inference about the population.
* Inferential statistics compares, test and predict the data.
* Make conclusions about the population that is beyond the data available.
* Possibility scores are used to represent the data.

**Inferential Analysis tools**

* The basic tools for inferential statistics are:
  + - * Hypothesis test
      * Analysis on variance(ANOVA)

**Inferential data Examples**

* Suppose you want to know the average height of all the men in a city with a population of so many million residence.
* It is not very practical to try and get the height of each man.

**Population and Sample**

* Population: Total number of population in a constituency.

Sample: Number of voters selected for opinion pool.

**Statistical Variable**

A variable is a data item, characteristics, number or quantity that can be measured or counted. It is called a variable because value may be vary between data units in a population and may change in value over time.

**3.11 Descriptive Statistics**

Descriptive statistics refers to a branch of statistics that involves summarizing, organizing, and presenting data meaningfully and concisely.

It focuses on describing and analyzing a dataset's main features and characteristics without making any generalizations or inferences to a larger population.

The primary goal of descriptive statistics is to provide a clear and concise summary of the data, enabling researchers or analysts to gain insights and understand patterns, trends, and distributions within the dataset.

This summary typically includes measures such as central tendency (e.g., mean, median, mode), dispersion (e.g., range, variance, standard deviation), and shape of the distribution (e.g., skewness, kurtosis).

Descriptive statistics also involves a graphical representation of data through charts, graphs, and tables, which can further aid in visualizing and interpreting the information.

Common graphical techniques include histograms, bar charts, pie charts, scatter plots, and box plots.By employing descriptive statistics, researchers can effectively summarize and communicate the key characteristics of a dataset, facilitating a better understanding of the data and providing a foundation for further statistical analysis or decision-making processes

**3.12 Types of Descriptive Statistics**

* Frequency Distribution or Distribution
  + Measures of Central Tendency
  + Measures of Spread (Variability)
  + Measures of Position
* Univariate Descriptive Statistics
* Bivariate Descriptive Statistics
* A frequency distribution is a representation, either in a graphical or tabular format, that displays the number of observations within a given interval.

**Frequency Distribution or Distribution**

The frequency is how often a value occurs in an interval while the distribution is the pattern of frequency of the variable. Datasets consist of a distribution of scores or values. Statisticians use graphs and tables to summarize the frequency of every possible value of a variable, rendered in percentages or numbers.

**Univariate Descriptive Statistics**

Univariate descriptive statistics examine only one variable at a time and do not compare variables. Rather, it allows the researcher to describe individual variables. As a result, this sort of statistic is also known as descriptive statistics.

**Bivariate Descriptive Statistics**

When using bivariate descriptive statistics, two variables are concurrently analyzed (compared) to see whether they are correlated. Generally, by convention, the independent variable is represented by the columns, and the rows represent the dependent variable. There are numerous real-world applications for bivariate data. Bivariate data analysis is a tool in the statistician's toolbox.

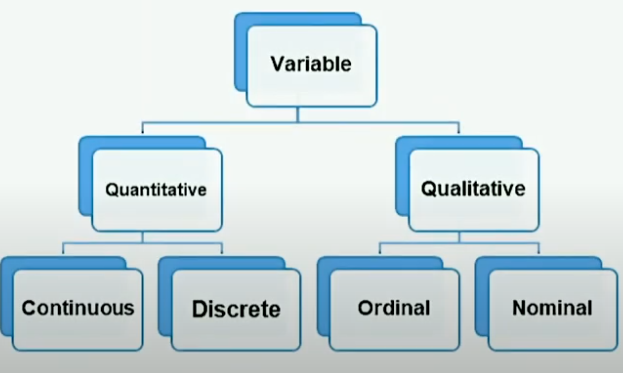


Fig. 3.2 Types of variables

**Quantitative variables**

* A variable is quantitative if its value or categories consists of a numbers and if differences between its categories can be expressed numerically.

Examples

* + - * Height of an individuals
      * Age of an individual
      * Population of a city
      * Number of square feet in a house.
      * Number of students in a class

**Quantitative Variables**

* Discrete Variables
  + A variable whose attributes are separate from one another.
  + It is a finite number of variables.
  + Discrete variable scan be measured as Nominal and Ordinal scale

**Quantitative Variables**

* Continuous Variables
  + A variable can be on any value between two specified values.
  + It is an infinite number of values.
  + Continuous variables is also known as Quantitative Variables.
  + Continuous variables are measured as interval and ratio scale.

**Qualitative Variables**

* Qualitative variable shave discrete categories, usually designed by words, label and non numeric differences between categories.
* Gender
* Breed of dogs
* Level of education
* Eye colour

**Qualitative Variables**

* Ordinal Variable

Ordinal data is a categorical, statistical data type where the variables have natural, ordered categories and the distance between categories are unknown.

**Ordinal variables example**

Ordinal data are categorical, statistical data type where the variables have natural, ordered categories and the distances between the categories are not known.

**Ordinal Variable example**

* Customer satisfaction level
  + - Happy
    - Satisfy
    - Some what satisfied
    - Not satisfied
    - Unhappy

**Ordinal Variable examples**

* Frequency of going to gym
  + - Never
    - Rarely
    - Sometimes
    - Often
    - Regular

**Nominal Variable**

* A nominal scale describes a variable with categories that don't have a natural order or ranking.

We can code nominal variables with numbers if you want, but the order is arbitrary and any calculations such as computing a mean, median or standard deviation would be meaningless.

**Nominal Variables examples**

* Gender
* Race
* Eye colour
* Language
* Material status

**3.13 Skewness**

Skewness can be defined as a statistical measure that describes the lack of symmetry or asymmetry in the probability distribution of a dataset.It quantifies the degree to which the data deviates from a perfectly symmetrical distribution, such as a normal (bell-shaped) distribution. Skewness is a valuable statistical term because it provides insight into the shape and nature of a dataset’s distribution. For example, understanding whether a dataset is positively or negatively skewed can be important in various fields, including finance, economics, and data analysis, as it can impact the interpretation of data and the choice of statistical techniques.

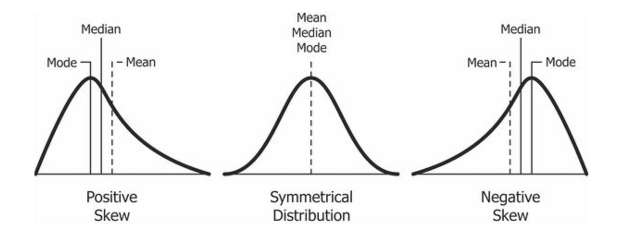


Fig.4 Types of skewness

**3.13.1 Tests of Skewness**

* Visual Inspection: - The simplest way to assess skewness is by creating a histogram or a density plot of the given data. If the plot is skewed to the left, it is negatively skewed, and if the plot is skewed to the right, it is positively skewed. If the plot is roughly symmetric, it has no skewness
* Skewness Coefficient (Pearson’s First Coefficient of Skewness):- This is a numerical measure of skewness, which determines the skewness when mean and mode are not equal. It is calculated as:
* Skewness as per Karl Pearson’s Measure:- In case of mean is greater than mode, the skewness will consist positive value. In case of mean is smaller than mode, the skewness will be a negative value. In case of equality of mean and mode, the skewness will be zero.
  + Skewness = Mean – Mode
* Quartiles are not equal to each other
  + Positive and Negative Skewness
* Positive skewness and negative skewness are two different ways that a dataset’s distribution can deviate from perfect symmetry (a normal distribution). They describe the direction of the skew or asymmetry in the data

**Positive Skewness (Right Skew)**

* In a positively skewed distribution, the tail on the right side (the larger values) is longer than the tail on the left side (the smaller values).
* This means that the majority of data points are concentrated on the left side of the distribution, and there are some extreme values on the right side.
* In the case of a positively skewed dataset,
* Mean > Median > Mode
* Examples of positively skewed data include income distribution (where most people earn a moderate income, but a few earn extremely high incomes), exam scores (where most students score in a certain range, but a few score exceptionally high), and stock market returns (where most days have modest returns, but a few days may have very high returns).

**Negative Skewness (Left Skew)**

* In a negatively skewed distribution, the tail on the left side (the smaller values) is longer than the tail on the right side (the larger values).
* This implies that most of the data points are concentrated on the right side of the distribution, with a few extreme values on the left side.
* In the case of a negatively skewed dataset,
* Mean < Median < Mode
* Examples of negatively skewed data include test scores on an easy test (where most students score well, but a few score very low), the age at retirement (where most people retire at a certain age, but a few retire exceptionally early), and the gestational age at birth (where most babies are born full-term, but a few are born prematurely).

**3.13.2 Measurement of Skewness**

* Karl Pearson’s Measure
* Bowler's Measure
* Kelly’s Measure

**Karl Pearson’s Measure**

* Karl Pearson’s Measure of Skewness uses the mean, median, and standard deviation of the given data set to quantify the asymmetry or lack of symmetry in the distribution.

It is a dimensionless number that provides valuable insights into the shape of a dataset’s distribution.

This measure is valuable in various fields of statistics and data analysis, helping researchers and analysts understand the direction and degree of skewness in their datasets, which can inform subsequent modelling and analytical decisions.

Skewness as per Karl Pearson’s Measure

Skewness = Mean – Mode

Coefficient of Skewness as per Karl Pearson’s Measure

1. With respect to Mean and Median:

2. With respect to Mean and Mode:

**Example of Karl Pearson’s Measure:**

Calculate Pearson’s skewness coefficient for a dataset of exam scores: 85, 88, 92, 94, 96, 98, 100, 100, 100, 100.

Solution:

Step 1: Calculation of Mean

Mean = 95.3

Step 2: Calculation of Median

Since there are 10 data points, the median is the average of the 5th and 6th values when sorted in ascending order:

Median = 97

Step 3: Calculation of standard deviation.

Thus, σ=√26.81

σ = ~5.

Step 4: Calculation of mode

It is clear from the data set that 100 is the most frequently occurring value in the data. Hence, mode of given data is 100.

Step 5: Substitute the values in the formulae

A. With respect to Mean and Median

Sk = -1.02

B. With respect to Mean and Mode

Sk = -0.94

**Bowley’s Measure**

Bowley’s Skewness Coefficient, named after the British economist Arthur Lyon Bowley, is a statistical measure used to assess the skewness or asymmetry in a probability distribution.

Unlike some other skewness measures that rely on deviations from the mean, Bowley’s Skewness Coefficient is based on quartiles.

This coefficient provides a simple and different way to understand the direction and magnitude of skewness in a dataset. Bowley’s Skewness Coefficient is especially useful when dealing with data that may not follow a normal distribution or when a robust measure of skewness is required.

Q1 is the first quartile (25th percentile),

Q2 is the second quartile (50th percentile, or median), and

Q3 is the third quartile (75th percentile).

**Coefficient of Bowley’s Measure**

* If B = 0, the distribution is perfectly symmetric about the mean (no skewness).
* If B < 0, the distribution is negatively skewed (left-skewed), meaning the tail on the left side of the distribution is longer or heavier.

If B > 0, the distribution is positively skewed (right-skewed), indicating that the tail on the right side of the distribution is longer or heavier.

**Example of Bowley’s Measure:**

Calculate Bowley’s Measure of Skewness for the following dataset representing the ages of a group of people in a sample: 20, 24, 28, 32, 35, 40, 42, 45, 50.

Solution:

Step 1: Calculate the median (Q2)

Q2= 35 (the middle value)

Step 2: Calculate the first quartile (Q1)

To find Q1, consider the values to the left of the median: 20, 24, 28, 32

Q1=(24+28)/2=26

Q1 = 26

Step 3: Calculate the third quartile (Q3)

To find Q3, consider the values to the right of the median: 40, 42, 45, 50.

Q3=(42+45)/2=43.5

Q3 = 43.5

**Step 4:** Substitute the above values in the formula

(26+43.5+2\*35)/(43.5-26)= -0.02

B= -0.02

**Kelly’s Measure**

Kelly’s measure of skewness is a way to quantify the degree of skewness in a distribution by comparing the values of certain percentiles (typically the 10th, 50th, and 90th percentiles) or dectiles (10th, 20th, …, 90th percentiles) of the dataset.

Specifically, it involves comparing the difference between the median (50th percentile) and the average of the 10th and 90th percentiles (or deciles) to assess the skewness of the data.

Coefficient of Skewness as per Kelly’s Measure

**Coefficient of Kelly’s Measure**

* If SKL is positive, it indicates positive skewness, meaning the distribution has a longer right tail.
* If SKL is negative, it indicates negative skewness, meaning the distribution has a longer left tail.

If SKL is close to zero, it suggests that the distribution is approximately symmetric.

**Example of Kelly’s Measure:**

Calculate Kelly’s Coefficient of Skewness for the following data:5, 7, 8, 9, 10, 12, 15, 16, 18, 20.

**Solution:**

**Step 1:** Find the 10th Percentile

To find the 10th percentile, we need to rank the data in ascending order and find the value below which 10% of the data falls. In this dataset, the 10th percentile corresponds to the value at position 1 since 10% of 10 data points is 1. So, the 10th percentile is **5**.

**P10 = 5**

Step 2: Find the 50th Percentile (Median)

Since there are 10 data points, the median is the average of the 5th and 6th values when sorted in ascending order

Median =(10+12)/2=11

P50 = 11

Step 3: Find the 90th Percentile

To find the 90th percentile, you need to identify the value below which 90% of the data falls. In this dataset, the 90th percentile corresponds to the value at position 9 since 90% of 10 data points is 9. So, the 90th percentile is 18.

P90 = 18

**Step 4:** Substitute the values in the formula.

 = 0.07

**3.13.3 Interpretation of Skewness**

* Direction of Skewness
  + Negative Skewness (Left Skewed)
  + Positive Skewness (Right Skewed)
  + Zero Skewness (Symmetric)
* Magnitude of Skewness

**Direction of Skewness:**

**Negative Skewness (Left Skewed):**If the skewness is negative, it indicates that the distribution is skewed to the left. In a left-skewed distribution:

* The tail on the left side (the smaller values) is longer and often contains outliers.
* The majority of data points are concentrated on the right side.

The mean is typically less than the median.

**Positive Skewness (Right Skewed):**A positive skewness indicates that the distribution is skewed to the right. In a right-skewed distribution:

* The tail on the right side (the larger values) is longer and may contain outliers.
* Most data points are concentrated on the left side.
* The mean is typically greater than the median.

**Zero Skewness (Symmetric):**

* A skewness value close to zero suggests a symmetric distribution where the data is evenly distributed on both sides of the mean.

This means there is no skewness.

**Magnitude of Skewness:**

* The magnitude of skewness provides information about the degree of skewness.
* If the skewness value is close to 0 (between -0.5 and 0.5), the distribution is approximately symmetric.
* If the skewness value is significantly negative (below -1), it suggests strong left skewness.

If the skewness value is significantly positive (above 1), it suggests strong right skewness.

**3.14 Kurtosis**

* It is also a characteristic of the frequency distribution. It gives an idea about the shape of a frequency distribution.
* Basically, the measure of kurtosis is the extent to which a frequency distribution is peaked in comparison with a normal curve.

It is the degree of peakedness of a distribution.

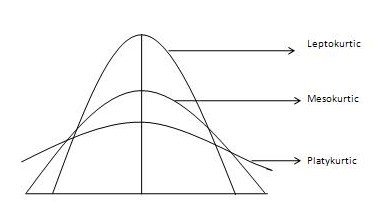


Fig.3.4 Kurtosis

3.14.1 **Types of kurtosis**

* Leptokurtic
* Mesokurtic
* Platykurtic
* **Leptokurtic:**Leptokurtic is a curve having a high peak than the normal distribution. In this curve, there is too much concentration of items near the central value.
* **Mesokurtic:**Mesokurtic is a curve having a normal peak than the normal curve. In this curve, there is equal distribution of items around the central value.
* **Platykurtic:**Platykurtic is a curve having a low peak than the normal curve is called platykurtic. In this curve, there is less concentration of items around the central value.

3.14.2 Difference Between Skewness and Kurtosis

|  |  |
| --- | --- |
| **Skewness** | **Kurtosis** |
| It indicates the shape and size of variation on either side of the central value. | It indicates the frequencies of distribution at the central value. |
| The measure differences of skewness tell us about the magnitude and direction of the asymmetry of a distribution. | It indicates the concentration of items at the central part of a distribution. |
| It indicates how far the distribution differs from the normal distribution. | It studies the divergence of the given distribution from the normal distribution. |
| The measure of skewness studies the extent to which deviation clusters is are above or below the average. | It indicates the concentration of items. |
| In an asymmetrical distribution, the deviation below or above an average is not equal. | No such distribution takes place. |
| **Skewness** | **Kurtosis** |
| It indicates the shape and size of variation on either side of the central value. | It indicates the frequencies of distribution at the central value. |

3.14.3 **How to Calculate Kurtosis?**

* Kurtosis can be calculated by dividing the fourth-order moment by the standard deviation of the population raised to the fourth power.
* Kurtosis is a measure 3 of the fourth moment of a probability distribution of a random variable.

It can be calculated as the ratio of the fourth moment to the square of the variance

**To calculate kurtosis in statistics, you can follow these steps:**

* **Compute the Mean (μ)**: Calculate the arithmetic mean of the dataset.
* **Compute the Variance (σ2)**: Calculate the variance of the dataset, which is the average of the squared differences from the mean.
* **Compute the Standard Deviation (σ)**: Take the square root of the variance to find the standard deviation.
* **Compute the Fourth Moment (μ4)**: Calculate the fourth moment of the dataset, which is the average of the fourth power of the differences from the mean.
* **Compute Kurtosis**: The formula for calculating kurtosis is:  
  Kurtosis = *μ*4/*σ*4​
* Sometimes, we might also see a version of kurtosis that subtracts 3 from this calculation. This is called excess kurtosis, and it subtracts 3 because the kurtosis of a normal distribution is 3.
* So the formula becomes:  
  Excess Kurtosis = (μ4/σ4​)​ − 3

This version is often used because it allows for easier comparison to the normal distribution, where excess kurtosis of 0 indicates normality.

3.14.4 **Measure of Kurtosis**

There are several measures of kurtosis, but the most commonly used one is the Pearson's moment coefficient of kurtosis, also known as simply kurtosis. It is defined as:

Kurtosis  
Where:

* ​ is the fourth central moment (moment about the mean), which is calculated as 𝜇4=𝐸[(𝑋−𝜇)4]μ4​=E[(X−μ)4] where 𝑋 is the random variable and 𝜇 is its mean.
* 𝜎 is the standard deviation.

3.15 **Box Plot**

Box Plot is a graphical method to visualize data distribution for gaining insights and making informed decisions.

Box plot is a type of chart that seprates a group of numerical data through their quartiles.

* Box plot is also known as a whisker plot, box-and-whisker plot, or simply a box-and whisker diagram.
* Box plot is a graphical representation of the distribution of a dataset. It displays key summary statistics such as the median, quartiles and potential outliers in a concise and visual manner.

By using Box plot you can provide a summary of the distribution, identify potential and compare different datasets in a compact and visual manner.

**Elements of Box Plot**

A box plot gives a five-number summary of a set of data which is-

**Minimum** – It is the minimum value in the dataset excluding the outliers.

**First Quartile (Q1)** – 25% of the data lies below the First (lower) Quartile.

**Median (Q2)**– It is the mid-point of the dataset. Half of the values lie below it and half above.

**Third Quartile (Q3)**– 75% of the data lies below the Third (Upper) Quartile.

**Maximum**– It is the maximum value in the dataset excluding the outliers.

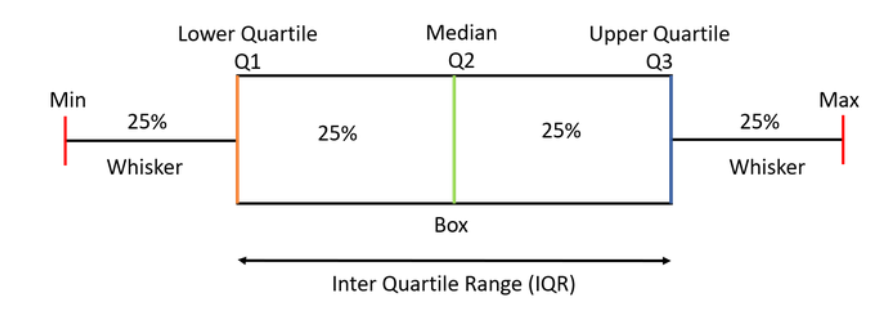


Fig.5 Box plot

The area inside the box (50% of the data) is known as the Inter Quartile Range. The IQR is calculated as –

IQR=Q3-Q1

Outlies are the data points below and above the lower and upper limit. The lower and upper limit is calculated as –

Lower limit=Q1-1.5\*IQR

Upper limit =Q3+1.5\*IQR

3.14.1 **How to create a box plots?**

Let us take a sample data to understand how to create a box plot.

Here are the runs scored by a cricket team in a league of 12 matches –***100, 120, 110, 150, 110, 140, 130, 170, 120, 220, 140, 110.***

To draw a box plot for the given data first we need to arrange the data in ascending order and then find the minimum, first quartile, median, third quartile and the maximum.

Ascending order

100,110,110,110,120,120,130,140,140,150,170,220

Median (Q2)=(120+130)/2=125

To find the First Quartile we take the first six values and find their median.

Q1=(110+110)/2=110

For the Third Quartile, we take the next six and find their median.

Q3=(140+150)/2=145

If the total number of values is odd then we exclude the Median while calculating Q1 and Q3.

Here since there were two central values we included them. Now, we need to calculate the Inter Quartile Range.

IQR=Q3-Q1=145-110=35

We can now calculate the Upper and Lower Limits to find the minimum and maximum values and also the outliers if any.

Lower limit=Q1-1.5\*IQR=110-1.5\*35=57.5

Upper limit=Q3+1.5\*IQR=145+1.5\*35=197.5

 Min=100,Max=220

**3.14.3 Use-Cases of Box Plot**

Box plots provide a visual summary of the data with which we can quickly identify the average value of the data, how dispersed the data is, whether the data is skewed or not (skewness).

The Median gives you the average value of the data.

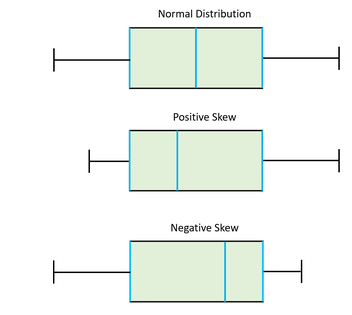


Fig. 6 **Use-Cases of Box Plot**

3.15 **Pivot Plot**

* A pivot table is a table of statistics that summarizes the data of a more extensive table (such as from a database, spreadsheet, or business intelligence program). This summary might include sums, averages, or other statistics, which the pivot table groups together in a meaningful way.

A pivot plot is a visual representation derived from pivot tables, which reorganize and summarize complex datasets to reveal insights. By aggregating data according to specified dimensions, pivot tables condense information, facilitating analysis from multiple perspectives. Pivot plots then translate these summaries into graphical forms like bar charts, line graphs, or heatmaps, making trends, patterns, and outliers easily identifiable. They are instrumental in data analysis, enabling quick comparison across different categories and time periods. This visual aid enhances decision-making processes by presenting intricate data relationships clearly and concisely, allowing stakeholders to grasp and act upon key insights efficiently.

A pivot plot is a powerful visual representation that stems from the use of pivot tables, a critical tool in data analysis for summarizing and reorganizing large datasets. Pivot tables facilitate the transformation of complex and voluminous data into an easily understandable format by aggregating data based on specific dimensions and metrics. These tables allow for the dynamic reorganization of data, enabling users to view it from various perspectives and thus extract valuable insights. When this summarized data is visualized using pivot plots, the benefits of pivot tables are significantly enhanced, providing a clear and immediate understanding of underlying patterns, trends, and anomalies within the data. The core concept behind pivot tables is their ability to aggregate data, such as summing up sales figures across different regions, averaging performance metrics over time, or counting the frequency of certain events within a dataset. This aggregation is performed by specifying the rows, columns, and values of interest, allowing for a multidimensional analysis. For example, a business analyst might use a pivot table to compare monthly sales across different product categories and regions, revealing not only the total sales figures but also identifying which products perform best in which regions and during which periods.

The visual nature of pivot plots makes them particularly effective in highlighting trends and patterns that might not be immediately evident in a tabular format. For example, a line graph generated from a pivot table can quickly reveal seasonal trends in sales data, such as peaks during holiday seasons or dips during off-peak periods. Similarly, a heatmap can show at a glance which product categories are underperforming across various regions, enabling more targeted decision-making.

Pivot plots are not only beneficial for analysts but also for communicating insights to stakeholders who may not have the same level of data literacy. Visual representations of data are generally easier to understand and interpret than raw tables or textual summaries. By providing a clear and concise visual summary, pivot plots help ensure that key insights are conveyed effectively, facilitating better-informed decision-making processes.

Furthermore, the interactive capabilities of many modern data visualization tools enhance the utility of pivot plots. Users can often interact with the plots, drilling down into specific data points, filtering data dynamically, and adjusting the dimensions and metrics to explore different aspects of the data. This interactivity makes pivot plots a flexible and versatile tool in exploratory data analysis, allowing users to uncover deeper insights through iterative exploration.

**3.16 Heat Map**

* Heatmap is defined as a graphical representation of data using colors to visualize the value of the matrix.
* In this, to represent more common values or higher activities brighter colors basically reddish colors are used and to represent less common or activity values, darker colors are preferred.
* Heatmap is also defined by the name of the shading matrix.

Heatmaps in Seaborn can be plotted by using the seaborn.heatmap() function.

**Important Parameters:**

* data: 2D dataset that can be represented into an nd-array.
* vmin, vmax: Values to anchor the colormap, otherwise they are inferred from the data and other keyword arguments.
* cmap: The mapping from data values to color space.
* annot: If True, write the data value in each cell.

A heatmap is a powerful data visualization tool that represents data values in a matrix format using color gradients. It is particularly useful for displaying large datasets where individual data points might be overwhelming or difficult to interpret. By using color to represent different values, heatmaps make it easy to spot patterns, trends, and anomalies at a glance, providing a visual summary that facilitates quick and effective analysis. Heatmaps are commonly used in various fields, including finance, biology, marketing, and web analytics, due to their ability to convey complex information clearly. In finance, for example, heatmaps can show the performance of different stocks over time, with colors indicating gains or losses. In biology, they are often used to represent gene expression data, where colors indicate the level of expression of different genes under various conditions. In marketing, heatmaps can illustrate customer behavior on websites, highlighting areas where users click most frequently, thereby revealing popular sections or potential issues with site navigation.

The construction of a heatmap involves plotting data points on a grid where each cell represents a specific data point. The color of each cell corresponds to the value of that data point, with a predefined color gradient used to indicate the range of values. For instance, in a heatmap showing temperature data, a gradient from blue to red might be used, with blue representing lower temperatures and red representing higher temperatures. This visual encoding allows viewers to immediately discern areas of high and low values without needing to interpret numerical data directly.

One of the key advantages of heatmaps is their ability to highlight correlations and patterns within the data. For example, in a heatmap displaying sales data across different regions and months, it might become evident that certain regions consistently outperform others or that there are seasonal spikes in sales. Such insights can be critical for strategic planning and decision-making. Heatmaps can also reveal outliers or anomalies, such as unusually high or low values that might indicate potential issues or opportunities that warrant further investigation. In addition to their analytical benefits, heatmaps are also valuable for communicating findings to a broader audience. Their intuitive visual format makes them accessible to individuals who may not have a strong background in data analysis, enabling them to grasp key insights quickly. This makes heatmaps an effective tool for presentations and reports, where conveying complex information succinctly is often essential.

Heatmaps also have interactive capabilities in modern data visualization platforms. Tools like Tableau and Power BI allow users to create interactive heatmaps where they can hover over cells to see detailed information, zoom in on specific areas, and apply filters to focus on subsets of the data. This interactivity enhances the exploratory analysis process, allowing users to dive deeper into the data and uncover more nuanced insights.

Heatmaps are an invaluable tool in the realm of data visualization, offering a visually intuitive way to represent complex data sets. By leveraging color to convey information, they make it easier to identify patterns, correlations, and anomalies, thereby facilitating more effective data analysis and communication. Whether used in finance, biology, marketing, or web analytics, heatmaps provide a clear and concise visual summary that supports informed decision-making and enhances the understanding of complex data.

**3.17 ANOVA**

ANOVA, or Analysis of Variance, is a statistical method for comparing means among three or more groups, crucial in understanding group differences and relationships in diverse fields.

ANOVA, or Analysis of Variance is a parametric statistical technique that helps in finding out if there is a significant difference between the mean of three or more groups.

It checks the impact of various factors by comparing groups (samples) based on their respective mean. ANOVA tests the null hypothesis that all group means are equal, against the alternative hypothesis that at least one group mean is different.

* The dependent variable is approximately normally distributed within each group. This assumption is more critical for smaller sample sizes.
* The samples are selected at random and should be independent of one another.
* All groups have equal standard deviations.
* Each data point should belong to one and only one group. There should be no overlap or sharing of data points between groups.

3.17.1 **Types of ANOVA**

**There are two main types of ANOVA:**

**One-way ANOVA:**

* This is the most basic form of ANOVA and is used when there is only one independent variable with more than two levels or groups.
* It assesses whether there are any statistically significant differences among the means of the groups.
* One-way ANOVA (Analysis of Variance) is a statistical method used to compare the means of three or more independent groups to determine if there are significant differences among them. It assesses the impact of a single factor on a continuous response variable by partitioning the total variability into between-group and within-group variability. If the between-group variability is significantly larger than the within-group variability, it suggests that at least one group mean is different from the others. This technique is widely used in experimental research to test hypotheses about group differences.

**Two-way ANOVA:**

* Extending the one-way ANOVA, two-way ANOVA involves two independent variables. It allows for the examination of the main effects of each variable as well as the interaction between them.
* The interaction effect explores whether the effect of one variable on the dependent variable is different depending on the level of the other variable.

Two-way ANOVA (Analysis of Variance) is a statistical method used to examine the influence of two different categorical independent variables on one continuous dependent variable. This technique not only assesses the main effects of each independent variable but also explores the interaction effect between them. By partitioning the total variability in the data into components attributed to each factor and their interaction, two-way ANOVA provides a detailed understanding of how multiple factors influence the response variable. The main effects analysis in two-way ANOVA evaluates whether changes in the levels of each independent variable significantly affect the dependent variable. For instance, in an experiment studying the effects of different teaching methods (first factor) and varying study durations (second factor) on students' test scores (dependent variable), the main effects analysis would determine if the teaching method and study duration individually influence test scores.

The interaction effect analysis investigates whether the effect of one independent variable on the dependent variable changes across the levels of the other independent variable. In the teaching methods example, the interaction effect would reveal if the effectiveness of teaching methods varies depending on study duration. Two-way ANOVA is a robust method for analyzing experimental data, enabling researchers to understand both the individual and combined effects of multiple factors on a response variable, thereby informing more nuanced interpretations and decisions.