MALLA REDDY COLLEGE OF ENGINEERING & TECHNOLOGY

DEPARTMENT OF DATA SCIENCE

Data Visualization in R

UNIT-I

Introduction to Statistics

Introduction to Statistics, Difference between inferential statistics and descriptive statistics, Random Variables, Sampling , Sample Statistics and Sampling Distributions.

Introduction to Data Visualization:

Data visualization and its importance, Characteristics of Effective Graphical Visual, Advantages and Disadvantages of using R for data Visualization, **Application areas,**

R and R studio Installation

UNIT – I

Introduction to Statistics

Statistics is a mathematical science that includes methods for collecting, organizing, analyzing and visualizing data in such a way that meaningful conclusions can be drawn.
Statistics is also a field of study that summarizes the data, interpret the data making decisions based on the data.
Statistics is composed of two broad categories:

1. Descriptive Statistics
2. Inferential Statistics

1. Descriptive Statistics

Descriptive statistics describes the characteristics or properties of the data. It helps to summarize the data in a meaningful data in a meaningful way. It allows important patterns to emerge from the data. Data summarization techniques are used to identify the properties of data. It is helpful in understanding the distribution of data. They do not involve in generalizing beyond the data.

Two types of descriptive statistics

1. Measures of Central Tendency: (Mean , Median , Mode)
2. Measures of data spread or dispersion (range, quartiles, variance and standard deviation)

Measures of Central Tendency: (Mean , Median , Mode)

A measure of central tendency is a single value that attempts to describe a set of data by identifying the central position within that set of data. The mean, median and mode are all valid measures of central tendency.

Mean (Arithmetic)

The mean (or average) is the most popular and well known measure of central tendency. It can be used with both discrete and continuous data, although its use is most often with continuous data.
The mean is equal to the sum of all the values in the data set divided by the number of values in the data set. So, if we have values in a data set and they have values $x_1, x_2, \ldots x_n$, the sample mean, usually denoted by $\bar{x}$.

$\bar{\bar{x}} = (x_1, x_2, \ldots x_n)/ n.$

An important property of the mean is that it includes every value in the data set as part of the calculation. In addition, the mean is the only measure of central tendency where the sum of the deviations of each value from the mean is always zero.

Median:

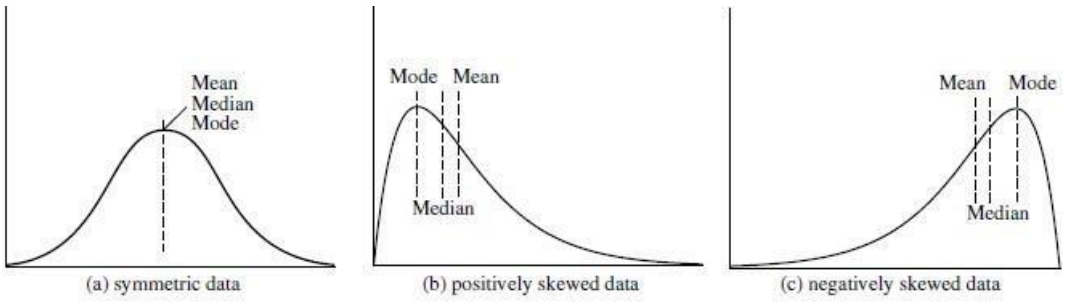
The median is the middlescore for a set of data that has been arranged in order of magnitude. The median is less affected by outliers and skewed data. It is a holistic measure. It is easy method of approximation of median value of a large data set.

Mode

The mode is the most frequent score in our data set. The mode is used for categorical data where we want to know which is the most common category occurring in the population. There are possibilities for the greatest frequency to correspond to different values. This results in more than one,two or more modes in a dataset. They are called as unimodal, bimodal and multimodal datasets. If each data occurs only once then the mode is equal to zero.

Unimodal frequency curve with symmetric data distribution , the mean median and mode are all the same.

In real applications the data is not symmetrical and they are asymmetric.It might be positively skewed or negatively skewed. If positively skewed then mode is smaller than median and in negatively skewed the mode occurs at a value greater than the median.



Mean, median, and mode of symmetric versus positively and negatively skewed data.

Measures of spread:

Measures of spread are the ways of summarizing a group of data by describing how scores are spread out. To describe this spread, a number of statistics are available to us, including the range, quartiles, absolute deviation, variance and standard deviation.

- The degree to which numerical data tend to spread is called the dispersion, or variance of the data. The common measures of data dispersion: Range, Quartiles, Outliers, and Boxplots.

Range : Range of the set is the difference between the largest (max()) and smallest (min()) values. Ex:Step 1: Sort the numbers in order, from smallest to largest: 7, 10, 21, 33, 43, 45, 45, 65, 67, 87, 98, 99

Step 2: Subtract the smallest number in the set from the largest number in the set:
99 – 7 = 92
The range is 92

Quartiles : Percentile : *k*th percentile of a set of data in numerical order is the value *xi h*aving the property that *k* percent of the data entries lie at or below *xi*

- The first quartile (*Q*1) is the 25th percentile;
- The third quartile (*Q*3) is the 75th percentile
- The distance between the first and third quartiles is the range covered by the middle half of the data.
- Interquartile range (*IQR*) and is defined as *IQR* = *Q*3 - Q1.
- Outliers is to single out values falling at least 1.*5 *IQR* above the third quartile or below the first quartile.
- *Five-number summary*: median, the quartiles *Q*1 and *Q*3, and the smallest and largest individual observations comprise the five number summary: *Minimum*; *Q*1; *Median*; *Q*3; *Maximum*

Example : Quartiles

- Start with the following data set:
- 1, 2, 2, 3, 4, 6, 6, 7, 7, 7, 8, 11, 12, 15, 15, 15, 17, 17, 18, 20
- There are a total of twenty data points in the set. There is an even number of data values, hence the median is the mean of the tenth and eleventh values.
- the median is: (7 + 8)/2 = 7.5.
- The median of the first half of the set is found between the fifth and sixth values of:
- 1, 2, 2, 3, 4, 6, 6, 7, 7, 7
- Thus the first quartile is found to equal *Q*1 = (4 + 6)/2 = 5
- To find the third quartile, examine the top half of the original data set. The median of
- 8, 11, 12, 15, 15, 15, 17, 17, 18, 20
- is (15 + 15)/2 = 15. Thus the third quartile *Q*3 = 15.

A small interquartile range indicates data that is clumped about the median. A larger interquartile range shows that the data is more spread out

Variance and Standard Deviation

The **variance** of N observations, $x_1, x_2, \ldots, x_N$, is

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2 = \frac{1}{N}\left[\sum x_i^2 - \frac{1}{N}(\sum x_i)^2\right],$$

where ẍ is the mean value of the observations
The standard deviation, σ, of the observations is the square root of the variance, σ^2

measures spread about the mean and should be used only when the mean is chosen σ as the measure of center.
σ =0 only when there is no spread, that is, when all observations have the same value.
Otherwise σ > 0.

For any population with mean μ and standard deviation σ:

- The **mean**, or center of the sampling distribution of $\bar{x}$, is equal to the population mean μ: $\mu_{\bar{x}} = \mu$.
- The **standard deviation** of the sampling distribution is $\sigma/\sqrt{n}$, where n is the sample size: $\sigma_{\bar{x}} = \sigma/\sqrt{n}$.

Application

Hypokalemia is diagnosed when blood potassium levels are below 3.5mEq/dl. Let's assume that we know a patient whose measured potassium levels vary daily according to a normal distribution $N(\mu = 3.8, \sigma = 0.2)$.
If only one measurement is made, what is the probability that this patient will be misdiagnosed with Hypokalemia?

$$z = \frac{(x - \mu)}{\sigma} = \frac{3.5 - 3.8}{0.2} = -1.5, \quad P(z < -1.5) = 0.0668 \approx 7\%$$

Instead, if measurements are taken on 4 separate days, what is the probability of a misdiagnosis?

$$z = \frac{(\bar{x} - \mu)}{\sigma/\sqrt{n}} = \frac{3.5 - 3.8}{0.2/\sqrt{4}} = -3, \quad P(z < -3) = 0.0013 \approx 0.1\%$$

Inferential Statistics – Definition and Types

Inferential statistics is generally used when the user needs to make a conclusion about the whole population at hand, and this is done using the various types of tests available. It is a technique which is used to understand trends and draw the required conclusions about a large population by taking and analyzing a sample from it. Descriptive statistics, on the other hand, is only about the smaller sized data set at hand – it usually does not involve large populations. Using variables and the relationships between them from the sample, we will be able to make generalizations and predict other relationships within the whole population, regardless of how large it is.

With inferential statistics, data is taken from samples and generalizations are made about a population.Inferential statistics use statistical models to compare sample data to other samples or to previous research.

There are two main areas of inferential statistics:

1. Estimating parameters:

This means taking a statistic from the sample data (for example the sample mean) and using it to infer about a population parameter (i.e. the population mean).There may be sampling variations because of chance fluctuations, variations in sampling techniques, and other sampling errors. Estimation about population characteristics may be influenced by such factors. Therefore, in estimation the important point is that to what extent our estimate is close to the true value.

Characteristics of Good Estimator: A good statistical estimator should have the following characteristics, (i) Unbiased (ii) Consistent (iii) Accuracy

i) Unbiased

An unbiased estimator is one in which, if we were to obtain an infinite number ofrandom samples of a certain size, the mean of the statistic would be equal to theparameter. The sample mean, (x) is an unbiased estimate of population mean (μ)because if we look at possible random samples of size N from a population, thenmean of the sample would be equal to μ.

ii) Consistent

A consistent estimator is one that as the sample size increased, the probability thatestimate has a value close to the parameter also increased. Because it is a consistentestimator, a sample mean based on 20 scores has a greater probability of beingcloser to (μ) than does a sample mean based upon only 5 scores

iii) Accuracy

The sample mean is an unbiased and consistent estimator of population mean (μ).But we should not over look the fact that an estimate is just a rough or approximatecalculation. It is unlikely in any estimate that (x) will be exactly equal to populationmean (μ). Whether or not x is a good estimate of (μ) depends upon the representativeness of sample, the sample size, and the variability of scores in the population.

2. **Hypothesis tests.** This is where sample data can be used to answer research questions. For example, we might be interested in knowing if a new cancer drug is effective. Or if breakfast helps children perform better in schools.

Inferential statistics is closely tied to the logic of hypothesis testing. We hypothesize that this value characterise the population of observations. The question is whether that hypothesis is reasonable evidence from the sample. Sometimes hypothesis testing is referred to as statistical decision-making process. In day-to-day situations we are required to take decisions about the population on the basis of sample information.

Statement of Hypothesis

A statistical hypothesis is defined as a statement, which may or may not be true about the population parameter or about the probability distribution of the parameter that we wish to validate on the basis of sample information. Most times, experiments are performed with random samples instead of the entire population and inferences drawn from the observed results are then generalised over to the entire population. But before drawing inferences about the population it should be always kept in mind that the observed results might have come due to chance factor. In order to have an accurate or more precise inference, the chance factor should be ruled out.

Null Hypothesis

The probability of chance occurrence of the observed results is examined by the null hypothesis (H0). Null hypothesis is a statement of no differences. The other way to state null hypothesis is that the two samples came from the same population. Here, we assume that population is normally distributed and both the groups have equal means and standard deviations.

Since the null hypothesis is a testable proposition, there is counter proposition to it known as alternative hypothesis and denoted by H1 . In contrast to null hypothesis, the alternative hypothesis (H1) proposes that

i) the two samples belong to two different populations,
ii) their means are estimates of two different parametric means of the respective population, and
iii) there is a significant difference between their sample means.

The alternative hypothesis (H1) is not directly tested statistically; rather its acceptance or rejection is determined by the rejection or retention of the null hypothesis. The probability 'p' of the null hypothesis being correct is assessed by a statistical test. If probability 'p' is too low, H0 is rejected and H1 is accepted.

It is inferred that the observed difference is significant. If probability 'p' is high, H0 is accepted and it is inferred that the difference is due to the chance factor and not due to the variable factor.

Level of Significance

The level of significance is defined as the probability of rejecting a null hypothesis by the test when it is really true, which is denoted as α. That is, P (Type I error) = α.

Confidence level:

Confidence level refers to the possibility of a parameter that lies within a specified range of values, which is denoted as *c*. Moreover, the confidence level is connected with the level of significance. The relationship between level of significance and the confidence level is $c=1-\alpha$.

The common level of significance and the corresponding confidence level are given below:

- The level of significance 0.10 is related to the 90% confidence level.
- The level of significance 0.05 is related to the 95% confidence level.
- The level of significance 0.01 is related to the 99% confidence level.

The rejection rule is as follows:

- If p-value $\leq$ level of significance (α), then reject the null hypothesis H_0.
- If p-value $>$ level of significance (α), then do not reject the null hypothesis H_0.

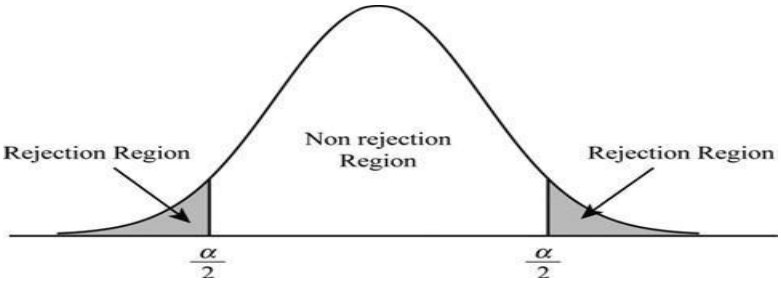
Rejection region:

The rejection region is the values of test statistic for which the null hypothesis is rejected.

Non rejection region:

The set of all possible values for which the null hypothesis is not rejected is called the rejection region.

The rejection region for two-tailed test is shown below:



The rejection region for one-tailed test is given below:

- In the left-tailed test, the rejection region is shaded in left side.
- In the right-tailed test, the rejection region is shaded in right side.

One-tail and Two-tail Test

Depending upon the statement in alternative hypothesis (H1), either a one-tail or twotail test is chosen for knowing the statistical significance. A one-tail test is a directional test. It is formulated to find the significance of both the magnitude and the direction (algebraic sign) of the observed difference between two statistics. Thus, in two-tailed tests researcher is interested in testing whether one sample mean is significantly higher (alternatively lower) than the other sample mean.

Types of Inferential Statistics Tests

There are many tests in this field, of which some of the most important are mentioned below.

1. Linear Regression Analysis

In this test, a linear algorithm is used to understand the relationship between two variables from the data set. One of those variables is the dependent variable, while there can be one or more independent variables used. In simpler terms, we try to predict the value of the dependent variable based on the available values of the independent variables. This is usually represented by using a scatter plot, although we can also use other types of graphs too.

2. Analysis of Variance

This is another statistical method which is extremely popular in data science. It is used to test and analyse the differences between two or more means from the data set. The significant differences between the means are obtained, using this test.

3. Analysis of Co-variance

This is only a development on the Analysis of Variance method and involves the inclusion of a continuous co-variance in the calculations. A co-variate is an independent variable which is continuous, and is used as regression variables. This method is used extensively in statistical modelling, in order to study the differences present between the average values of dependent variables.

4. Statistical Significance (T-Test)

A relatively simple test in inferential statistics, this is used to compare the means of two groups and understand if they are different from each other. The order of difference, or how significant the differences are can be obtained from this.

5. Correlation Analysis

Another extremely useful test, this is used to understand the extent to which two variables are dependent on each other. The strength of any relationship, if they exist, between the two variables can be obtained from this. You will be able to understand whether the variables have a strong correlation or a weak one. The correlation can also be negative or positive, depending upon the variables. A negative correlation means that the value of one variable decreases while the value of the other increases and positive correlation means that the value both variables decrease or increase simultaneously.

Differences between Descriptive and Inferential Statistics

Descriptive Statistics	Inferential Statistics
Concerned with describing the target population	Make inferences from the sample and generalize them to the population
Organise, analyse, present the data in a meaningful way	Compare, tests and predicts future outcomes
The analysed results are in the form of graphs, charts etc	The analysed results are the probability scores
Describes the data which is already known	Tries to make conclusions about the population beyond the data available
Tools: Measures of central tendency and measures of spread	Tools: Hypothesis tests, analysis of variance etc

Random Variables

A random variable, X, is a variable whose possible values are numerical outcomes of a random phenomenon. There are two types of random variables, discrete and continuous.

Example of Random variable

- A person's blood type
- Number of leaves on a tree
- Number of times a user visits LinkedIn in a day
- Length of a tweet.

Discrete Random Variables :

A discrete random variable is one which may take on only a countable number of distinct values such as 0,1,2,3,4,..... Discrete random variables are usually counts. If a random variable can take only a finite number of distinct values, then it must be discrete. Examples of discrete random variables include the number of children in a family, the Friday night attendance at a cinema, the number of patients in a doctor's surgery, the number of defective light bulbs in a box of ten.

The ***probability distribution*** of a discrete random variable is a list of probabilities associated with each of its possible values. It is also sometimes called the probability function or the probability mass function

Suppose a random variable X may take k different values, with the probability that $X = x_i$ defined to be $P(X = x_i) = p_i$. The probabilities p_i must satisfy the following:

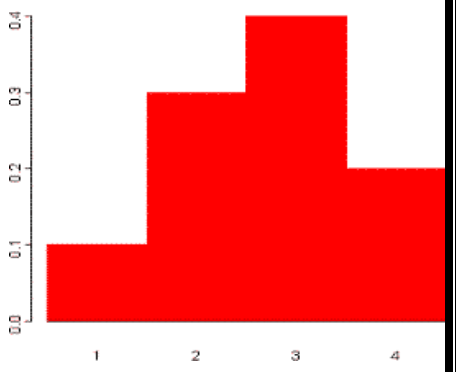
1: $0 \leq p_i \leq 1$ *for each* i

2: $p_1 + p_2 + ... + p_k = 1$.

Example

Suppose a variable X can take the values 1, 2, 3, or 4. The probabilities associated with each outcome are described by the following table:

Outcome	1	2	3	4
Probability	0.1	0.3	0.4	0.2



The probability that X is equal to 2 or 3 is the sum of the two probabilities: $P(X = 2 \text{ or } X = 3) = P(X = 2) + P(X = 3) = 0.3 + 0.4 = 0.7$. Similarly, the probability that X is greater than 1 is equal to $1 - P(X = 1) = 1 - 0.1 = 0.9$, by the complement rule.

Continuous Random Variables

A ***continuous random variable*** is one which takes an infinite number of possible values. Continuous random variables are usually measurements. Examples include height, weight, the amount of sugar in an orange, the time required to run a mile.

A continuous random variable is not defined at specific values. Instead, it is defined over an *interval* of values, and is represented by the ***area under a curve*** (known as an *integral*). The probability of observing any single value is equal to 0, since the number of values which may be assumed by the random variable is infinite.

Suppose a random variable *X* may take all values over an interval of real numbers. Then the probability that *X* is in the set of outcomes *A*, *P(A)*, is defined to be the area above *A* and under a curve. The curve, which represents a function *p(x)*, must satisfy the following:

1: *The curve has no negative values (p(x) ≥ 0 for all x)*

2: *The total area under the curve is equal to 1.*

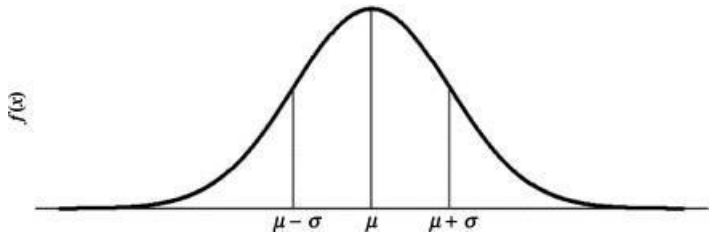
A curve meeting these requirements is known as a ***density curve***.

All random variables (discrete and continuous) have a ***cumulative distribution function***. It is a function giving the probability that the random variable *X* is less than or equal to *x*, for every value *x*. For a discrete random variable, the cumulative distribution function is found by summing up the probabilities.

Normal Probability Distribution

The Bell-Shaped Curve

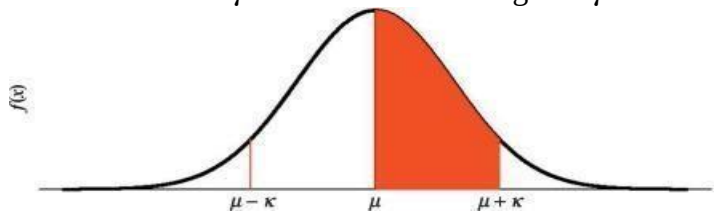
The **Bell-shaped Curve** is commonly called the **normal curve** and is mathematically referred to as the Gaussian probability distribution. Unlike Bernoulli trials which are based on discrete counts, the **normal distribution** is used to determine the probability of a continuous random variable.



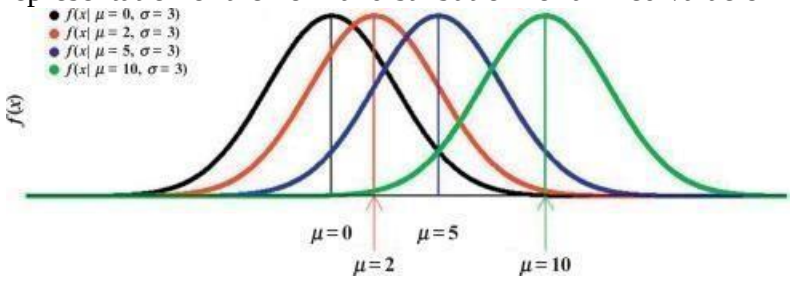
The **normal** or Gaussian Probability Distribution is most popular and important because of its unique mathematical properties which facilitate its application to practically any physical problem in the real world.The constants $\boldsymbol{\mu}$ and $\boldsymbol{\sigma^2}$ are the parameters;

- "$\boldsymbol{\mu}$" is the population true mean (or expected value) of the subject phenomenon characterized by the continuous random variable, ***X***,
- "$\boldsymbol{\sigma^2}$" is the population true variance characterized by the continuous random variable, ***X***.
- Hence, "$\boldsymbol{\sigma}$" the population standard deviation characterized by the continuous random variable ***X***;
- the points located at μ−σ and μ+σ are the points of inflection; that is, where the graph changes from cupping up to cupping down

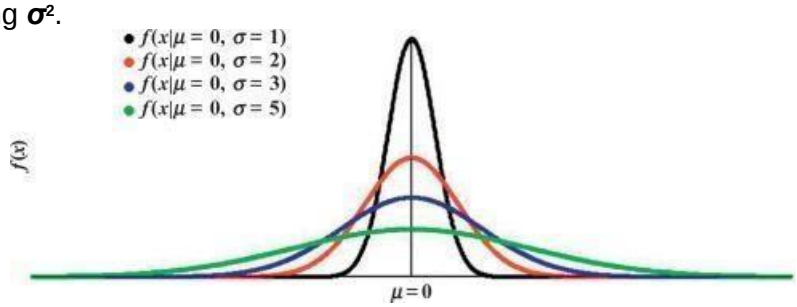
The **normal curve graph of the normal probability distribution**) is **symmetric** with respect to the mean $\boldsymbol{\mu}$ as the **central position**. That is, the area between $\boldsymbol{\mu}$ and $\boldsymbol{\kappa}$ units to the left of $\boldsymbol{\mu}$ is equal to the area between $\boldsymbol{\mu}$ and $\boldsymbol{\kappa}$ units to the right of $\boldsymbol{\mu}$.



There is not a unique **normal probability distribution.**The figure below is a graphical representation of the normal distribution for a fixed value of σ2 with μ varying.



The figure below is a graphical representation of the **normal distribution** for a fixed value of $\boldsymbol{\mu}$ with varying $\boldsymbol{\sigma^2}$.



SAMPLING and SAMPLING DISTRIBUTION

Sampling is a process used in statistical analysis in which a predetermined number of observations are taken from a larger population.It helps us to make statistical inferences about the population.A population can be defined as a whole that includes all items and characteristics of the research taken into study. However, gathering all this information is time consuming and costly. We therefore make inferences about the population with the help of samples.

Random sampling:

In data collection, every individual observation has equal probability to be selected into a sample. In random sampling, there should be no pattern when drawing a sample.

Probability sampling:

It is the sampling technique in which every individual unit of the population has greater than zero probability of getting selected into a sample.

Non-probability sampling:

It is the sampling technique in which some elements of the population have no probability of getting selected into a sample.

Cluster samples:

It divides the population into groups (clusters). Then a random sample is chosen from the clusters.

Systematic sampling : select sample elements from an ordered frame. A sampling frame is just a list of participants that we want to get a sample from.

Stratified sampling : sample each subpopulation independently. First, divide the population into homogeneous (very similar) subgroups before getting the sample. Each population member only belongs to one group. Then apply simple random or a systematic method within each group to choose the sample.

Sampling Distribution

A sampling distribution is a probability distribution of a statistic. It is obtained through a large number of samples drawn from a specific population.It is the distribution of all possible values taken by the statistic when all possible samples of a fixed size n are taken from the population.

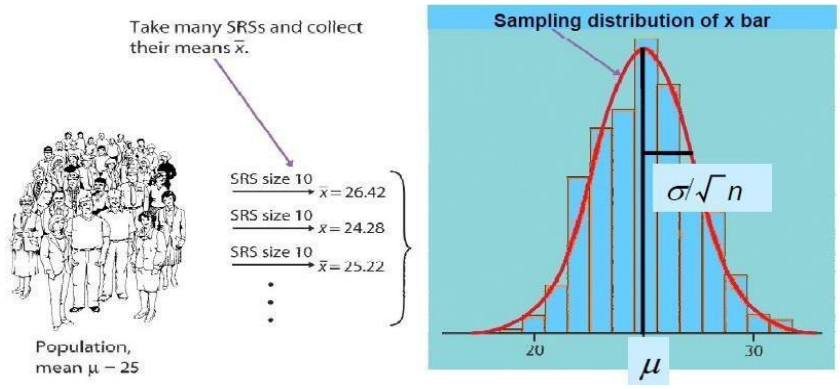
Sampling Distributions and Inferential Statistics

Sampling distributions are important for inferential statistics. A population is specified and the sampling distribution of the mean and the range were determined. In practice, the process proceeds the other way: the sample data is collected and from these data we estimate parameters of the sampling distribution. This knowledge of the sampling distribution can be very useful.

- Knowing the degree to which means from different samples would differ from each other and from the population mean (this would give an idea of how close the particular sample mean is likely to be to the population mean)
- The most common measure of how much sample means differ from each other is the standard deviation of the sampling distribution of the mean. This standard deviation is called the standard error of the mean.
- If all the sample means were very close to the population mean, then the standard error of the mean would be small. On the other hand, if the sample means varied considerably, then the standard error of the mean would be large.

Sampling distribution of the sample mean

1. We take many random samples of a given size n from a population with mean μ and standard deviation σ.
2. Some sample means will be above the population mean μ and some will be below, making up the sampling distribution.



Introduction to Data Visualization

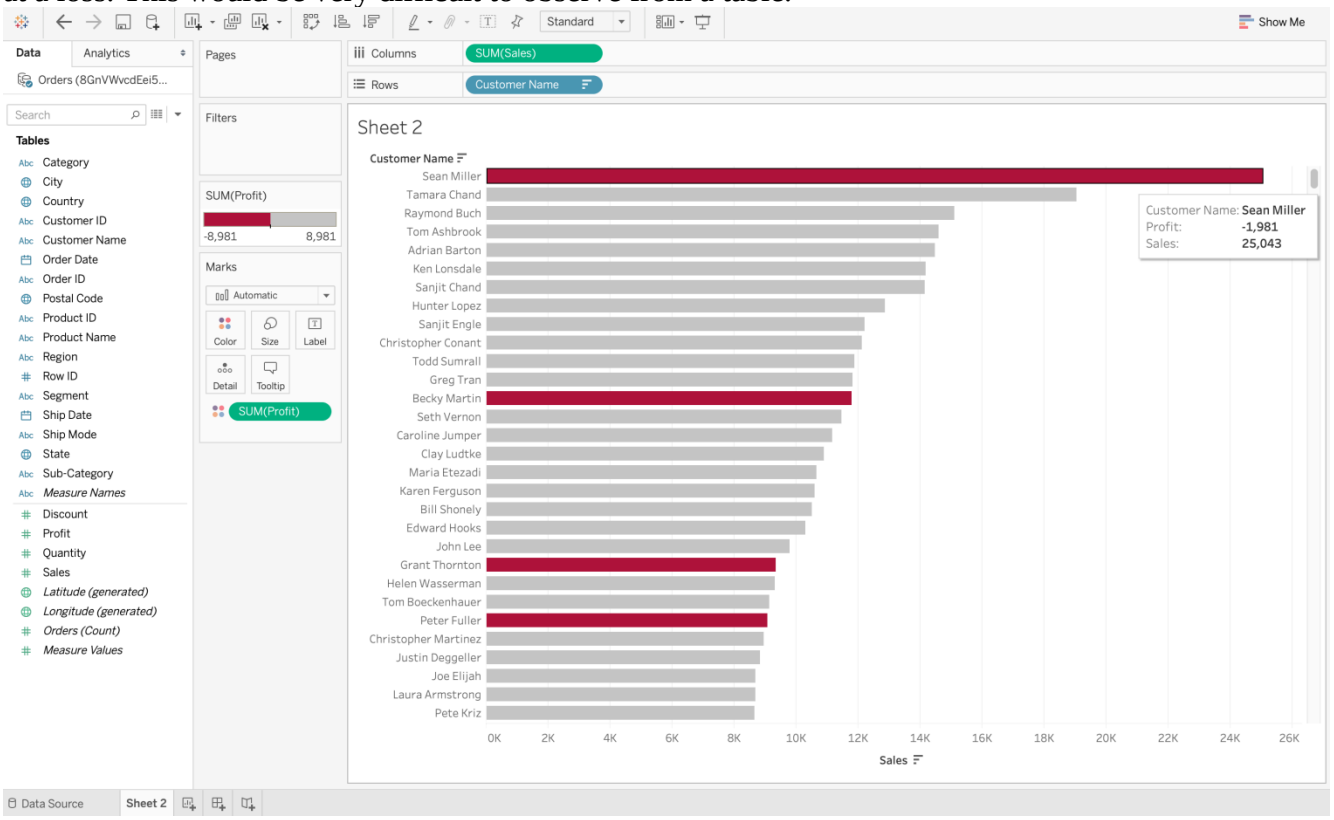
What is Data Visualization?

- Data visualization is the process of creating visual representations of data sets to better understand the underlying information.
- These visuals can take many different forms, but often include charts, graphs, and maps. The goal of data visualization is to make complex data more accessible and understandable, and it can be used in a variety of contexts, from scientific research to business analytics.
- Data visualization is sometimes used interchangeably with data analysis, but the two terms are not quite the same. Data analysis focuses on understanding the underlying meaning of data sets, while data visualization focuses on creating visual representations of that data
- Data visualization can be a valuable tool for both data analysts and laypeople alike, as it can help to make complex information more digestible

Why is Data Visualization Important?

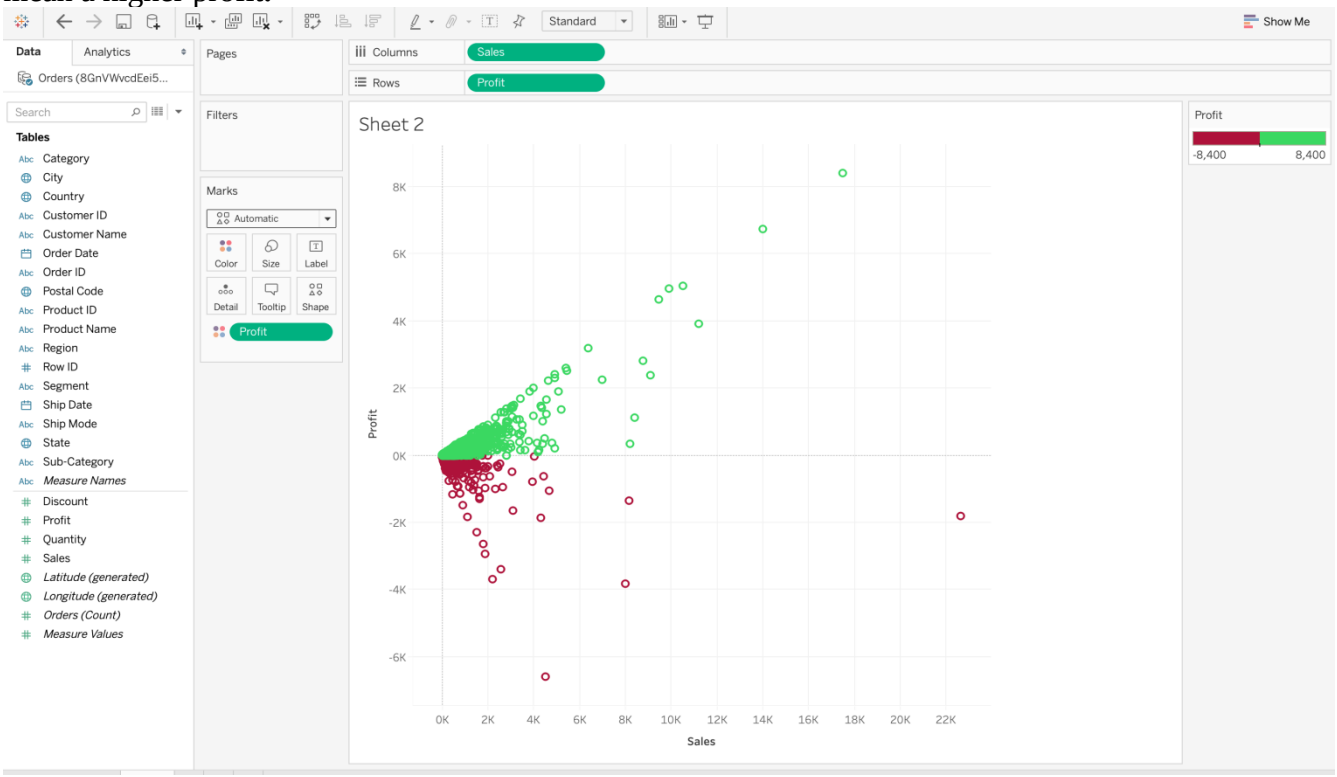
1. Data Visualization Discovers the Trends in Data

The most important thing that data visualization does is discover the trends in data. After all, it is much easier to observe data trends when all the data is laid out in front of you in a visual form as compared to data in a table. For example, the screenshot below on Tableau demonstrates the sum of sales made by each customer in descending order. However, the colour red denotes loss while grey denotes profits. So it is very easy to observe from this visualization that even though some customers may have huge sales, they are still at a loss. This would be very difficult to observe from a table.



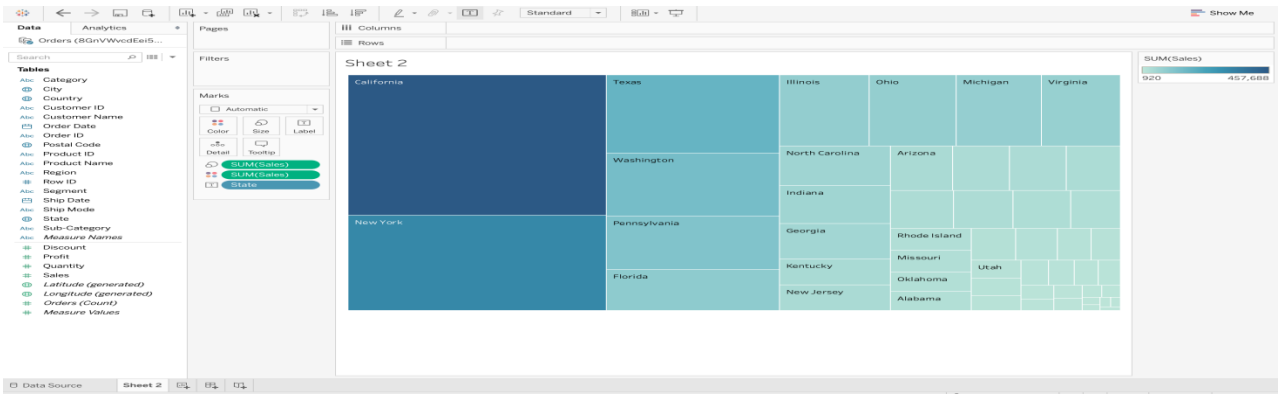
2. Data Visualization Provides a Perspective on the Data

Data Visualization provides a perspective on data by showing its meaning in the larger scheme of things. It demonstrates how particular data references stand with respect to the overall data picture. In the data visualization below, the data between sales and profit provides a data perspective with respect to these two measures. It also demonstrates that there are very few sales above 12K and higher sales do not necessarily mean a higher profit**.**



3. Data Visualization Puts the Data into the Correct Context

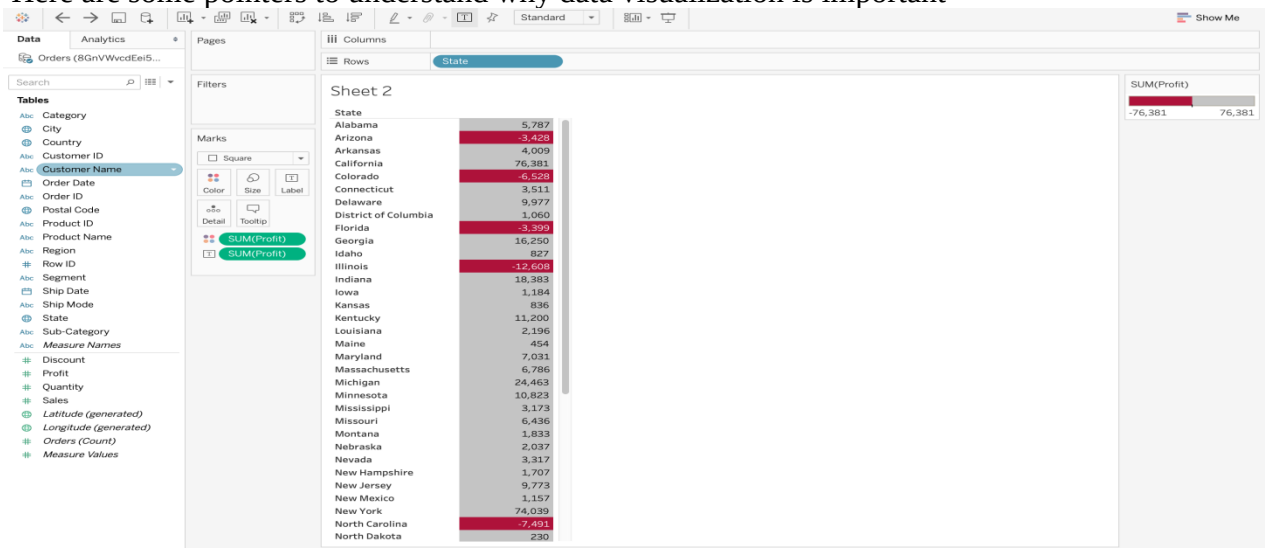
Since context provides the whole circumstances of the data, it is very difficult to grasp by just reading numbers in a table. In the below data visualization on Tableau, a TreeMap is used to demonstrate the number of sales in each region of the United States. It is very easy to understand from this data visualization that California has the largest number of sales out of the total number since the rectangle for California is the largest. But this information is not easy to understand outside of context without data visualization**.**



4. Data Visualization Saves Time

In the screenshot below, it is very easy to identify the states that have suffered a net loss rather than a profit. This is because all the cells with a loss are colored red using a heat map, so it is obvious states have suffered a loss. Compare this to a normal table where you would need to check each cell to see if it has a negative value to determine a loss. Obviously, data visualization saves a lot of time in this situation!

Here are some pointers to understand why data visualization is important



5. Data Visualization Tells a Data Story

Data visualization is also a medium to tell a data story to the viewers. The visualization can be used to present the data facts in an easy-to-understand form while telling a story and leading the viewers to an inevitable conclusion. This data story, like any other type of story, should have a good beginning, a basic plot, and an ending that it is leading towards. For example, if a data analyst has to craft a data visualization for company executives detailing the profits on various products, then the data story can start with the profits and losses of various products and move on to recommendations on how to tackle the losses.

Importance of Data Visualization in different areas:

1. Importance of Data Visualization in Healthcare

Healthcare is an industry that relies heavily on data. From patient medical records to insurance claims, there is a lot of data that needs to be collected, analyzed, and interpreted. Data visualization plays an important role in healthcare by allowing doctors and other medical professionals to make better-informed decisions. For example, let's say a hospital wanted to reduce the number of patient readmissions. They could use data visualization to look at readmission rates over time, identify which types of patients are more likely to be readmitted, and develop interventions to target those high-risk patients.

2. Importance of Data Visualization in Analytics

Analytics is the process of turning data into insights. Data visualization is a key part of analytics because it allows analysts to take a huge dataset and distill it down into something that can be easily understood and interpreted.

Without data visualization, analysts would be stuck looking at raw data all day long without being able to identify any patterns or trends. But with data visualization, analysts can quickly see relationships between

variables and make better-informed decision

3. Importance of Data Visualization in Business Intelligence

Business intelligence (BI) is the process of transforming raw data into actionable insights. Data visualization plays an important role in BI because it allows businesses to see their data in a new light and make better-informed decisions about their overall strategy.

For example, let's say a company wanted to increase sales by 10% this year. They could use data visualization to track sales over time, identify which products are selling well and which ones are not, and develop **marketing** campaigns and targeted promotions accordingly.

4. Importance of Data Visualization in Data Science

Data science is all about extracting insights from large datasets. Data visualization plays an important role in data science by allowing scientists to visualize their data and find patterns that are otherwise impossible to discover.

For example, let's say a scientist was studying a disease and wanted to find out which genetic factors were associated with it. They could use data visualization techniques like cluster analysis or heat maps to try to find groups of genes that are similar to the disease group and then further investigate those genes.

5. Importance of Data Visualization in Machine Learning

Machine learning is a subfield of artificial intelligence that deals with the design and development of algorithms that can learn from data. Data visualizations are important in machine learning because they can be used to understand complex datasets and identify patterns. Machine learning algorithms can then be developed to automatically detect these patterns.

For example: Google's PageSpeed Insights tool uses machine learning to automatically analyze website performances and provides recommendations on how to improve them.

6. Importance of Data Visualization in IoT

The Internet of Things (IoT) is a network of physical devices connected to the internet that are able to collect and share data. IoT devices can include everything from fitness trackers and smart thermostats to industrial machines and self-driving cars. Data visualization is important for IoT because it helps make sense of the enormous amount of data being generated by these devices.

For example: Data visualization can help detect issues and problems with the IoT devices so that the company can take corrective action before they cause major disruptions.

7. Importance of Data Visualization in Big Data Analytics

Big data analytics can be very overwhelming because of the sheer volume of data that is involved. Data visualization can help us make sense of all this data by helping us identify correlations and structures that we would not be able to see otherwise.

For example, if we are looking at a dataset with millions of rows, it would be very difficult to find patterns without data visualization. But by utilizing a visualization tool, we can easily find patterns in the data.

8. Importance of Data Visualization in Business Analytics

Business analytics deals with understanding and analyzing business data so that businesses can make better decisions. Data visualization plays an important role in business analytics because it helps businesses see the latest trends and patterns.

For example, if we are looking at sales data over time, it might be difficult to spot trends without data visualization. However, if we use a tool like Tableau or Google Charts to visualize the data ,we would be

able to see any ups or downs in sales much more easily.

Benefits of data visualization

1. **Helps You See Trends and Patterns**
2. **Helps You Make Better Decisions**
3. **Helps You Communicate Your Findings More Effectively**
4. **Data Visualization Helps Users See Outliers**
5. **Data Visualization is Easy to Use**

Disadvantages of data visualization

1. **Can be time-consuming:** Creating visualizations can be a time-consuming process, especially when dealing with large and complex datasets. This can slow down the machine learning workflow and reduce productivity.
2. **Can be misleading:** While data visualization can help identify patterns and relationships in data, it can also be misleading if not done correctly. Visualizations can create the impression of patterns or trends that may not actually exist, leading to incorrect conclusions and poor decision-making.
3. **Can be difficult to interpret:** Some types of visualizations, such as those that involve 3D or interactive elements, can be difficult to interpret and understand. This can lead to confusion and misinterpretation of the data.
4. **May not be suitable for all types of data:** Certain types of data, such as text or audio data, may not lend themselves well to visualization. In these cases, alternative methods of analysis may be more appropriate.
5. **May not be accessible to all users:** Some users may have visual impairments or other disabilities that make it difficult or impossible for them to interpret visualizations. In these cases, alternative methods of presenting data may be necessary to ensure accessibility.

Top Data Visualization Tools

The following are the 10 best Data Visualization Tools

1. Tableau
2. Looker
3. Zoho Analytics
4. Sisense
5. IBM Cognos Analytics
6. Qlik Sense
7. Domo
8. Microsoft Power BI
9. Klipfolio
10. SAP Analytics Cloud

Data Visualization Libraries

R:

- ggplot2
- Plotly
- Leaflet
- Esquisse
- Lattice

Advantages of Data Visualization in R:

R has the following advantages over other tools for data visualization:

- R offers a broad collection of visualization libraries along with extensive online guidance on their usage

Data Visualization Libraries in R

  - ggplot2
  - Plotly
  - Leaflet
  - Esquisse
  - Lattice

- R also offers data visualization in the form of 3D models and multipanel charts.
- Through R, we can easily customize our data visualization by changing axes, fonts, legends, annotations, and labels.

Disadvantages of Data Visualization in R:

R also has the following disadvantages:

- R is only preferred for data visualization when done on an individual standalone server.
- Data visualization using R is slow for large amounts of data as compared to other counterparts.

Application Areas:

- Presenting analytical conclusions of the data to the non-analysts departments of your company.
- Health monitoring devices use data visualization to track any anomaly in blood pressure, cholesterol and others.
- To discover repeating patterns and trends in consumer and marketing data.
- Meteorologists use data visualization for assessing prevalent weather changes throughout the world.
- Real-time maps and geo-positioning systems use visualization for traffic monitoring and estimating travel time.

R overview and Installation

R is a programming language and software environment for statistical analysis, graphics representation and reporting. R was created by Ross Ihaka and Robert Gentleman at the University of Auckland, New Zealand, and is currently developed by the R Development Core Team.

The core of R is an interpreted computer language which allows branching and looping as well as modular programming using functions. R allows integration with the procedures written in the C, C++, .Net, Python or FORTRAN languages for efficiency.

R is freely available under the GNU General Public License, and pre-compiled binary versions are provided for various operating systems like Linux, Windows and Mac.

R is free software distributed under a GNU-style copy left, and an official part of the GNU project called **GNUs**.

To Install R:

1. Open an internet browser and go to www.r-project.org.
2. Click the "download R" link in the middle of the page under "Getting Started."
3. Select a CRAN location (a mirror site) and click the corresponding link.
4. Click on the "Download R for Windows" link at the top of the page.
5. Click on the "install R for the first time" link at the top of the page.
6. Click "Download R for Windows" and save the executable file somewhere on computer. Run the .exe file and follow the installation instructions.
7. Now that R is installed, next step is to download and install RStudio.

To Install RStudio

1. Go to www.rstudio.com and click on the "Download RStudio" button.
2. Click on "Download RStudio Desktop."
3. Click on the version recommended for your system, or the latest Windows version, and save the executable file. Run the .exe file and follow the installation instructions.