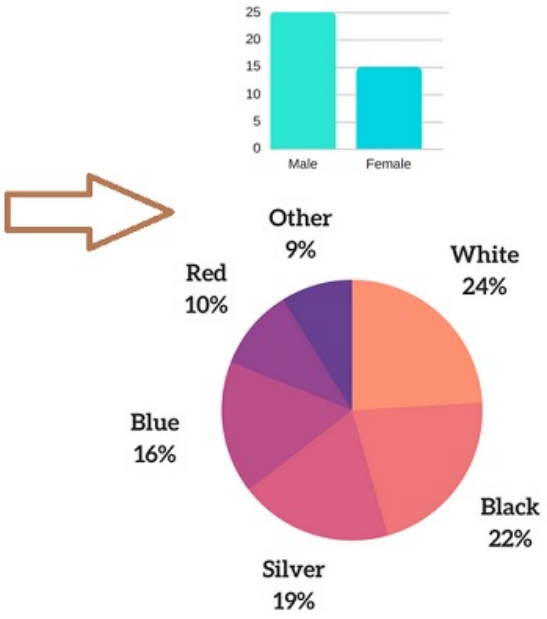
Descriptive statistics

- Descriptive statistics describes, shows, and summarizes the features of a dataset.
- Descriptive statistics are just the representation of the data (sample) available and not based on any theory of probability.
- Charts and graphs are often used to present descriptive statistics.
- It helps Data analysts to understand the data better.

	A	B	C	D
1	Respondent Number	Age	Gender	Favorite Car Color
2	1	22	M	White
3	2	37	F	Silver
4	3	45	F	Black
5	4	62	F	Gray
6	5	28	M	Red
7	6	45	M	Green
8	7	88	F	Brown
9	8	61	M	White
10	9	95	M	Black
11	10	27	M	White
12	11	39	F	Green
13	12	43	M	Brown
14	13	55	F	Black
15	14	59	F	White

RAW DATA

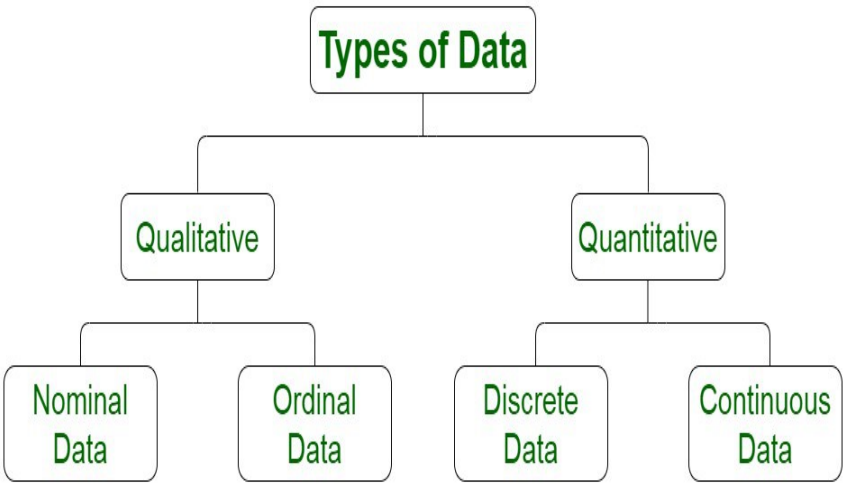


Descriptive Statistics

Data:

- A data set is a collection of related information or Records.
- Data is the most important part of Machine Learning, Artificial Intelligence, and Data Analytics, without data it is impossible to train any model and all modern research and automation will go in worthless.
- Data can be any value, text, sound, or picture i.e., structured or unstructured data (raw data).
- Each row of a data set is called a **record**. Each data set also has multiple **attributes**, each of which gives information on a specific characteristic.

- Attributes can also be termed as feature, variable, dimension or field.



Qualitative Data:

- Qualitative Data can't be measured or counted in the **form of numbers** i.e., sorted by category, not by number. It is also known as **Categorical Data**.
- Qualitative Data may consist of audio, images, symbols, or text.

  Examples: Gender: Male, Female. Product: Good or Bad.
- Qualitative data tells about the perception of people. Which helps market researchers understand the customers' tastes and then design their ideas and strategies accordingly?
- The Qualitative data is further classified into two parts:

  1. Nominal Data

  2. Ordinal Data

1. Nominal Data:

- Nominal Data is used to label variables without any order or quantitative value.
- Numerical task such as Arithmetic operations can't be performed on Nominal data.
- Nominal Data don't have any meaningful order; they are distributed to distinct categories.

Examples:

1. The color of hair can be considered nominal data, as one color can't be compared with another color.

2. Marital status (Single, Married)
3. Nationality (Indian, German, American)
4. Gender (Male, Female)

Ordinal data:

- Ordinal data have natural ordering where a number is present in some kind of order by their position on the scale.

  Examples:

  1. Feedback System on a scale 1 to 10.
  2. Assigning Ranks 1, 2, 3.
  3. Grading in Exam: A, B, C, D.
- Ordinal data is used for observation like customer satisfaction, happiness, etc.
- Since ordering is possible in case of ordinal data, median, and quartiles can be identified Mean can still not be calculated.

Quantitative data

- Quantitative data relates to information about the quantity of an object.
- Quantitative data can be expressed in numerical values, which make it countable and include statistical data analysis.
- Quantitative data also known as a **Numerical data**.
- Quantitative data can be used for statistical manipulation and represented on a wide variety of graphs and charts such as bar graphs, histograms, scatter plots, boxplot, pie charts, etc.

Examples:

- Height or weight of a person: 175cm
- Room Temperature: 29 centigrade
- Marks: 59, 80, 60…
- Time: 2PM, 6AM

The Quantitative data is further classified into two parts:

1. Discrete Data
2. Continuous Data

1. Discrete Data

- Discrete means **distinct or separate**. The discrete data are countable and have finite values, their **subdivision is not possible** i.e., can't be broken into decimal or fraction values.
- The discrete data contain the values that fall under **integers or whole numbers.**
- Represented mainly by a bar graph, scatter plot, etc.

Examples:

- Total no. of students present in a class.
- Numbers of employees in a company.
- Days in a week.

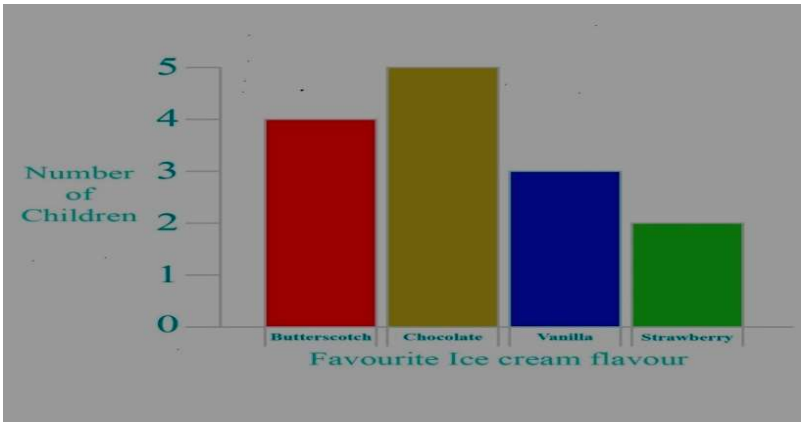


Figure: Bar Plot

2. Continuous data

- Continuous data is data that can take any value within a **range** such as weight, length, temperature, speed, etc.

- Continuous data can be in the form of **decimal or fractional numbers**.
- Continuous data can be measured on an infinite scale; it can take any value between two numbers, no matter how small.
- Represented mainly by a Line plot, Histogram plot**,** etc.

Examples:

- Height of a person
- Speed of a vehicle
- Time-taken to finish the work

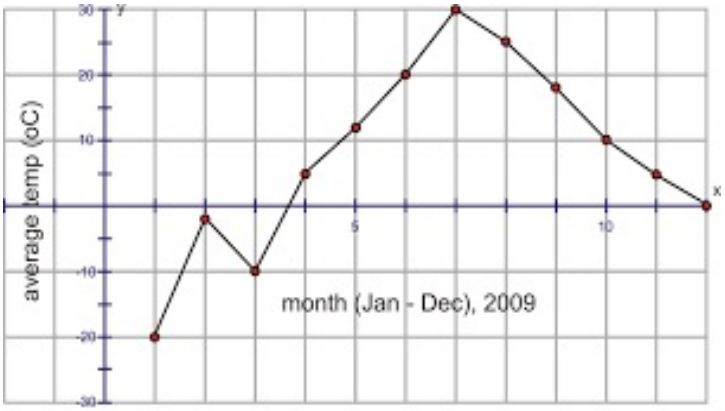


Figure: Line Plot

Interval data

- Interval data is numeric data for which not only the order is known, but the exact difference between values is also known.
- The distance between the two points is equal i.e., equal interval between adjacent values.
- Example: Celsius temperature, IQ Score, Income Ranges.
- The problem with interval values data is that they don't have a "true zero". That means in regards to our example, that there is no such thing as no temperature. With interval data, we can add and subtract, but we cannot multiply, divide or calculate ratios. Because there is no true zero, a lot of descriptive and inferential statistics can't be applied.

Ratio Data:

- Ratio values are also ordered units that have the same difference. Ratio values are the same as interval values, with the difference that they do have an absolute zero. Good examples are height, weight, length etc.

- If there is a true meaning for 0, then we can call it ratio data type. Ratio Data variables can be added, subtracted, multiplied, or divided.
- Ratio variables can be discrete (i.e. expressed in finite, countable units) or continuous (potentially taking on infinite values). Here are some examples of ratio data:
    - Weight in grams (continuous)
    - Number of employees at a company (discrete)
    - Speed in miles per hour (continuous)
    - Age in years (continuous)
    - Income in dollars (continuous)

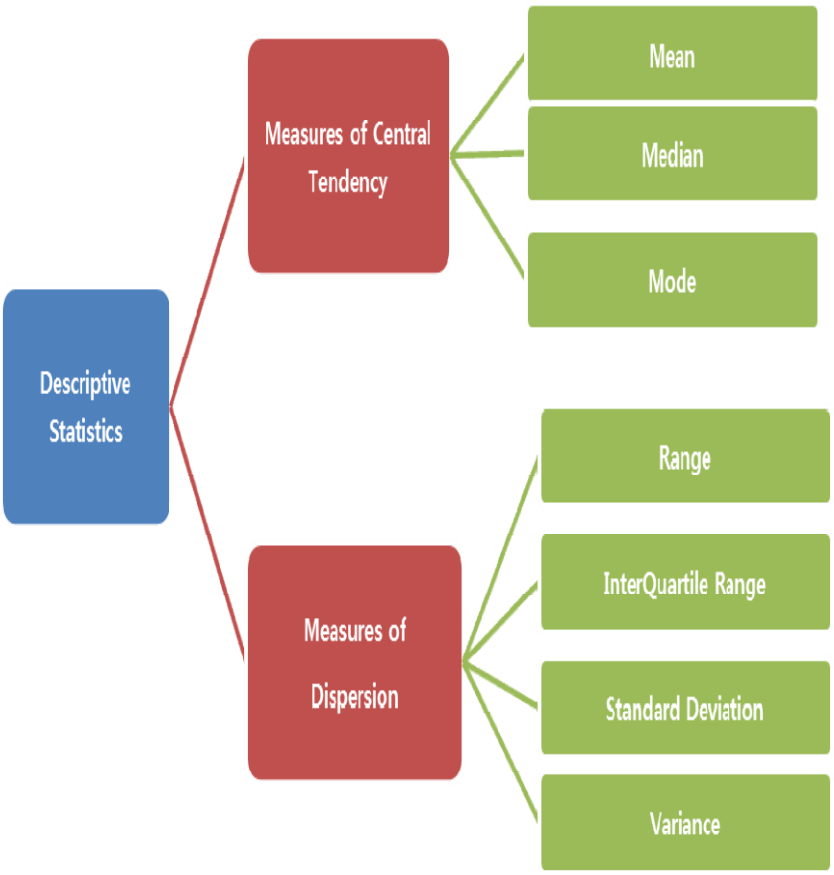
Types of data on the basis of measurement

Scale	True Zero	Equal Intervals	Order	Category	Example
Nominal	No	No	No	Yes	Marital Status, Sex, Gender, Ethnicity
Ordinal	No	No	Yes	Yes	Student Letter Grade, NFL Team Rankings
Interval	No	Yes	Yes	Yes	Temperature in Fahrenheit, SAT Scores, IQ, Year
Ratio	Yes	Yes	Yes	Yes	Age, Height, Weight

NOTE: In general, nominal and ordinal attributes are discrete. On the other hand, interval and ratio attributes are continuous, barring a few exceptions, e.g. 'count' attributes.

Descriptive statistics which can be obtained using interval data include:

1. **Central tendency**: Mode, median, and mean
2. **Dispersion**: Range, Standard deviation and Variance



Central tendency:

➢ Central tendency is defined as "the statistical measure that identifies a single value as representative of an entire distribution.

➢ Central tendency is a one-number summary of the data that typically describes the center of the data. This one-number summary is of three types i.e., measures of central tendency.

- Mean
- Median
- Mode

1. Mean:

Mean is the arithmetic average of a data set. This is found by adding the numbers in a data set and dividing by the number of observations in the data set.

- The Mean

$$\mu_x = \sum_{i=1}^{N} \frac{x_i}{N} = \frac{x_1 + x_2 + \ldots + x_N}{N}$$

Where,

∑-- represents the summation
x -- represents observations
N -- represents the no. of observations.

- It is also known as the **arithmetic mean**, and it is the most common measure of central tendency.
- Arrangement or order of the numbers does not affect calculation for mean.

2. Median:

- The median is the middle number in a data set when the numbers are listed in either ascending or descending order.
- If the total number of observations (n) is an **odd number**, then the formula is given below:

$$Median = \left(\frac{n+1}{2}\right)^{th} observation$$

- If the total number of the observations (n) is an **even number**, then the formula is given below:

$$Median = \frac{\left(\frac{n}{2}\right)^{th} observation + \left(\frac{n}{2}+1\right)^{th} observation}{2}$$

3. Mode:

- The mode is the value that occurs the most often in a data set.
- Mode is the most frequently occurring sample.

Problem 1)

Find the mean, median, mode, and range for the following list of values:
13, 18, 13, 14, 13, 16, 14, 21, 13.

Solution:

Data given : 13, 18, 13, 14, 13, 16, 14, 21, 13

The **mean** is the average.

Mean = {13 + 18 + 13 + 14 + 13 + 16 + 14 + 21 + 13} / {9} = **15**

The **median** is the middle value, so **rewrite** the list in **ascending order** as given below:

13, 13, 13, 13, 14, 14, 16, 18, 21

There are nine numbers in the list, so the middle one will be

{9 + 1} / {2} = {10} / {2} = 5 = 5th number

Hence, the median is **14**.

The **mode** is the number that is repeated more often than any other,

So **13** is the mode.

Difference between Mean and Median	
Mean	**Median**
The average arithmetic of a given set of numbers is called Mean.	The method of separating the higher sample with the lower value, usually from a probability distribution is termed as the median
The application for the mean is for normal distributions	The primary application for the median is skewed distributions.
There are a lot of external factors that limit the use of Mean.	It is much more robust and reliable for measuring the data for uneven data.
Mean can found by calculated by adding all the values and dividing the total by the number of values.	Median can be found by listing all the numbers available in the set in arranging the order and then finding the number in the centre of the distribution.
Mean is considered as an arithmetic average.	Median is considered as a positional average.
It is highly sensitive to outlier data	It is not much sensitive to the outlier data.
It defines the central value of the data set.	It defines the centre of gravity of the midpoint of the data set.

Understanding Data Distribution & measures central tendency:

Let's consider a simple dataset to understand in better way.

Dataset1 = [1,2,2,3,3,3,4,4,4,4,5,5,5,5,5,6,6,6,6,6,6,7,7,7,7,7,7,7,8,8,8,8,8,8,8,8,9,9,9,9,9,9,9,10,10,10,10,10,10,11,11,11,11,11,12,12,12,12,13,13,13,14,14,15].

Example 1:

```
import warnings
warnings.filterwarnings('ignore')
import numpy as np
from scipy import stats
Dataset1 = [1,2,2,3,3,3,4,4,4,4,5,5,5,5,5,6,6,6,6,6,6,7,7,7,7,7,7,7,8,8,8,8,8,8,8,8,9,9,9,9,9,9,9,10,10,10,10,10,10,11,11,11,11,11,12,12,12,12,13,13,13,14,14,15]
```

x1=np.mean(Dataset1)
print(" The Average of Dataset is:",x1)
x2=np.median(Dataset1)
print(" The middle value of Dataset is:",x2)
x3 = stats.mode(Dataset1)
print(" Most frequently used value of Dataset is ",x3)

Let's understand with a plot:

import numpy as np

import matplotlib.pyplot as plt

Dataset1=[1,2,2,3,3,3,4,4,4,4,5,5,5,5,5,6,6,6,6,6,6,7,7,7,7,7,7,7,8,8,8,8,8,8,8,8,9,9,9,9,9,9,9,10,

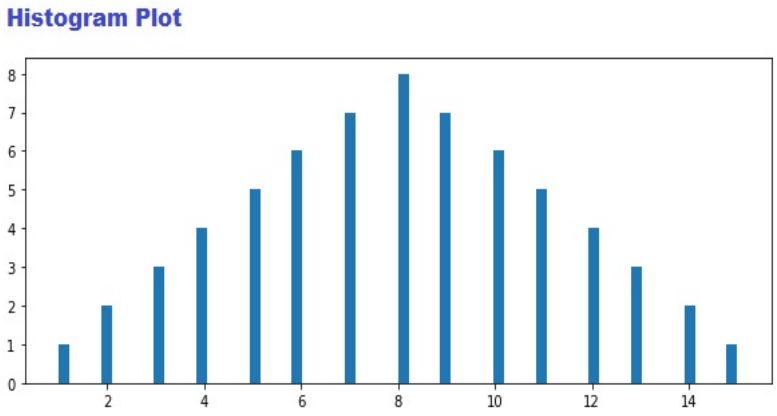
10,10,10,10,10,11,11,11,11,11,12,12,12,12,13,13,13,14,14,15]

y=Dataset1

plt.figure(figsize=(10, 4))

plt1 = plt.hist(Dataset1,bins=64)

plt.show()



Observation 1

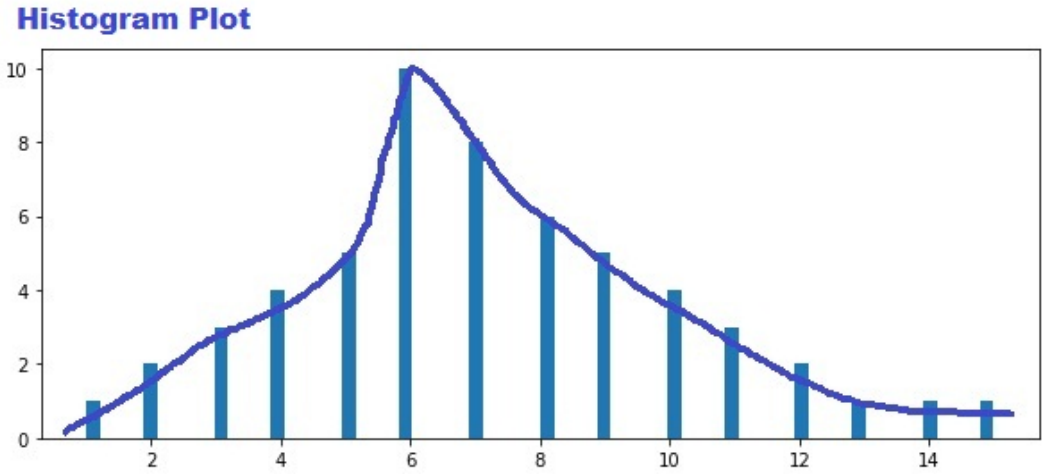
- If mean, mode and median are exactly the same then the data is distributed **symmetrically** i.e., **normal distribution**.
- In a normal distribution, data is symmetrically distributed with no skew. Most values cluster around a central region.

Example 2:

```
import warnings
warnings.filterwarnings('ignore')
import numpy as np
from scipy import stats
Dataset1 =
[1,2,2,3,3,3,4,4,4,4,5,5,5,5,5,6,6,6,6,6,6,6,6,6,6,7,7,7,7,7,7,7,7,8,8,8,8,8,8,9,9,9,9,9,10,10,10,10,
11,11,11,12,12,13,14,15]
x1=np.mean(Dataset1)
print(" The Average of Dataset is:",x1)
x2=np.median(Dataset1)
print(" The middle value of Dataset is:",x2)
x3 = stats.mode(Dataset1)
print(" Most frequently used value of Dataset is ",x3)
```

Understanding with Plot

```
import numpy as np
import matplotlib.pyplot as plt
Dataset1 =
[1,2,2,3,3,3,4,4,4,4,5,5,5,5,5,6,6,6,6,6,6,6,6,6,6,7,7,7,7,7,7,7,7,8,8,8,8,8,8,9,9,9,9,9,10,10,10,10,
11,11,11,12,12,13,14,15]
y=Dataset1
plt.figure(figsize=(10, 4))
plt1 = plt.hist(Dataset1,bins=64)
plt.show()
```



Observation 2

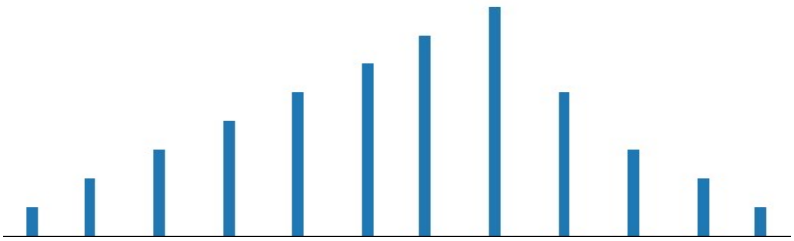
- If **mode < median < mean** then positively skewed distribution,
- In a positively skewed distribution, there's a cluster of lower scores and a spread out tail on the right.
- It is Observed in histogram, distribution is skewed to the right, and the central tendency of dataset is on the lower end of possible scores.

Example 3

```
import warnings
warnings.filterwarnings('ignore')
import numpy as np
from scipy import stats
Dataset1 =
[1,2,2,3,3,3,4,4,4,4,5,5,5,5,5,6,6,6,6,6,6,7,7,7,7,7,7,7,8,8,8,8,8,8,8,8,8,9,9,9,9,9,10,10,10,11,11,12
]
x1=np.mean(Dataset1)
print(" The Average of Dataset is:",x1)
x2=np.median(Dataset1)
print(" The middle value of Dataset is:",x2)
x3 = stats.mode(Dataset1)
print(" Most frequently used value of Dataset is ",x3)
```

Plotting:

```
import numpy as np
import matplotlib.pyplot as plt
Dataset1 =
[1,2,2,3,3,3,4,4,4,4,5,5,5,5,5,6,6,6,6,6,6,7,7,7,7,7,7,7,8,8,8,8,8,8,8,8,9,9,9,9,9,10,10,10,11,11,12
]
y=Dataset1
plt.figure(figsize=(10, 4))
plt1 = plt.hist(Dataset1,bins=64)
plt.show()
```



Observation 3

- In a negatively skewed distribution, **mean < median < mode**.
- In a negatively skewed distribution, there's a cluster of higher scores and a spread out tail on the left.
- In this histogram, distribution is skewed to the left, and the central tendency of dataset is towards the higher end of possible scores.

When should you use the mean, median or mode?

- For **normally distributed data**, all **three measures** of central tendency will give you the

same answer so they can all be used but **mean** is the most widely used for normal distributed data.

- In **skewed distributions**, the **median** is the best measure because it is unaffected by extreme outliers or non-symmetric distributions of scores. The mean and mode can vary in skewed distributions.
- The **mode** can be of used for any level measurement, but it's most meaningful for nominal and ordinal levels.
- The **median** can only be used on data that can be ordered – that is, from ordinal, interval and ratio levels of measurement.
- The **mean** can only be used on interval and ratio levels of measurement because it requires equal spacing between adjacent values or scores in the scale.

Levels of measurement	Examples	Measure of central tendency
Nominal	•Gender: Male , Female •Nationality: Indian	•Mode
Ordinal	•Assigning Ranks •IQ Score	•Mode •Median
Interval and ratio	•Reaction time •Test score •Temperature	•Mode •Median •Mean