Project Report

on

Title: Presentation & Analysis of data

Prepared by

Raj Shekhar Karmaker - 011193149

Course: Probability & Statistics (MATH 2205)

Section: A

Table of Contents

A report representation

An overview on methodology and analysis of data

Abstract

This report presents the results of a statistical analysis conducted on a dataset of customer satisfaction survey responses and other secondary explores. The survey was conducted by a retail company in order to better understand their customers' preferences and experiences. The analysis included descriptive statistics, correlation analysis, and regression analysis. The findings suggest that there is a strong positive correlation between customer satisfaction and customer loyalty, as well as a significant relationship between customer satisfaction and certain demographic variables such as age and income. The regression analysis further indicates that customer satisfaction is a significant predictor of future purchase intentions. These findings can help the retail company make data-driven decisions in order to improve customer satisfaction and loyalty, ultimately leading to increased sales and revenue.

Introduction

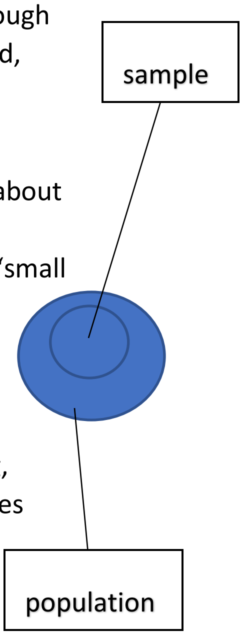
Introduction, Definition, objectives

# What is Statistics?

Statistics is the discipline that concerns the collection, organization, analysis, interpretation, and presentation of data. Data are individual facts or items of information, may be qualitative or quantitative.

# Primary & Secondary Data: Primary data are the original data derived from your research endeavors. Secondary data are data derived from your primary data. Primary data is information collected through original or first-hand research. For example, surveys and focus group discussions. On the other hand, secondary data is information which has been collected in the past by someone else. For example, researching the internet, newspaper articles and company reports.

# Population & Sample: Population A population consists of all the items or individuals or subjects about which you want to draw a conclusion. So, the population is the "large group" in which you are interested. Sample A sample is the portion of a population selected for analysis. The sample is the "small group" for whom we have (or plan to have) data, often randomly selected.



# Sample and Parameter: Parameter is a numerical measure that describes a characteristic of a population. Statistic is a numerical measure that describes a characteristic of a sample

# BRANCHES OF STATISTICS Descriptive Statistics: The branch of statistics that focuses on collecting, summarizing, and presenting a set of data. Inferential Statistics: The branch of statistics that analyzes sample data to draw conclusions about a population.

# Variable: A characteristic of an individual that will be analyzed using statistics Categorical (qualitative) variables have values that can only be placed into categories, such as "yes" and "no"; major; architectural style; etc. Numerical (quantitative) variables have values that represent quantities.

- Discrete variables arise from a counting process Examples: Number of printing errors per page on a book. Number of customers arriving at a restaurant

- Continuous variables arise from a measuring process Examples: Height of a person, Weight of a person, Time a customer waits in a bank queue.

Unit-1:

Data summarization & presentation

Data summarization is the first step in statistics, it is aimed at extracting useful information. Summary statistics are used to summarize a set of observations, to communicate the largest amount of information as simply as possible. Data can be summarized numerically as a table (tabular summarization), or visually as a graph (data visualization).

Data presentation refers to the process of organizing and displaying data in a way that is easy to understand and interpret. The goal of data presentation is to effectively communicate the key insights and findings from the data analysis to the intended audience.

There are several methods of data presentation, including:

Tables: Tables are a common way to present data in a structured format, particularly when dealing with numerical data. Tables can be used to display summary statistics, such as means, medians, and standard deviations, as well as individual data points.

Graphs and charts: Graphs and charts are visual representations of data that can help to convey complex information in a clear and concise manner. Common types of graphs and charts include bar charts, line graphs, scatter plots, and pie charts.

Infographics: Infographics are a combination of text, images, and data visualizations that are used to present complex information in a more engaging and memorable way. Infographics are particularly useful when presenting data to a non-technical audience.

Dashboards: Dashboards are interactive displays of data that allow users to quickly and easily explore and analyze data. Dashboards can be customized to display relevant information in real-time, making them particularly useful for monitoring key performance indicators (KPIs) and other metrics.

When presenting data, it is important to consider the intended audience and to select a method of presentation that is appropriate for their level of technical expertise and familiarity with the subject matter. The goal is to ensure that the data is presented in a way that is easy to understand and that effectively communicates the key insights and findings.

We will be learning across the following data presentations:

- Pie chart
- Bar chart
- Histogram
- Cumulative Frequency Distribution
- Frequency Distribution
- Stem & leaf diagram

Pie chart

A pie chart is a circular statistical graphic, which is divided into slices to illustrate numerical proportion. Or A Pie Chart is a type of graph that displays data in a circular graph. The pieces of the graph are proportional to the fraction of the whole in each category. In other words, each slice of the pie is relative to the size of that category in the group as a whole. The entire “pie” represents 100 percent of a whole, while the pie “slices” represent portions of the whole.

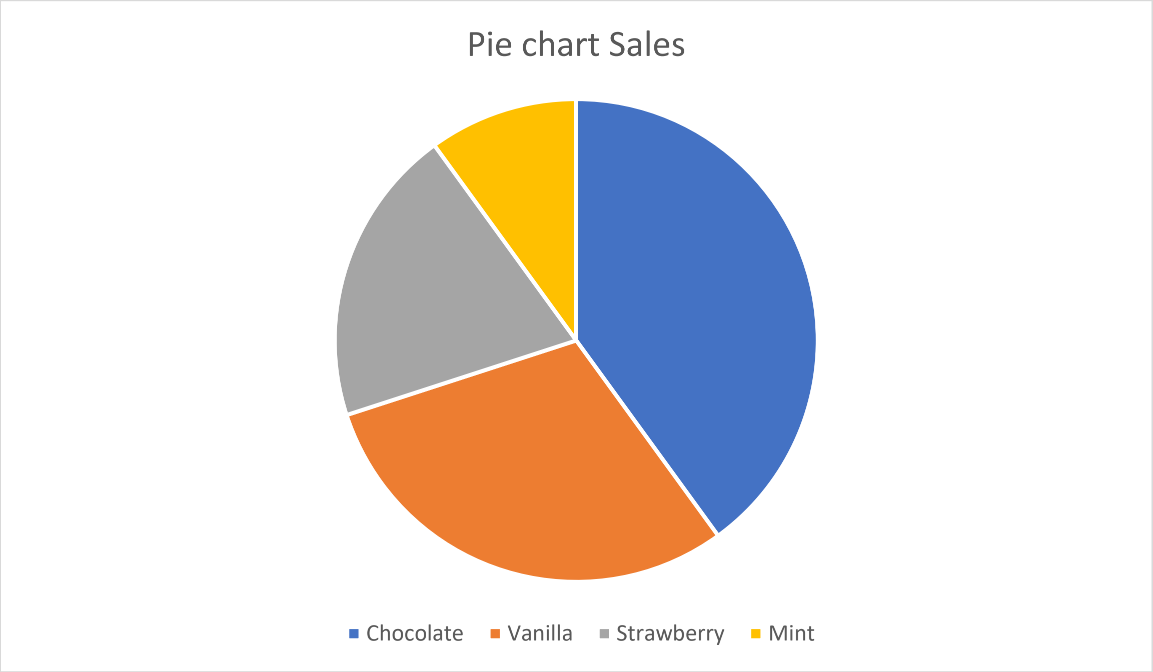
Suppose we want to create a pie chart showing the distribution of ice cream flavors sold at a shop. We might have the following data:

Chocolate: 40%

Vanilla: 30%

Strawberry: 20%

Mint: 10%



The resulting pie chart would show each flavor as a slice of the pie, with the size of each slice proportional to the percentage of sales for that flavor.

Bar chart

A bar chart or bar graph is a chart or graph that presents categorical data with rectangular bars with heights or lengths proportional to the values that they represent. The bars can be plotted vertically or horizontally. A vertical bar chart is sometimes called a column chart.

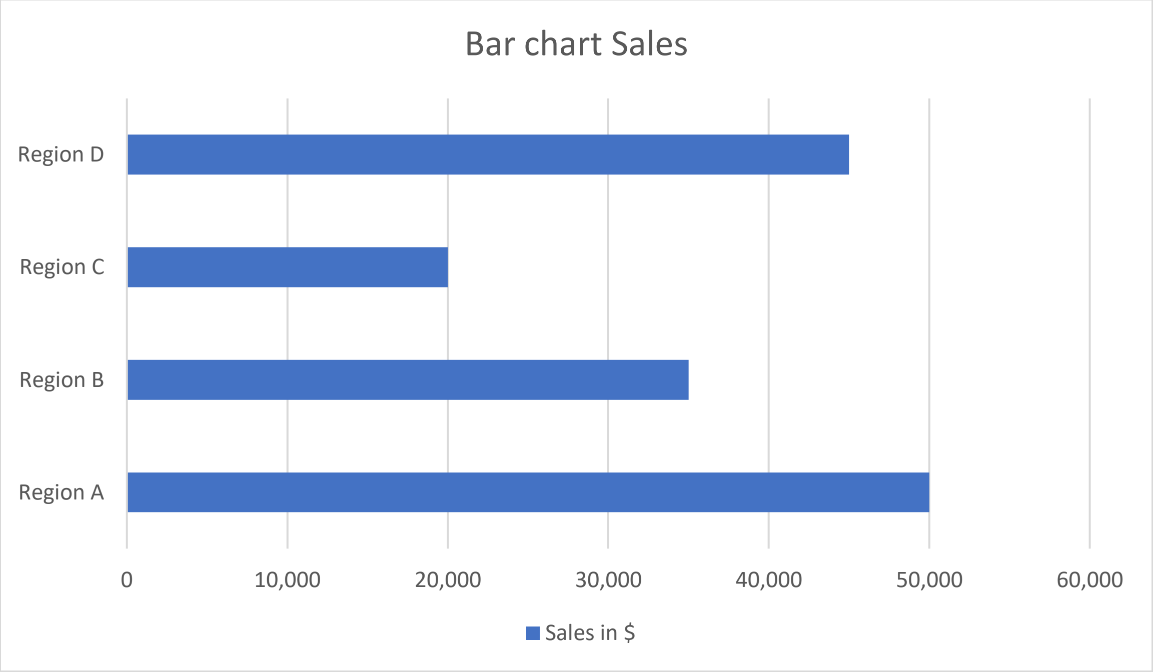
Suppose we want to create a bar chart showing the sales figures for a particular product in different regions. We might have the following data:

Region A: $50,000

Region B: $35,000

Region C: $20,000

Region D: $45,000



The resulting bar chart would show a vertical bar for each region, with the height of each bar proportional to the sales figure for that region. The regions would be listed on the horizontal axis, with the sales figures listed on the vertical axis.

Histogram

A histogram is a graphical presentation of data using rectangular bars of different heights. In a histogram, there is no space between the rectangular bars. A two-dimensional graphical representation of a continuous frequency distribution is called a histogram. In histogram, the bars are placed continuously side by side with no gap between adjacent bars. That is, in histogram rectangles are erected on the class intervals of the distribution. The areas of rectangle are proportional to the frequencies.

Suppose we want to create a histogram showing the distribution of ages for a sample of 50 people. We might have the following data:

20, 25, 30, 35, 28, 40, 42, 29, 26, 33

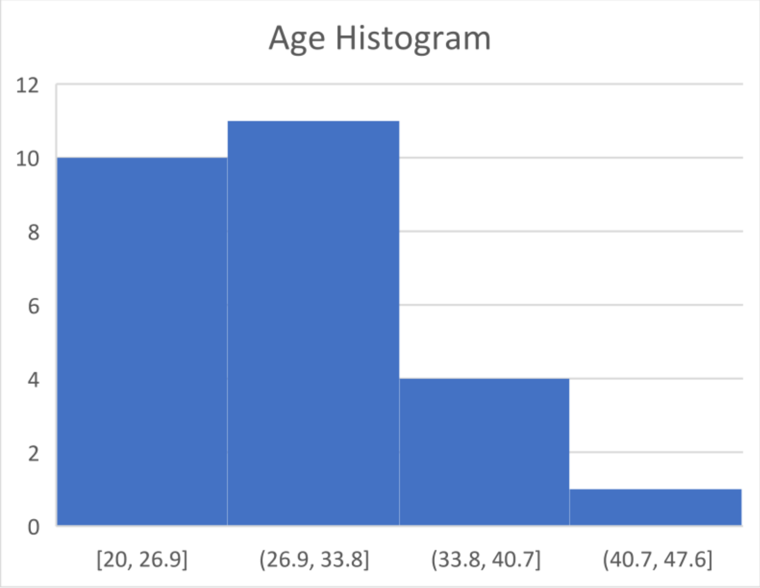
22, 27, 31, 38, 39, 24, 23, 36, 32, 27

28, 43, 41, 34, 27, 22, 25, 31, 26, 30

35, 29, 26, 33, 37, 28, 31, 27, 24, 39

28, 32, 26, 30, 35, 37, 26, 24, 31, 29

To create a histogram from this data, you would group the ages into intervals, known as bins. You could choose a bin size of 5, and group the ages as follows:



The resulting histogram would show the distribution of ages in the sample and provide insights into the age range and frequency of the sample.

Bin 1: 20-26

Bin 2: 26-30

Bin 3: 30-34

Bin 4: 34-39

Bin 5: 40-44

Frequency distribution

A frequency distribution is a summary of data that shows the number of times a particular value or range of values occurs in a dataset. It is a way to organize and analyze data by counting the number of times each value appears.

A frequency distribution can be represented in a table or a graph. In a frequency table, the values are listed in one column, and the frequency or count of each value is listed in another column. The frequency can be expressed as either an absolute frequency (the actual number of occurrences) or a relative frequency (the proportion or percentage of occurrences).

For example, let's say we have a dataset of exam scores for a class of 30 students. The scores range from 60 to 100. We can create a frequency distribution table for these scores by grouping them into intervals (e.g., 60-69, 70-79, 80-89, 90-100) and counting the number of scores in each interval.

Here's an example of what the frequency distribution table might look like:

Score Range	Frequency
60-69	4
70-79	7
80-89	12
90-100	7

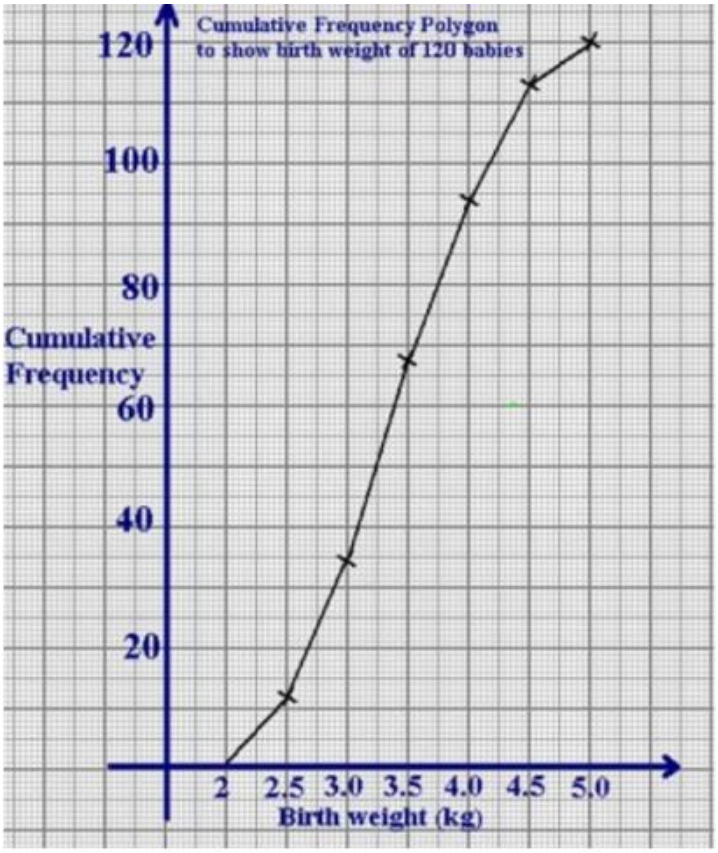
From this table, we can see that the most common score range is 80-89, which occurred 12 times in the dataset. We can also see that the total number of scores is 30, which is the sum of the frequencies in each row.

A frequency distribution graph can also be created to visualize the data. A histogram is a common type of frequency distribution graph that displays the frequency of each value or range of values as a bar. The height of each bar represents the frequency of the corresponding value or range of values.

Cumulative Frequency Distribution

A curve that represents the cumulative frequency distribution of grouped data on a graph is called a Cumulative Frequency Curve or an Ogive. Representing cumulative frequency data on a graph is the most efficient way to understand the data and derive results.

Birth Weight (kg)	2.0-2.5	2.5-3.0	3.0-3.5	3.5-4.0	4.0-4.5	4.5-5.0
Frequency	12	22	33	27	18	8
Cumulative Frequency	12	34	67	94	112	120



Stem-and-Leaf Diagram

A stem and leaf diagram is a method of organizing numerical data based on the place value of the numbers. In a stem and leaf Plot each data value is split into a "stem" (the first digit or digits) and a "leaf" (usually the last digit). Like in this example:

Female (leaves)			Stem	Male (leaves)		
		7 6	0	1 4 5		
	9 8	8 4 2	1	0 2 3	4 4 8	
	6 6 5	5 5 4	2	2 7 7	9	
6	3 3 3	2 1 0	3	0 0 0	3 6 8	
			4	0		

Key: 3 |1|4 represents

13 Female

14 Male

Unit-2:

Measures of location/Central Tendency

- Mean
- Median
- Mode
- Arithmetic Mean, Geometric Mean, Harmonic Mean

Decile, Percentile

Mean

The mean is the sum of the observations divided by the number of observations. It is interpreted as the balance point of the distribution.

In Words

The formula

$$\bar{\mathbf{x}} = \Sigma x/n$$

is short for “sum the values on the variable x and divide by the sample size.”

The mean is a statistical measure of central tendency that represents the average value of a set of data. It is also called the arithmetic mean, and it is calculated by adding up all the values in a dataset and dividing by the number of values.

The formula for calculating the mean is:

mean = (sum of all values) / (number of values)

For example, let's say we have a dataset of five numbers: 2, 4, 6, 8, and 10. To find the mean of this dataset, we add up all the values and divide by the total number of values:

mean = (2 + 4 + 6 + 8 + 10) / 5

mean = 30 / 5

mean = 6

So the mean of this dataset is 6. This means that, on average, the values in the dataset are around 6.

The mean is a useful measure of central tendency because it takes into account all the values in a dataset and is not affected by extreme values or outliers as much as other measures such as the median or mode. However, it is important to note that the mean can be influenced by skewed data or extreme values, and it may not always be the best measure of central tendency depending on the distribution of the data.

We shall study 3 types of mean methodologies:

- Arithmetic Mean
- Geometric Mean
- Harmonic Mean

Arithmetic Mean (A.M)

The arithmetic mean is a statistical measure of central tendency that represents the typical or average value of a set of data. It is also simply called the mean.

The arithmetic mean is calculated by adding up all the values in a dataset and then dividing by the number of values. It is the sum of the values divided by the count of the values. The formula for calculating the arithmetic mean is:

mean = (sum of values) / (number of values)

For example, if we have a dataset of 5 numbers: 2, 4, 6, 8, and 10, we can calculate the arithmetic mean by adding up all the values and dividing by the total number of values:

mean = (2 + 4 + 6 + 8 + 10) / 5

mean = 30 / 5

mean = 6

So the arithmetic mean of this dataset is 6.

The arithmetic mean is a useful measure of central tendency because it takes into account all the values in a dataset and is easy to understand and calculate. However, it can be sensitive to extreme values or outliers in the data, which can skew the value of the mean. In some cases, other measures of central tendency, such as the median or mode, may be more appropriate to use.

Geometric Mean (G.M)

The geometric mean is a measure of central tendency that is calculated by taking the nth root of the product of n positive numbers. It is a type of average that is commonly used to calculate growth rates, ratios, and other exponential changes.

The formula for calculating the geometric mean is:

geometric mean = (x1 * x2 * x3 * ... * xn)^(1/n)

where x1, x2, x3, ..., xn are the n positive numbers in the dataset, and n is the number of values in the dataset.

For example, suppose we have a dataset of three numbers: 2, 4, and 8. To find the geometric mean of this dataset, we multiply these numbers together and then take the cube root (since n=3):

geometric mean = (2 * 4 * 8)^(1/3)

geometric mean = 32^(1/3)

geometric mean = 3.174

So the geometric mean of this dataset is approximately 3.174.

The geometric mean is useful for calculating average rates of change over time or for data that follows an exponential pattern, such as population growth or investment returns. However, it is less commonly used than the arithmetic mean because it can be sensitive to extreme values and can only be calculated for datasets that contain positive numbers.

Harmonic Mean (H.M)

The harmonic mean is a type of average that is used to calculate the average rate of change, such as speed, when the data involves rates or ratios. It is the reciprocal of the arithmetic mean of the reciprocals of a set of numbers.

The formula for calculating the harmonic mean is:

harmonic mean = n / (1/x1 + 1/x2 + 1/x3 + ... + 1/xn)

where x1, x2, x3, ..., xn are the n positive numbers in the dataset, and n is the number of values in the dataset.

For example, let's say we have a dataset of three numbers that represent the speeds of a car in three segments of a journey: 40 mph, 60 mph, and 80 mph. To find the harmonic mean of these speeds, we calculate the sum of the reciprocals of the speeds, divide by the number of speeds, and then take the reciprocal of the result:

harmonic mean = 3 / (1/40 + 1/60 + 1/80)

harmonic mean = 3 / (0.025 + 0.017 + 0.013)

harmonic mean = 3 / 0.055

harmonic mean = 54.55

So the harmonic mean of the speeds in this dataset is approximately 54.55 mph.

The harmonic mean is useful for calculating the average rate of change, such as the average speed over a journey that involves different speeds or rates. However, it is less commonly used than the arithmetic mean and may not be appropriate for datasets that contain extreme values or outliers.

Note:

- A.M > G.M > H.M
- Hence, or otherwise, A.M = G.M = H.M

x	f	f.x
20	3	60
30	4	120
40	2	80
50	1	50
	Σf = 10	Σfx = 310

$\bar{x} = \Sigma fx / \Sigma f$

$= 310 / 10$

$= 31$

Median

The median is a measure of central tendency that represents the middle value of a dataset. It is the value that separates the higher half of a dataset from the lower half.

To find the median of a dataset, we first need to arrange the values in order from lowest to highest. If the dataset contains an odd number of values, the median is the middle value. If the dataset contains an even number of values, the median is the average of the two middle values.

For example, let's say we have a dataset of seven numbers: 2, 4, 6, 8, 10, 12, and 14. To find the median of this dataset, we first arrange the numbers in order:

2, 4, 6, 8, 10, 12, 14

Since there are seven numbers, which is an odd number, the median is the middle number, which is 8. So the median of this dataset is 8.

Now let's consider a different dataset with eight numbers (even): 1, 3, 5, 7, 9, 11, 13, and 15. To find the median of this dataset, we arrange the numbers in order:

1, 3, 5, 7, 9, 11, 13, 15

Since there are eight numbers, which is an even number, the median is the average of the two middle numbers, which are 7 and 9. So the median of this dataset is:

median = (7 + 9) / 2

median = 8

So the median of this dataset is also 8.

The median is a useful measure of central tendency because it is not influenced by extreme values or outliers as much as the mean, and it provides a better representation of the typical value in the dataset in some cases.

The Median Is Resistant to Outliers

The median is resistant. The mean is not.

Outlier

An outlier is an observation that falls well above or well below the overall bulk of the data.

The mean can be highly influenced by an outlier

In statistics, an outlier is an observation that is significantly different from other observations in a dataset. Outliers can occur due to measurement error, data entry errors, or true variation in the data.

Outliers can have a significant impact on the results of statistical analyses, especially when calculating the mean or variance. Outliers can cause the mean to be biased and not representative of the typical value in the dataset.

It is important to identify outliers in a dataset and determine whether they should be included in the analysis or removed. One way to identify outliers is to use a box plot, which displays the distribution of the data and any values that fall outside of the whiskers of the plot can be considered outliers. Another approach is to use statistical methods, such as the z-score or interquartile range (IQR), to identify values that fall significantly outside of the expected range of values for the dataset.

In some cases, outliers may be valid observations and should be retained in the analysis. For example, in medical studies, outliers may represent rare cases that are important to study. However, in other cases, outliers may be due to errors or anomalies and should be removed from the analysis. The decision to remove outliers should be made carefully and with a clear rationale.

Here is an example of an outlier:

Let's say we have a dataset of exam scores for a class of students. The dataset contains the following scores: 75, 80, 85, 90, 95, 100, and 200. The score of 200 is much higher than the other scores and is an outlier.

If we calculate the mean of this dataset, it would be:

mean = (75 + 80 + 85 + 90 + 95 + 100 + 200) / 7

mean = 114.3

The mean score of 114.3 is significantly higher than most of the scores in the dataset, and this is largely due to the presence of the outlier score of 200.

If we remove the outlier score of 200 and calculate the mean again, it would be:

mean = (75 + 80 + 85 + 90 + 95 + 100) / 6

mean = 88.3

The mean score of 88.3 is a better representation of the typical score in the dataset, without the influence of the outlier.

In this example, the outlier score of 200 is likely due to an error in data entry or measurement, and should be removed from the analysis. However, it is important to note that outliers may also be valid observations that should be retained in the analysis, depending on the context and the goals of the study.

Σ (xi – 5) = a ; where n=25, a=375

Σ (xi – 5) = 375

Σ xi = 375 + (25*5)

= 500

$\bar{x}$ = Σ xi /n

= 500/25

= 20

Note:

When data are very close to each other (clustered) then mean is the appropriate measures of central tendency, i.e, when the data are highly deviated, **mean** is not an appropriate total, here comes **median** and **mode**.

Mode

In statistics, the mode is the value that appears most frequently in a dataset. In other words, it is the value that has the highest frequency or count in the dataset.

The mode can be used to describe the central tendency of a dataset, along with the mean and median. Unlike the mean and median, which are affected by extreme values or outliers in the data, the mode is not influenced by extreme values.

A dataset can have one mode (unimodal), two modes (bimodal), or more than two modes (multimodal). If no value appears more frequently than any other value in the dataset, then the dataset has no mode.

The mode is commonly used in descriptive statistics and can be useful in identifying the most common value or category in a dataset. For example, the mode can be used to describe the most common type of car owned by people in a particular city, or the most common color of eyes in a group of people.

Suppose we have a dataset of the ages of people in a group:

[23, 26, 19, 20, 23, 25, 28, 23, 24, 20, 21, 23, 22, 20, 23]

To find the mode of this dataset, we need to identify the value that appears most frequently. In this case, the value 23 appears 4 times, which is more than any other value in the dataset. Therefore, the mode of this dataset is 23.

So, the mode of this dataset is 23, which indicates that the age of 23 is the most common age in this group.

- Unimodal class
- Bimodal class
- Multimodal class

Unimodal class

In statistics, a unimodal class refers to a frequency distribution in which there is a single value or range of values that occurs most frequently.

A frequency distribution can be represented graphically by a histogram, in which the range of values in the dataset is divided into a set of intervals or classes, and the number of observations falling within each class is shown by the height of a corresponding bar.

If the histogram has a single peak, or mode, it is said to be unimodal. This indicates that the values in the dataset are clustered around a central value or range of values, with relatively fewer observations occurring at higher or lower values.

A unimodal class is often used to describe a normal distribution, which is a common type of distribution that appears in many natural and social phenomena. The normal distribution is characterized by a symmetrical bell-shaped curve, with a single peak at the mean value of the dataset.

Suppose we have a dataset of the heights of a group of people, and we create a histogram to represent the frequency distribution of the heights. The heights are divided into classes of 5 cm, and the frequency of observations falling within each class is shown by the height of a corresponding bar in the histogram.

Height (cm)	Frequency
150-154	4
155-159	8
160-164	12
165-169	16
170-174	14
175-179	10
180-184	6
185-189	2

In this example, we can see that the histogram has a single peak at the class interval 165-169 cm, which has the highest frequency of 16 observations. Therefore, the histogram is unimodal, and the class interval 165-169 cm can be described as the unimodal class.

This indicates that the heights of the people in this group are clustered around the central value of 165-169 cm, with relatively fewer observations occurring at higher or lower heights.

-Has only 1 mode

Bimodal class

In statistics, a bimodal class refers to a frequency distribution in which there are two values or ranges of values that occur with the highest frequency.

A bimodal class occurs when a dataset has two distinct modes, or peaks, in its frequency distribution. This can occur when the data comes from two different populations, or when there are two distinct phenomena that are affecting the data.

A bimodal class can be represented graphically by a histogram, in which the range of values in the dataset is divided into a set of intervals or classes, and the number of observations falling within each class is shown by the height of a corresponding bar. In a bimodal histogram, there are two distinct peaks that represent the two modes.

Bimodal classes are less common than unimodal classes in many datasets, but they can be important to identify because they may indicate that the data is coming from two different sources, or that there are two different patterns or processes at work.

An example of a bimodal class might be a dataset of the ages of people in a country, where there are two distinct age groups that are more common than others, such as young children and retirees.

Suppose we have a dataset of the test scores of a group of students in a class, and we create a histogram to represent the frequency distribution of the scores. The test scores are divided into classes of 10 points, and the frequency of observations falling within each class is shown by the height of a corresponding bar in the histogram.

Test Score	Frequency
50-59	4
60-69	10
70-79	16
80-89	20
90-99	10
100-109	4

In this example, we can see that the histogram has two distinct peaks at the class intervals 80-89 and 70-79, which have the highest frequency of 20 and 16 observations, respectively. Therefore, the histogram is bimodal, and the class intervals 80-89 and 70-79 can be described as the bimodal classes.

This indicates that the test scores of the students in this class are clustered around two central values, one in the range of 70-79 and the other in the range of 80-89. This might indicate that there are two groups of students with different levels of preparation or ability, or that the test had two distinct parts that tested different skills or knowledge.

-Has 2 mode.

Multimodal class

In statistics, a multimodal class refers to a frequency distribution in which there are more than two values or ranges of values that occur with the highest frequency.

A multimodal class occurs when a dataset has multiple distinct modes, or peaks, in its frequency distribution. This can occur when the data comes from multiple populations or there are several underlying processes that are affecting the data.

A multimodal class can be represented graphically by a histogram, in which the range of values in the dataset is divided into a set of intervals or classes, and the number of observations falling within each class is shown by the height of a corresponding bar. In a multimodal histogram, there are multiple peaks that represent the different modes.

Multimodal classes can be more challenging to analyze than unimodal or bimodal classes, as they can be more difficult to interpret and may require more sophisticated statistical techniques to analyze. However, they can also be more informative, as they may reveal important patterns or trends in the data.

An example of a multimodal class might be a dataset of the salaries of employees in a company, where there are several distinct groups of employees with different salaries, such as entry-level workers, mid-level managers, and senior executives.

Suppose we have a dataset of the annual incomes of a group of people, and we create a histogram to represent the frequency distribution of the incomes. The incomes are divided into classes of $10,000, and the frequency of observations falling within each class is shown by the height of a corresponding bar in the histogram.

Annual Income	Frequency
$20,000-$29,999	12
$30,000-$39,999	22
$40,000-$49,999	32
$50,000-$59,999	38
$60,000-$69,999	30
$70,000-$79,999	26
$80,000-$89,999	18
$90,000-$99,999	12

In this example, we can see that the histogram has multiple peaks at the class intervals of $50,000-$59,999, $40,000-$49,999, and $30,000-$39,999. Therefore, the histogram is multimodal, and these class intervals can be described as the multimodal classes.

This indicates that the annual incomes of the people in this dataset are clustered around multiple central values, and there may be several underlying factors that are influencing their incomes, such as education level, occupation, or location. For example, there may be a group of people with higher incomes who live in urban areas and work in high-paying professions, while another group with lower incomes who live in rural areas and work in lower-paying jobs. Understanding these underlying factors can be important for developing policies and interventions to address income inequality.

-Has more than 2 mode.

Decile

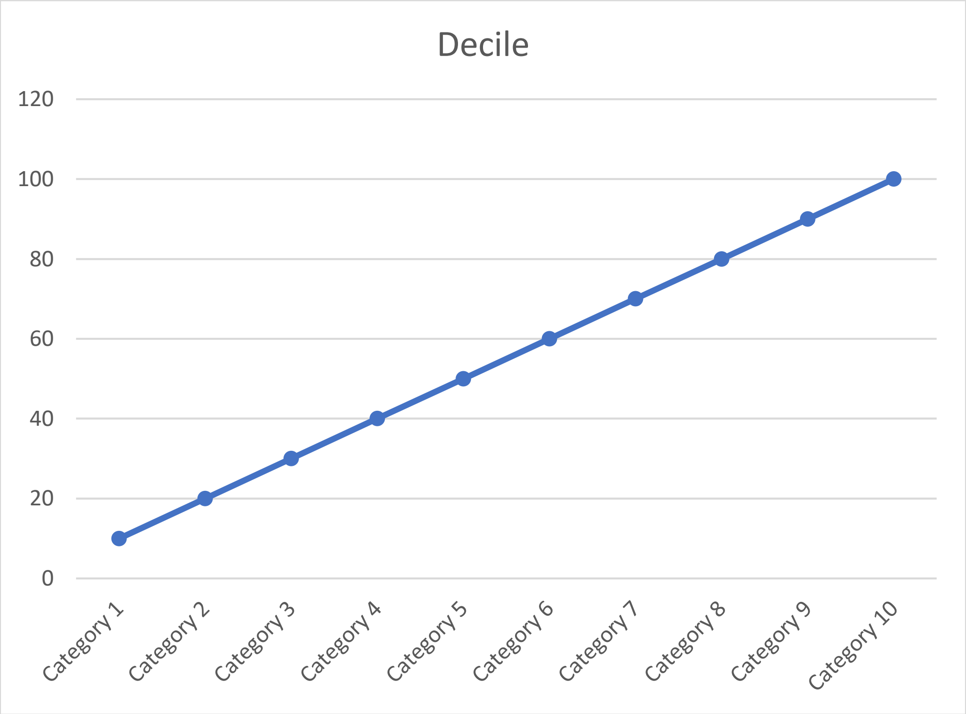
In statistics, a decile is a type of quantile that divides a dataset into ten equal parts. Each decile represents a particular point in the distribution of the data.

To calculate the deciles of a dataset, you first need to order the values from lowest to highest. Then, you divide the ordered dataset into ten equal parts by finding the values that mark the boundaries between each decile.

For example, to find the 3rd decile of a dataset, you would order the values from lowest to highest and then identify the value that marks the point at which 30% of the data lies below it. Similarly, to find the 7th decile, you would identify the value that marks the point at which 70% of the data lies below it.

Deciles can be useful for understanding the distribution of data, as they allow you to identify important points in the data that divide it into equal parts. For example, the 5th decile represents the median, or middle value, of the dataset, while the 1st and 9th deciles represent the values at which 10% and 90% of the data lies below, respectively.

Deciles are often used in conjunction with other measures of central tendency and dispersion, such as the mean and standard deviation, to provide a more complete picture of the distribution of data.



Percentile

In statistics, a percentile is a measure that indicates the value below which a given percentage of observations in a dataset falls. Percentiles are often used to describe the distribution of data and to compare individual observations to the rest of the dataset.

To calculate a percentile, you first need to order the values in the dataset from smallest to largest. Then, you identify the rank of the observation in question within the dataset, as a proportion of the total number of observations. Finally, you multiply this proportion by 100 to get the percentile.

For example, if a student scored in the 80th percentile on a test, this means that they performed better than 80% of the other students who took the test. Similarly, if a person's height is in the 75th percentile for their age and gender, this means that their height is greater than that of 75% of people in their demographic group.

Percentiles can be useful for understanding the distribution of data and identifying outliers or extreme values. The most commonly used percentiles are the quartiles, which divide the dataset into quarters, with the 25th percentile representing the lower quartile (Q1), the 50th percentile representing the median (Q2), and the 75th percentile representing the upper quartile (Q3). Other commonly used percentiles include the deciles (dividing the dataset into tenths) and the 90th and 95th percentiles (indicating values below which 90% and 95% of the data falls, respectively).

Unit-3:

- Measures of dispersion/Variability Range
- Inter Quartile Range (IQR), Mean Deviation
- Standard Deviation
- Variance,
- 5 number summary
- Box & Whisker Plot,
- Outlier (*Covered in Unit-2*)

Measures of dispersion/Variability Range

Measures of dispersion, also known as measures of variability, are statistical tools used to describe how spread out a set of data is. One simple measure of dispersion is the range.

The range is simply the difference between the largest and smallest values in a dataset. To calculate the range, you subtract the smallest value in the dataset from the largest value. For example, if the values in a dataset are 1, 2, 3, 4, and 5, the range is 5 - 1 = 4.

The range is a quick and easy way to describe the spread of a dataset, but it has some limitations. One limitation is that it is sensitive to outliers, or extreme values that are far from the rest of the data. Outliers can make the range appear larger than it actually is, and they can also make the range less useful as a measure of dispersion.

Another limitation of the range is that it only takes into account the two most extreme values in the dataset, and it does not provide any information about the spread of the data between those values. To get a more complete picture of the dispersion of a dataset, other measures of variability such as the variance, standard deviation, or interquartile range may be used.

Quartile

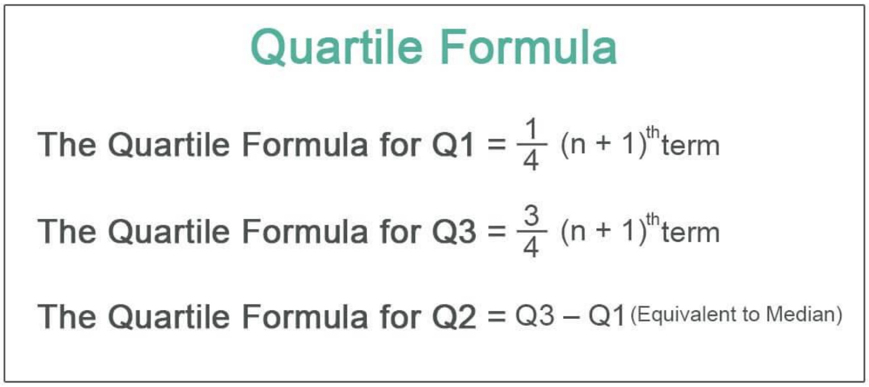
Quartiles are values that divide a dataset into four equal parts, each containing 25% of the data points. There are three quartiles, known as Q1, Q2 (the median), and Q3.

The first quartile, Q1, represents the 25th percentile of the dataset. It is the value below which 25% of the data points fall. Q1 is also known as the lower quartile.

The second quartile, Q2, represents the 50th percentile of the dataset. It is the value below which 50% of the data points fall. Q2 is also known as the median.

The third quartile, Q3, represents the 75th percentile of the dataset. It is the value below which 75% of the data points fall. Q3 is also known as the upper quartile.

Quartiles are often used to analyze the distribution of a dataset and to identify potential outliers. The interquartile range (IQR), which is the difference between the third and first quartiles (Q3-Q1), can also be used to measure the spread of a dataset, with a larger IQR indicating a greater degree of variability.



Quartile Formula

The Quartile Formula for Q1 $= \frac{1}{4}(n + 1)^{th}$term

The Quartile Formula for Q3 $= \frac{3}{4}(n + 1)^{th}$term

The Quartile Formula for Q2 $= Q3 - Q1$ (Equivalent to Median)

Range

Range is a simple measure of dispersion that is defined as the difference between the largest and smallest values in a dataset. To calculate the range, you simply subtract the smallest value in the dataset from the largest value.

Lowest value ↔ Highest value

For example, if a dataset consists of the values 2, 4, 6, 8, and 10, the range is 10 - 2 = 8. This means that the largest value in the dataset is 8 units larger than the smallest value.

The range is a quick and easy way to describe the spread of a dataset, but it has some limitations. One limitation is that it is sensitive to outliers, or extreme values that are far from the rest of the data. Outliers can make the range appear larger than it actually is, and they can also make the range less useful as a measure of dispersion.

Another limitation of the range is that it only takes into account the two most extreme values in the dataset, and it does not provide any information about the spread of the data between those values. To get a more complete picture of the dispersion of a dataset, other measures of variability such as the variance, standard deviation, or interquartile range may be used.

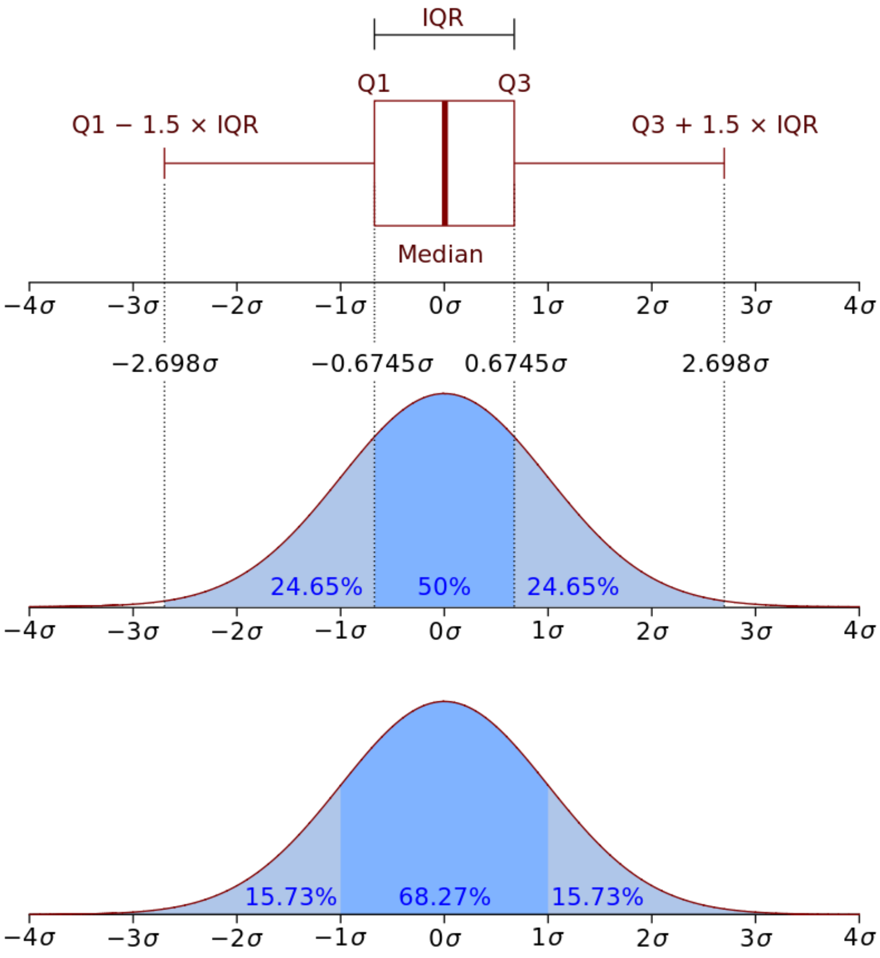
Inter Quartile Range (IQR)

The interquartile range (IQR) is a measure of variability that is based on dividing a dataset into quarters, or quartiles. The IQR represents the range of the middle 50% of the data and is calculated as the difference between the upper quartile (Q3) and the lower quartile (Q1).

To calculate the IQR, you first need to find the median of the dataset. Then, you divide the dataset into two halves: the lower half, consisting of all the values less than or equal to the median, and the upper half, consisting of all the values greater than or equal to the median. Next, you find the median of each half to get the lower quartile (Q1) and the upper quartile (Q3). Finally, you calculate the IQR as the difference between Q3 and Q1.

For example, if a dataset consists of the values 1, 2, 3, 4, 5, 6, 7, 8, 9, and 10, the median is 5.5. The lower half of the data consists of the values 1, 2, 3, 4, and 5, and the upper half consists of the values 6, 7, 8, 9, and 10. The median of the lower half is 3, and the median of the upper half is 8. Therefore, the lower quartile (Q1) is 3 and the upper quartile (Q3) is 8. The IQR is calculated as Q3 - Q1 = 8 - 3 = 5.

The IQR is a useful measure of variability because, unlike the range, it is less sensitive to outliers or extreme values that can distort the distribution of the data. It also provides a measure of the spread of the middle 50% of the data, which can give a better understanding of the central tendency of the dataset.



Deviation

In statistics, deviation refers to the amount by which a data point or a set of data points differs from a central value, such as the mean or median. The deviation is typically calculated as the difference between the observed value and the central value.

For example, if the mean of a dataset is 5 and a data point has a value of 7, then the deviation of that data point from the mean is 7 - 5 = 2. Similarly, if the median of a dataset is 4 and a data point has a value of 6, then the deviation of that data point from the median is 6 - 4 = 2.

The deviation can be either positive or negative, depending on whether the data point is above or below the central value. When calculating the deviation for a set of data points, it is common to use absolute values to ensure that the deviation is always positive.

Deviation is an important concept in statistics because it is used to measure the variability or spread of a dataset. Other measures of variability that use deviations include the mean deviation, variance, and standard deviation.

Mean Deviation

The mean deviation is a measure of variability that describes the average distance between each data point and the mean of the dataset. It is also known as the average deviation or the mean absolute deviation (MAD).

To calculate the mean deviation, you first find the mean of the dataset. Then, for each data point in the dataset, you calculate the absolute difference between the data point and the mean. Finally, you take the average of all of these absolute differences to get the mean deviation.

Mathematically, the formula for calculating the mean deviation is:

mean deviation = (sum of absolute deviations from the mean) / (number of data points)

Or, **mean deviation = $\Sigma(|xi - \bar{x}|) / n$**

Where:

xi = each data point in the dataset

$\bar{x}$ = mean of the dataset

n = number of data points in the dataset

For example, suppose a dataset consists of the values 2, 4, 6, 8, and 10. The mean of the dataset is (2+4+6+8+10)/5 = 6. The absolute deviations from the mean are |2-6| = 4, |4-6| = 2, |6-6| = 0, |8-6| = 2, and |10-6| = 4. The sum of these absolute deviations is 4 + 2 + 0 + 2 + 4 = 12. Therefore, the mean deviation is 12/5 = 2.4.

The mean deviation is a useful measure of variability because it takes into account the distance of each data point from the mean, and it gives an indication of how spread out the data is. However, like the range, it is also sensitive to outliers and extreme values. To overcome this limitation, other measures of variability such as the variance or standard deviation may be used.

Standard deviation

Standard deviation is a commonly used measure of variability or spread in a dataset. It measures the degree of dispersion or variability of a set of data points relative to the mean.

The standard deviation is calculated as the square root of the variance, which is the average of the squared deviations of each data point from the mean. In other words, the standard deviation is the square root of the average squared distance of each data point from the mean.

Mathematically, the formula for calculating the standard deviation is:

standard deviation = sqrt(variance)

where

variance = sum of $(xi - \bar{x})^2$ / (n - 1)

and

$\bar{x}$ = mean of the dataset

xi = each data point in the dataset

n = number of data points in the dataset

The standard deviation has the same unit of measurement as the data points in the dataset. A small standard deviation indicates that the data points are tightly clustered around the mean, while a large standard deviation indicates that the data points are more spread out.

The standard deviation is widely used in statistics and data analysis because it provides a measure of how much the data points deviate from the mean, which can help identify outliers, assess the precision of estimates, and evaluate the significance of differences between groups or variables.

$$\sigma = \sqrt{\frac{\sum(x_i - \mu)^2}{N}}$$

$$\text{Standard Deviation Formula} = \sqrt{\frac{\Sigma_{i\text{-}1}^{n} (x_i - \bar{x})^2}{N - 1}}$$

Variance

In statistics, variance is a measure of how spread out a set of data is. It is calculated by finding the average of the squared differences from the mean of the data.

The formula for variance is:

Variance = (1/n) * Σ(xi - x̄)^2

where n is the number of data points, xi is each individual data point, x̄ is the mean of the data, and Σ represents the sum of all values.

Variance is a useful tool for understanding the variability of a dataset. It tells us how much the individual data points differ from the mean, and can help us make inferences about the overall distribution of the data. A larger variance indicates a wider spread of the data, while a smaller variance indicates a more tightly clustered set of data points.

$$\sigma^{2} = \frac{\sum(xi - \bar{x})^{2}}{N}$$

5 number summary

The five-number summary is a descriptive statistics tool that summarizes the distribution of a dataset using five key values. These values include the minimum, the first quartile (Q1), the median, the third quartile (Q3), and the maximum.

The five-number summary is often used to create box plots, which provide a visual representation of the distribution of the data.

Here's how to calculate the five-number summary:

Minimum: the smallest value in the dataset.

Q1: the value that separates the lowest 25% of the data from the rest of the dataset.

Median: the value that separates the lowest 50% of the data from the highest 50% of the data.

Q3: the value that separates the lowest 75% of the data from the rest of the dataset.

Maximum: the largest value in the dataset.

In summary, the five-number summary provides a quick snapshot of the range, central tendency, and spread of the data in a dataset.

Let's say we have the following dataset of 10 numbers:

{ 3, 7, 8, 12, 15, 16, 19, 22, 25, 30 }

To calculate the five-number summary, we need to find the minimum, Q1, median, Q3, and maximum values.

Minimum: The smallest number in the dataset is 3.

Q1: The median of the lower half of the dataset is (7 + 8) / 2 = 7.5.

Median: The median of the entire dataset is (15 + 16) / 2 = 15.5.

Q3: The median of the upper half of the dataset is (22 + 25) / 2 = 23.5.

Maximum: The largest number in the dataset is 30.

So, the five-number summary for this dataset is:

Minimum: 3

Q1: 7.5

Median: 15.5

Q3: 23.5

Maximum: 30

We can use the five-number summary to create a box plot, which provides a visual representation of the distribution of the data. The box plot would show the minimum and maximum values as whiskers, a box from Q1 to Q3, and a line at the median value.

Skewness

In statistics, skewness is a measure of the asymmetry of a probability distribution. A distribution is said to be skewed if it is not symmetric around its mean. Skewness can be either positive or negative, or even zero, indicating different types and degrees of asymmetry.

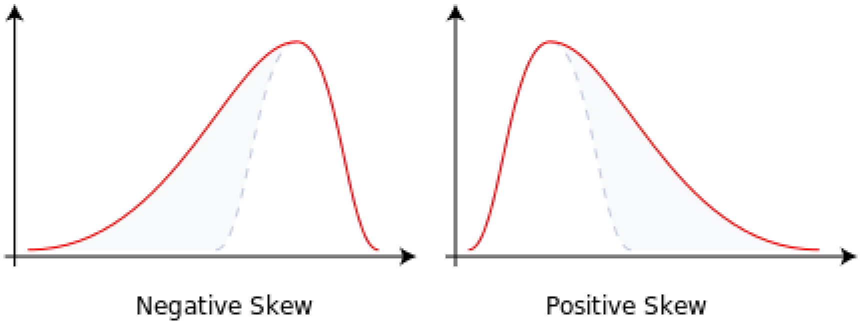
A positively skewed distribution has a long right tail, meaning that the majority of the data is on the left-hand side of the distribution, and there are some extreme values on the right-hand side. In contrast, a negatively skewed distribution has a long left tail, meaning that the majority of the data is on the right-hand side of the distribution, and there are some extreme values on the left-hand side.

The skewness of a distribution can be quantified using a skewness coefficient. One commonly used coefficient is Pearson's moment coefficient of skewness, which is calculated as:

Skewness = (3 * (mean - median)) / standard deviation

A positive value of skewness indicates a positively skewed distribution, while a negative value indicates a negatively skewed distribution. A skewness value of zero indicates a symmetric distribution.

Skewness is important in statistics because it can affect the interpretation of statistical analyses. For example, if a dataset is highly skewed, the mean may not accurately represent the central tendency of the data, and it may be more appropriate to use the median instead.



Mean<Median<Mode

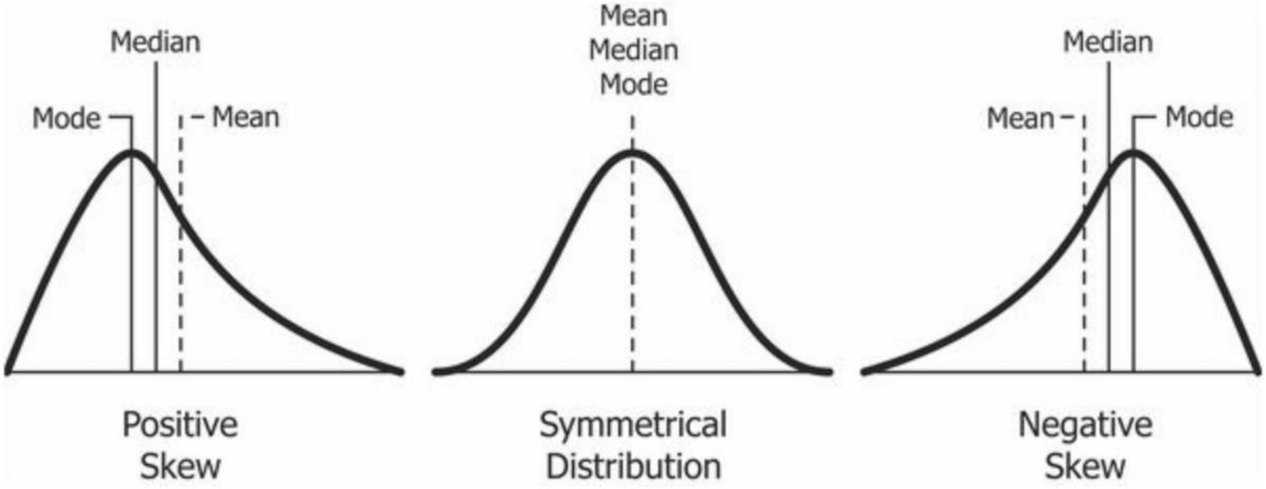
Q3-Q2 < Q2-Q1

Mean>Median>Mode

Q3-Q2 >Q2-Q1

Mean=Median=Mode

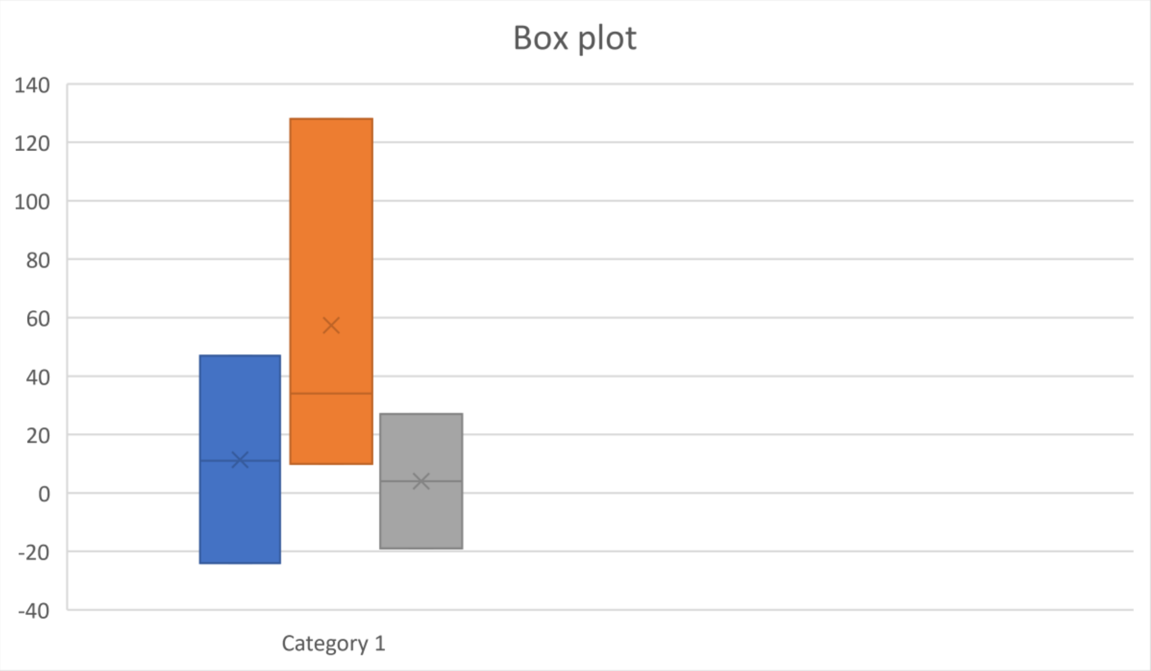
Q3-Q2 =Q2-Q1



Box and Whisker Plot

A box and whisker plot, also known as a box plot, is a graphical representation of a dataset's five-number summary. It is a useful tool for visualizing the distribution, central tendency, and spread of a dataset.

A box plot consists of a rectangular box and two "whiskers" that extend from the box. The box spans the interquartile range (IQR), which is the distance between the first quartile (Q1) and the third quartile (Q3). The whiskers extend to the minimum and maximum values in the dataset, or to a certain distance from the box known as the "fences."



Here's how to create a box plot:

Calculate the five-number summary of the dataset: minimum, Q1, median, Q3, and maximum.

Draw a vertical axis and label it with the minimum and maximum values in the dataset.

Draw a horizontal axis and label it with the variable being measured.

Draw a box that spans from Q1 to Q3.

Draw a vertical line inside the box at the median.

Draw whiskers that extend from the box to the minimum and maximum values in the dataset or to the fences.

Plot any outliers as individual points beyond the whiskers.

Box plots provide a quick and easy way to compare multiple datasets or to visualize the distribution of a single dataset. They can also reveal potential outliers and help identify differences in the spread and shape of the data.

Suppose we have a dataset of 20 numbers representing the weights of 20 apples, in ounces:

{ 2.5, 3.0, 3.5, 4.0, 4.5, 4.5, 5.0, 5.0, 5.5, 6.0, 6.0, 6.5, 7.0, 7.0, 7.5, 8.0, 8.5, 9.0, 9.5, 10.0 }

To create a box plot of this dataset, we can follow the steps I mentioned earlier:

Calculate the five-number summary:

Minimum: 2.5

Q1: 4.5

Median: 6.0

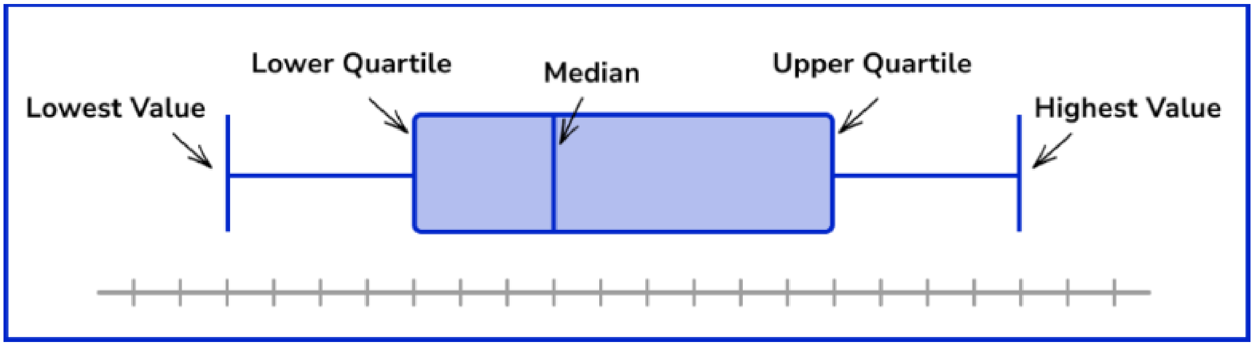
Q3: 7.5

Maximum: 10.0

Draw a vertical axis and label it with the minimum and maximum values in the dataset: 2.5 to 10.0.

Draw a horizontal axis and label it "Apple Weights (ounces)."

Draw a box from Q1 to Q3, with a horizontal line at the median:



Note that there could be or not be outliers in a dataset, so we don't need to plot any points beyond the whiskers.

The resulting box plot shows the distribution of the apple weights, with the majority of the apples weighing between 4.5 and 7.5 ounces. The median weight is around 6 ounces, and the range of the dataset is from 2.5 to 10.0 ounces.

Unit-4:

Correlation & Regression

- Correlation coefficient, line of best fit
- Simple regression line
- Rank correlation/ Spearman Rank

Correlation

In statistics, correlation refers to the relationship between two variables. More specifically, it measures the strength and direction of the linear association between two variables.

Correlation can be either positive or negative, or even zero. A positive correlation means that as one variable increases, the other variable also tends to increase. A negative correlation means that as one variable increases, the other variable tends to decrease. A zero correlation means that there is no linear relationship between the variables.

Correlation is usually quantified using a correlation coefficient, such as Pearson's correlation coefficient (r) or Spearman's rank correlation coefficient (rho). These coefficients range from -1 to +1, with a value of -1 indicating a perfect negative correlation, +1 indicating a perfect positive correlation, and 0 indicating no correlation.

For example, suppose we have a dataset of the heights and weights of 50 people, and we want to measure the correlation between these variables. We can use Pearson's correlation coefficient to calculate this:

Calculate the means and standard deviations of the heights and weights.
For each person, subtract their height and weight from the respective mean and divide by the standard deviation, to obtain standardized scores.
Multiply each person's standardized height score by their standardized weight score, and sum these products.
Divide the sum by the product of the sample size and the sample standard deviation of the heights and weights, to obtain Pearson's correlation coefficient (r).
If the resulting r value is positive, it indicates a positive correlation between height and weight, meaning that taller people tend to weigh more. If the r value is negative, it indicates a negative correlation between height and weight, meaning that taller people tend to weigh less. If the r value is zero, it indicates no correlation between height and weight.

Correlation is an important concept in statistics because it helps us understand the relationship between variables and can inform decisions about how to analyze and interpret data.

Correlation coefficient

The correlation coefficient is a statistical measure that quantifies the strength and direction of the linear relationship between two variables. The most commonly used correlation coefficient is Pearson's correlation coefficient, which is denoted by the symbol "r".

The value of r ranges between -1 and +1. A value of +1 indicates a perfect positive correlation, meaning that as one variable increases, the other variable also increases in a linear fashion. A value of -1 indicates a perfect negative correlation, meaning that as one variable increases, the other variable decreases in a linear fashion. A value of 0 indicates no correlation, meaning that there is no linear relationship between the two variables.

To calculate Pearson's correlation coefficient, you need to have paired data for two variables. Here are the steps to calculate r:

Calculate the mean and standard deviation of both variables.
For each pair of data points, subtract the mean of each variable from the respective data point.
Multiply the resulting deviations for each pair of data points.
Sum the products from step 3.
Divide the sum from step 4 by the product of the standard deviations of both variables.
The resulting value is the correlation coefficient, r.

For example, if we have the following dataset of height and weight of 10 individuals:

Height (inches)	Weight (lbs)
60	115
65	120
70	150
63	123
67	135
68	158
69	160
72	170
74	185
62	120

We can calculate the correlation coefficient between height and weight using the formula above to find that r = 0.71, indicating a positive correlation between height and weight. This means that taller people tend to weigh more.

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

The formula for Pearson's correlation coefficient (r) between two variables X and Y with n data points is:

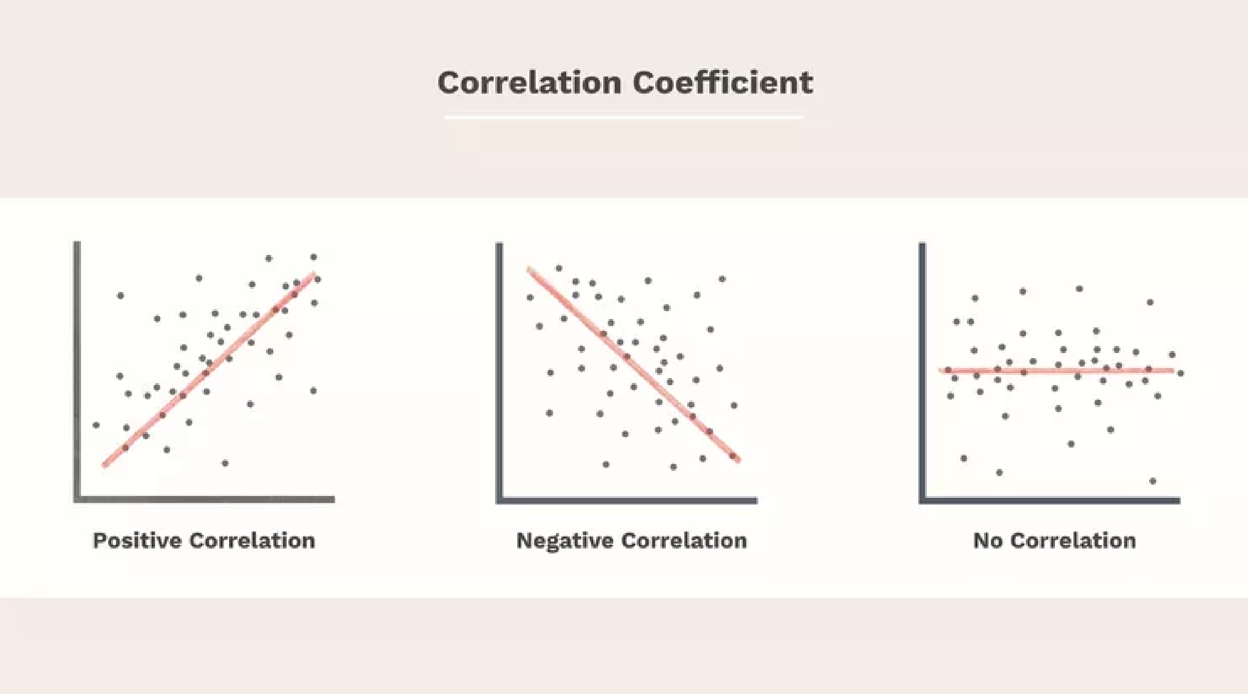
$$r = (\Sigma(Xi - \bar{X})(Yi - \bar{Y})) / (\sqrt{(\Sigma(Xi - \bar{X})^2)}\ \sqrt{(\Sigma(Yi - \bar{Y})^2)})$$

where:

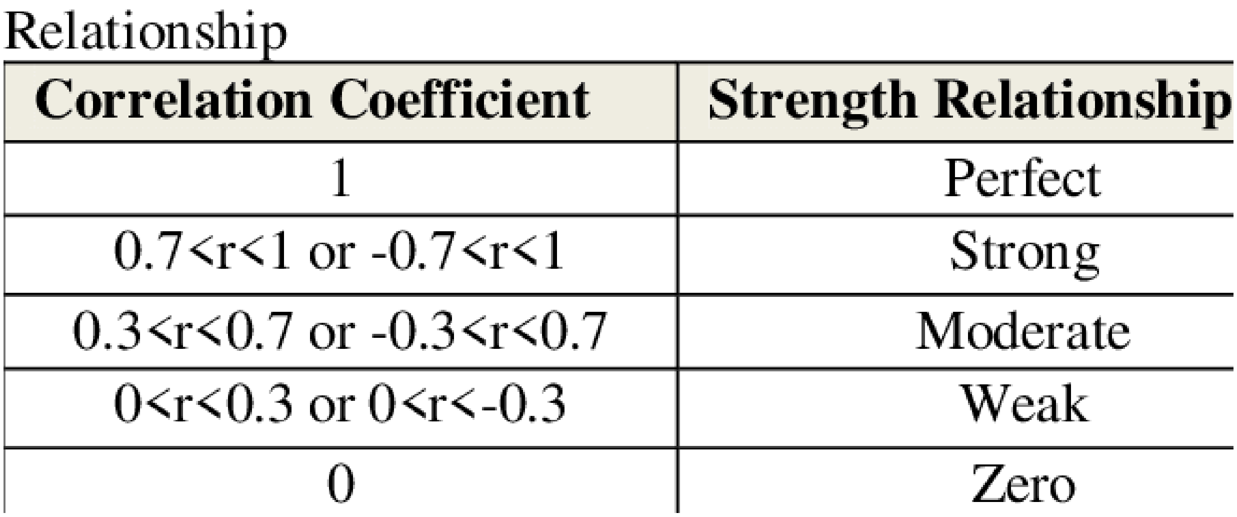
Xi and Yi are the ith observations of the two variables X and Y, respectively
X̄ and Ȳ are the means of the X and Y observations, respectively
This formula calculates the covariance between the two variables, which measures how much the two variables vary together, and then scales this value by the product of their standard deviations. The resulting correlation coefficient r ranges from -1 (perfect negative correlation) to 1 (perfect positive correlation), with 0 indicating no correlation.

Note that this formula assumes that the relationship between X and Y is linear, which may not always be the case in real-world data. Other types of correlation coefficients, such as Spearman's rank correlation coefficient, can be used to measure the strength of non-linear relationships between variables.

Strength of correlation

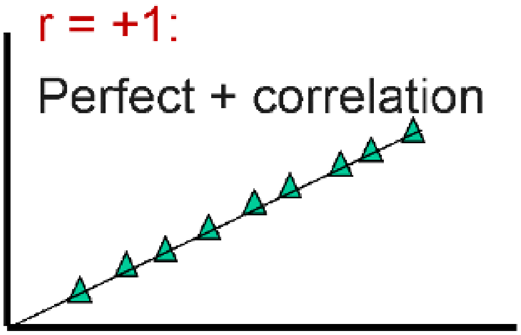


Relationships



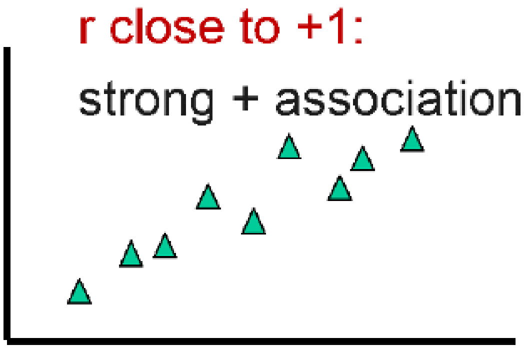
Relationship

Correlation Coefficient	Strength Relationship
1	Perfect
0.7<r<1 or -0.7<r<1	Strong
0.3<r<0.7 or -0.3<r<0.7	Moderate
0<r<0.3 or 0<r<-0.3	Weak
0	Zero



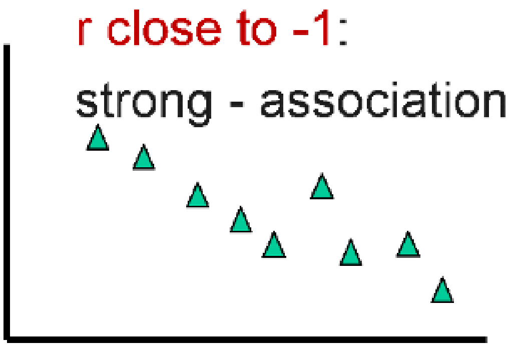
r = +1:

Perfect + correlation



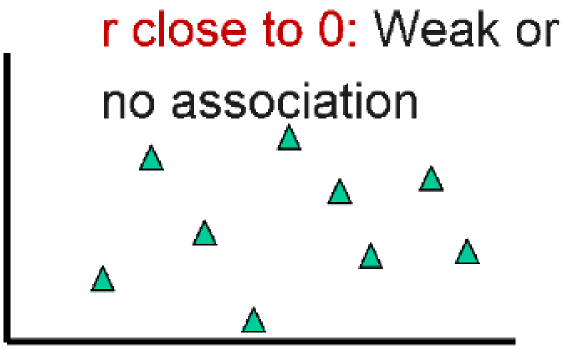
r close to +1:

strong + association

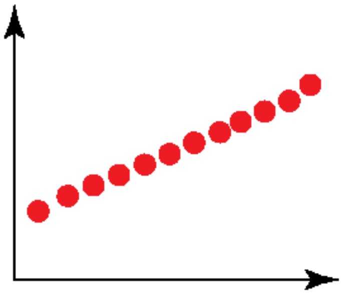


r close to -1:

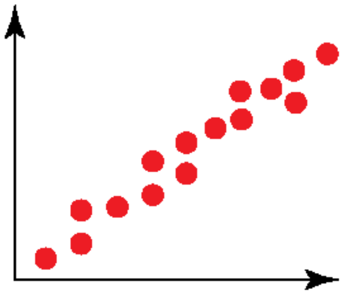
strong - association



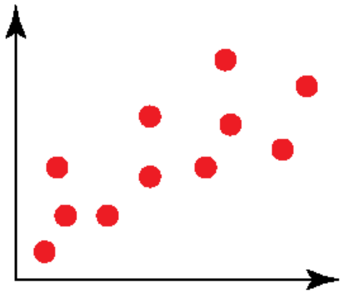
r close to 0: Weak or no association



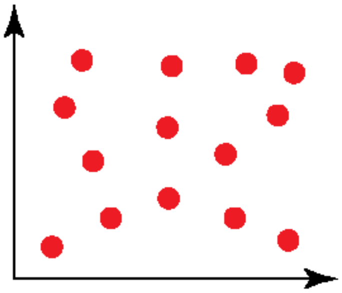
Perfect
Positive
Correlation



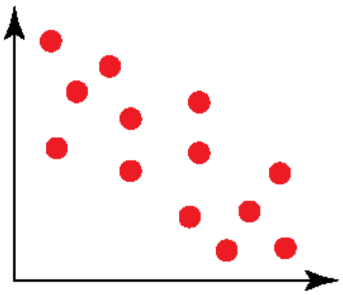
Strong
Positive
Correlation



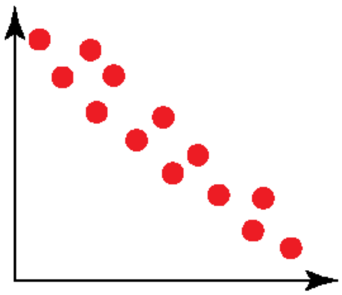
Weak
Positive
Correlation



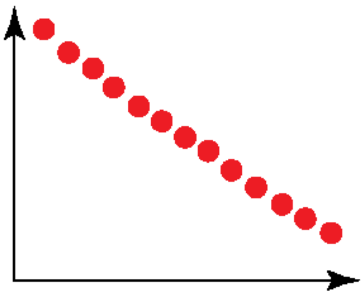
No
Correlation



Weak
Negative
Correlation



Strong
Negative
Correlation



Perfect
Negative
Correlation

Best Fit Line

In statistics, the best fit line, also known as the regression line, is a line that represents the linear relationship between two variables. The line is called the "best fit" because it is the line that best fits the pattern of the data, in the sense that it minimizes the differences between the observed data points and the predicted values of the line.

To find the best fit line, you can use a method called linear regression. The simplest form of linear regression is known as simple linear regression, which involves fitting a line to a dataset of two variables (X and Y). The best fit line is the line that minimizes the sum of the squared differences between the observed values of Y and the predicted values of Y based on the X values. This line can be expressed as:

Y = a + bX

where:

Y is the dependent variable (the variable being predicted)

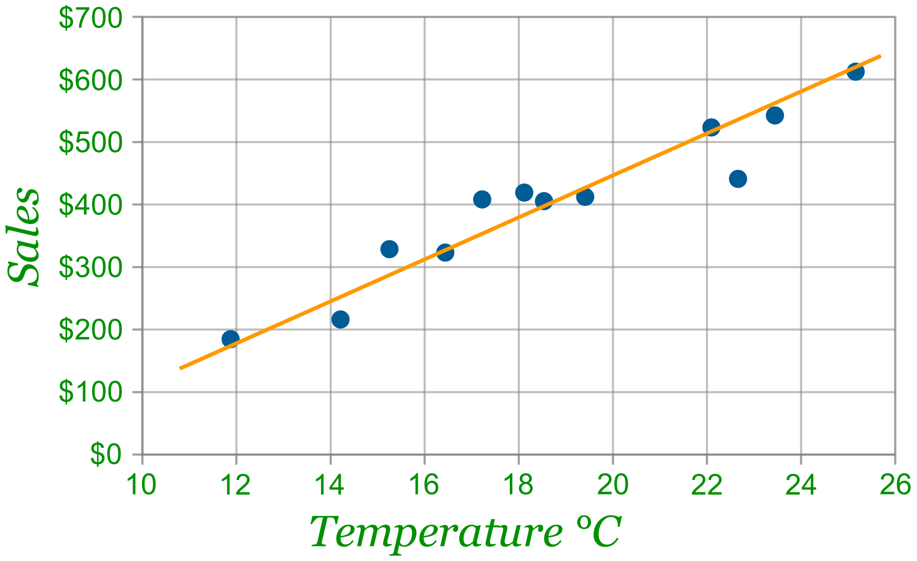
X is the independent variable (the variable used to predict Y)

a is the y-intercept (the value of Y when X is 0)

b is the slope (the change in Y for a one-unit increase in X)

The values of a and b can be estimated using statistical software or by hand using formulas. Once the values of a and b are known, the best fit line can be plotted on a scatter plot of the data, allowing us to visualize the linear relationship between the two variables.

The best fit line can be useful for making predictions about future values of Y based on known values of X, as well as for understanding the strength and direction of the linear relationship between the two variables.



Scatter plot

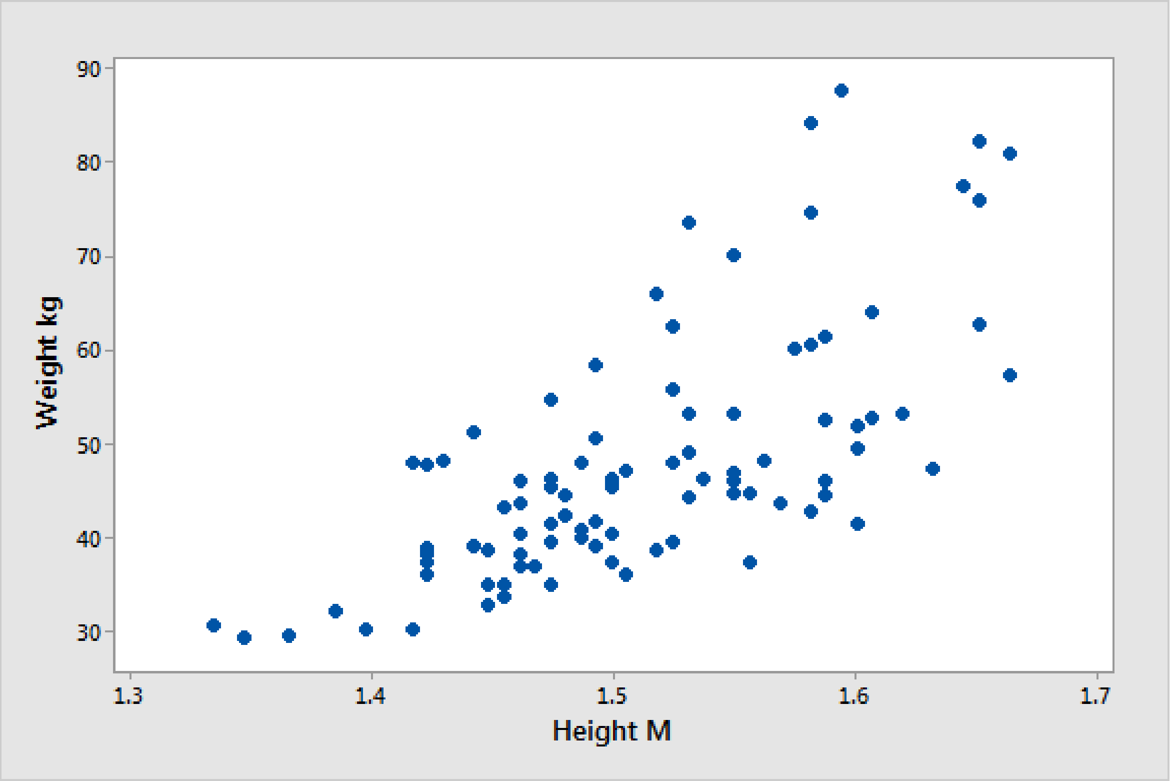
A scatter plot is a graph that displays the relationship between two quantitative variables. It is a useful tool for identifying patterns and trends in data, as well as for examining the strength and direction of the relationship between two variables.

In a scatter plot, each data point represents a single observation of the two variables being plotted. The x-axis typically represents the independent variable, while the y-axis represents the dependent variable. Each data point is plotted at the intersection of its x-value and y-value.

The pattern of the data in a scatter plot can reveal a variety of different relationships between the two variables being plotted. If the data points form a roughly linear pattern, this suggests a strong linear relationship between the two variables. If the data points form a curved pattern, this suggests a non-linear relationship between the two variables. If the data points form a cluster or grouping, this suggests a strong correlation between the two variables within that grouping.

Scatter plots can also be used to identify outliers, which are data points that are significantly different from the rest of the data. Outliers may be due to measurement errors or other factors that cause a particular observation to deviate from the overall pattern of the data.

Overall, scatter plots are a powerful tool for visualizing relationships between two quantitative variables and identifying patterns and trends in data.



Regression

Regression is a statistical method for modeling and analyzing the relationship between two or more variables. It is commonly used to predict the value of one variable based on the values of other variables. In particular, regression analysis is used to estimate the relationship between a dependent variable (also called the outcome or response variable) and one or more independent variables (also called predictors or explanatory variables).

The most common type of regression is linear regression, which involves fitting a straight line to the data points. The equation for a simple linear regression model is:

$y = a + bx + e$

where:

y is the dependent variable (the variable being predicted)

x is the independent variable (the variable used to predict y)

a is the intercept (the value of y when x is 0)

b is the slope (the change in y for a one-unit increase in x)

e is the error term (the part of y that is not explained by x)

The goal of linear regression is to estimate the values of a and b that best fit the data, in the sense that they minimize the sum of the squared differences between the observed values of y and the predicted values of y based on x. Once the values of a and b are estimated, the regression equation can be used to predict the value of y for any given value of x.

Regression analysis can also be used to model more complex relationships between variables, including nonlinear relationships and interactions between multiple predictors. In addition to linear regression, there are many other types of regression models, including logistic regression, Poisson regression, and Cox regression, each of which is designed to model different types of data and relationships between variables.

Regression analysis is a statistical method used to analyze the relationship between one or more independent variables (also called predictors or explanatory variables) and a dependent variable (also called the outcome or response variable). The goal of regression analysis is to identify the nature and strength of the relationship between the independent and dependent variables and to use that information to make predictions about the dependent variable.

Regression analysis can be conducted using various statistical methods, such as simple linear regression, multiple linear regression, logistic regression, and others. The choice of method depends on the nature of the data and the research question being addressed.

In simple linear regression, there is a single independent variable and a single dependent variable, and the relationship between the two variables is assumed to be linear. The regression model estimates the slope and intercept of the line that best fits the data and uses these estimates to predict the value of the dependent variable for a given value of the independent variable.

In multiple linear regression, there are multiple independent variables and a single dependent variable, and the relationship between the independent variables and the dependent variable is assumed to be linear. The regression model estimates the coefficients of each independent variable and uses these estimates to predict the value of the dependent variable for a given combination of the independent variables.

Logistic regression is used when the dependent variable is binary (i.e., it can take on only two values, such as yes/no or success/failure), and the relationship between the independent variables and the probability of the dependent variable taking on one value or the other is assumed to be a logistic function.

Regression analysis can be used in many fields, such as finance, economics, social sciences, and health care, to make predictions about future outcomes, understand the relationships between variables, and identify the factors that influence a particular outcome.

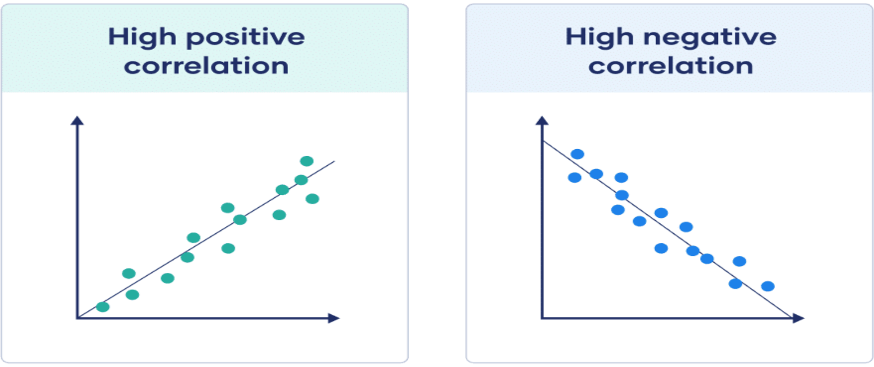
Regression Analysis:

- Linear Regression
- Multiple Linear Regression
- Non-linear Regression

Linear Regression

Linear regression is a statistical method used to analyze the relationship between a dependent variable (also called the response or outcome variable) and one or more independent variables (also called predictors or explanatory variables). The method assumes that the relationship between the variables is linear, meaning that the dependent variable changes in a constant and predictable way with changes in the independent variable(s).

In linear regression, the goal is to find the line that best fits the data points in a scatterplot. This line is called the regression line or the line of best fit. The regression line represents the relationship between the dependent variable and the independent variable(s). The regression line can be used to predict the value of the dependent variable for any given value of the independent variable(s).



Scribbr

The equation for a simple linear regression model is:

$$Y = \beta 0 + \beta 1X + \varepsilon$$

where:

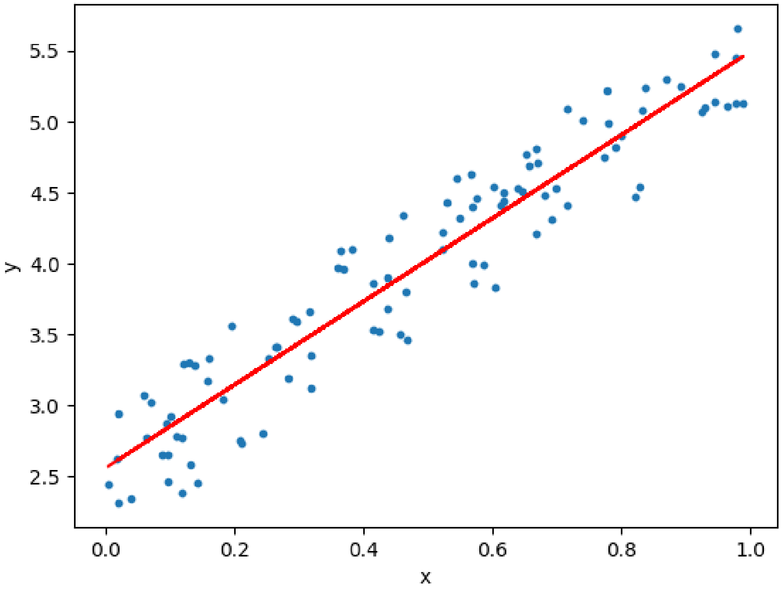
Y is the dependent variable
X is the independent variable
β0 is the intercept (the value of Y when X is 0)
β1 is the slope (the change in Y for a one-unit increase in X)
ε is the error term (the part of Y that is not explained by X)
The goal of linear regression is to estimate the values of β0 and β1 that minimize the sum of the squared differences between the observed values of Y and the predicted values of Y based on X. Once the values of β0 and β1 are estimated, the regression line can be used to predict the value of Y for any given value of X.

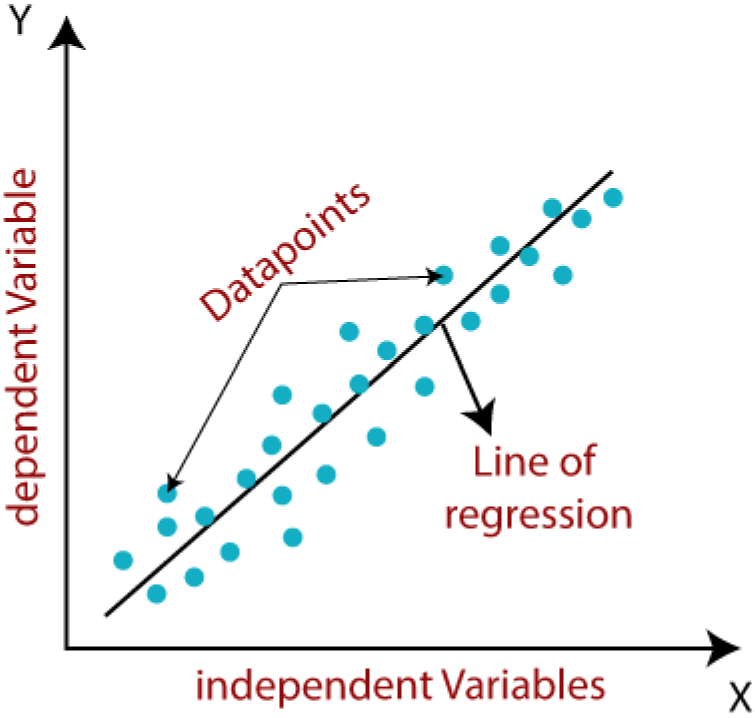
Linear regression can be extended to multiple linear regression, where there are multiple independent variables. The equation for multiple linear regression is:

$$Y = \beta 0 + \beta 1X1 + \beta 2X2 + \ldots + \beta nXn + \varepsilon$$

where X1, X2, ..., Xn are the n independent variables and β1, β2, ..., βn are the corresponding regression coefficients.

Linear regression is widely used in various fields, such as finance, economics, social sciences, and engineering, to model relationships between variables and make predictions.


5.5
5.0
4.5
4.0
3.5
3.0
2.5
y
0.0
0.2
0.4
0.6
0.8
1.0
x


Y
dependent Variable
Datapoints
Line of
regression
independent Variables
X

Multiple linear regression

is a statistical method used to analyze the relationship between a dependent variable (also called the response or outcome variable) and two or more independent variables (also called predictors or explanatory variables). The method extends simple linear regression by allowing for more than one predictor variable, and assumes that the relationship between the variables is linear.

The equation for a multiple linear regression model is:

$$Y = \beta 0 + \beta 1X1 + \beta 2X2 + ... + \beta nXn + \varepsilon$$

where:

Y is the dependent variable

X1, X2, ..., Xn are the independent variables

β0 is the intercept (the value of Y when all X variables are 0)

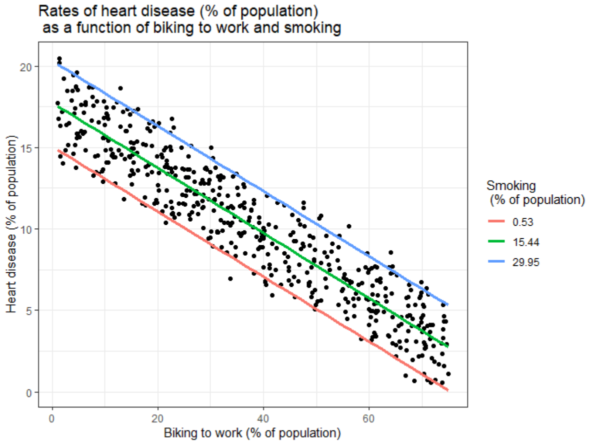
β1, β2, ..., βn are the regression coefficients (the change in Y for a one-unit increase in each X variable, holding all other variables constant)

ε is the error term (the part of Y that is not explained by the X variables)

The goal of multiple linear regression is to estimate the values of β0, β1, β2, ..., βn that minimize the sum of the squared differences between the observed values of Y and the predicted values of Y based on the X variables. Once the values of β0, β1, β2, ..., βn are estimated, the regression equation can be used to predict the value of Y for any combination of X variable values.

Multiple linear regression is useful when there are multiple factors that may influence the dependent variable, and when those factors are not highly correlated with each other. For example, multiple linear regression could be used to predict a person's salary based on their age, education level, years of work experience, and location.

Multiple linear regression is widely used in various fields, such as finance, economics, social sciences, and engineering, to model relationships between variables and make predictions. However, it is important to remember that correlation does not imply causation, and multiple linear regression only identifies associations between variables, not causality.



Non-linear Regression

Nonlinear regression is a statistical method used to model relationships between variables that are not linear. In contrast to linear regression, nonlinear regression allows for more complex functional relationships between the dependent variable and the independent variables.

The equation for a nonlinear regression model is:

$Y = f(X, \beta) + \varepsilon$

where:

Y is the dependent variable

X is the independent variable

β is a vector of parameters that determine the functional form of the relationship between X and Y

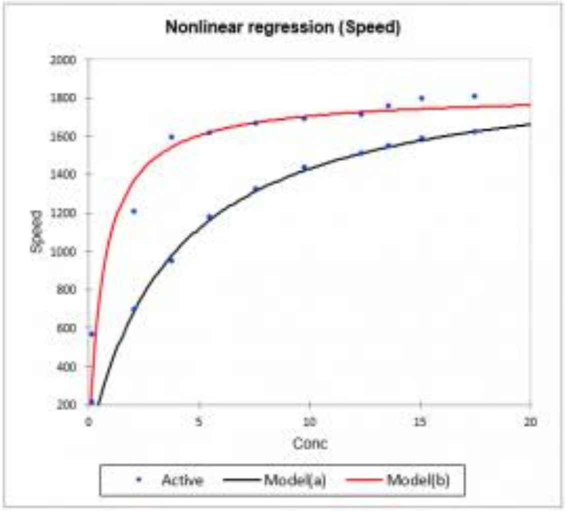
f is a nonlinear function of X and β

ε is the error term

The goal of nonlinear regression is to estimate the values of the parameters β that minimize the sum of the squared differences between the observed values of Y and the predicted values of Y based on X and the nonlinear function f. This optimization process is often done using numerical methods, such as the Gauss-Newton algorithm or the Levenberg-Marquardt algorithm.

Nonlinear regression can be used to model a variety of functional relationships, such as exponential, logarithmic, power, and polynomial functions. However, nonlinear regression can be more complex than linear regression, as it requires more computational resources and may require more complex optimization algorithms.

Nonlinear regression is widely used in various fields, such as biology, engineering, physics, and economics, to model complex relationships between variables and make predictions. However, it is important to choose the appropriate nonlinear function for the data being analyzed and to ensure that the model is not overfitting the data.



Rank correlation/Spearman rank-correlation of coefficient

Rank correlation, also known as nonparametric correlation, is a statistical method used to measure the strength and direction of the relationship between two variables, when the data is measured on ordinal scales or when the data does not meet the assumptions of the parametric correlation methods.

Spearman's rank correlation coefficient, or Spearman's rho (ρ), is a popular rank correlation method used to measure the degree of association between two variables. Spearman's rho is based on the ranks of the observations, rather than their actual values.

The formula for Spearman's rho is:

$$\rho = 1 - (6\Sigma d^2) / (n(n^2 - 1))$$

where:

d is the difference between the ranks of the two variables
n is the number of observations
Spearman's rho ranges from -1 to 1, with a value of 1 indicating a perfect positive rank correlation, 0 indicating no correlation, and -1 indicating a perfect negative rank correlation. A positive value of Spearman's rho indicates that higher ranks of one variable tend to be associated with higher ranks of the other variable, while a negative value indicates that higher ranks of one variable tend to be associated with lower ranks of the other variable.

Spearman's rank correlation coefficient is widely used in various fields, such as psychology, sociology, and business, to analyze the relationship between variables measured on ordinal scales or when the data does not meet the assumptions of the parametric correlation methods. It is a robust method, as it is less sensitive to outliers than the Pearson correlation coefficient, which is based on the actual values of the variables.

Let's say we want to measure the rank correlation between the hours of exercise per week and the number of cups of coffee consumed per day, for a group of 10 individuals. Here are the observed data:

Individual	Exercise (hours/week)	Coffee (cups/day)
1	5	2
2	2	4
3	4	1
4	3	3
5	6	2
6	1	5
7	2	4
8	5	1
9	3	3
10	4	2

To calculate Spearman's rank correlation coefficient, we first need to rank the observations for each variable:

Individual	Exercise (hours/week)	Rank	Coffee (cups/day)	Rank
1	5	6	2	7
2	2	2	4	4
3	4	4	1	1
4	3	3	3	6
5	6	8	2	7
6	1	1	5	10
7	2	2	4	4
8	5	6	1	1
9	3	3	3	6
10	4	4	2	7

Next, we calculate the difference between the ranks for each individual and square them:

Individual	Exercise (hours/week)	Rank	Coffee (cups/day)	Rank	d	d^2
1	5	6	2	7	-1	1
2	2	2	4	4	-2	4
3	4	4	1	1	0	0
4	3	3	3	6	-3	9
5	6	8	2	7	1	1
6	1	1	5	10	-9	81
7	2	2	4	4	-2	4
8	5	6	1	1	5	25
9	3	3	3	6	-3	9
10	4	4	2	7	-3	9
Total	-	-	-	-	-	143

Finally, we substitute the values into the formula for Spearman's rho:

$\rho = 1 - (6\Sigma d^2) / (n(n^2 - 1))$
$= 1 - (6 \times 143) / (10 \times (10^2 - 1))$
$= 1 - 858 / 990$
$= -0.8687$

Therefore, the rank correlation coefficient between exercise and coffee consumption is -0.8687, which indicates a strong negative rank correlation between the two variables. This means that individuals who exercise more tend to consume fewer cups of coffee per day, while individuals who exercise less tend to consume more cups of coffee per day.

Unit-5:
Conclusion

Statistics - Making Conclusions
Using statistics to make conclusions about a population is called statistical inference.

Statistical Inference

Statistics from the data in the sample is used to make conclusions about the whole population. This is a type of statistical inference.

Probability theory is used to calculate the certainty that those statistics also apply to the population.

When using a sample, there will always be some uncertainty about what the data looks like for the population.

Uncertainty is often expressed as confidence intervals.

Confidence intervals are numerical ways of showing how likely it is that the true value of this statistic is within a certain range for the population.

Hypothesis testing is a another way of checking if a statement about a population is true. More precisely, it checks how likely it is that a hypothesis is true is based on the sample data.

Some examples of statements or questions that can be checked with hypothesis testing:

People in the Netherlands taller than people in Denmark
Do people prefer Pepsi or Coke?
Does a new medicine cure a disease?
Note: Confidence intervals and hypothesis testing are closely related and describe the same things in different ways. Both are widely used in science.

Causal Inference

Causal inference is used to investigate if something causes another thing.

For example: Does rain make plants grow?

If we think two things are related we can investigate to see if they correlate. Statistics can be used to find out how strong this relation is.

Even if things are correlated, finding out of something is caused by other things can be difficult. It can be done with good experimental design or other special statistical techniques.

Note: Good experimental design is often difficult to achieve because of ethical concerns or other practical reasons.

A report representation

An overview on methodology and analysis of data