Applied Data Science with Python

Statistics Fundamentals

Learning Objectives

By the end of this lesson, you will be able to:

- Understand the fundamentals of statistics and its terminology
- Compare different types of statistics
- Analyze data categorization and its types
- Calculate the measures of central tendency and dispersion

Learning Objectives

By the end of this lesson, you will be able to:

- Analyze random variables and sets
- Examine statistical fundamentals like skewness, covariance, and correlation
- Calculate the partition of variants using Python

Business Scenario

A retailer is having difficulty assessing the sales performance of its products across several locations. To address this, it collects sales data for each product and store, including the number of units sold and revenue generated.

It intends to measure the central tendency and variability of sales data by employing statistical metrics such as mean, median, mode, standard deviation, and skewness. Measures of shape, such as kurtosis, will be used to determine if the sales distribution is heavy-tailed or light-tailed. Also, by analyzing the covariance and correlation between product sales and store location, it will identify significant relationships or dependencies.

This information will help the company make informed decisions regarding product inventory, pricing adjustments, and marketing initiatives to improve overall sales success.

Importance of Statistics for Data Science

Discussion: Statistics for Data Science

Duration: 10 minutes

- What is statistics, and why is it important?
- What are the different types of statistics?

Importance of Statistics

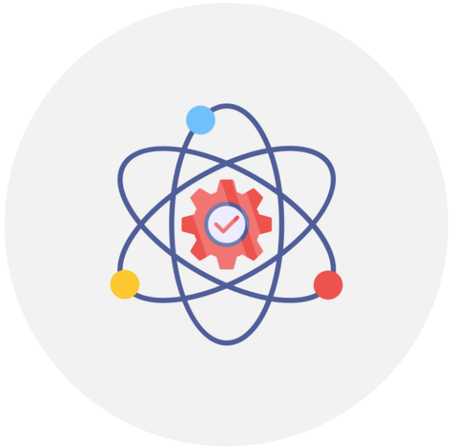
Knowledge of the core concepts of statistics helps in accurately analyzing data, making informed decisions, and effectively utilizing statistical techniques in various fields.

This knowledge also helps in building ensemble methods.

Ensemble methods combine different machine learning algorithms to get the highest possible level of accuracy.

Data Science

Data science is a combination of computer science and statistics.



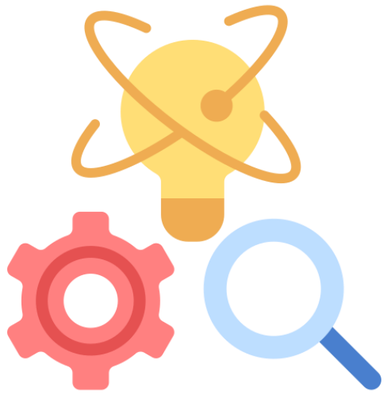
A thorough understanding of statistics is vital for developing a strong intuition for machine learning algorithms.

Otherwise, the algorithms resemble black boxes that give output at a certain level of accuracy.

The user of the black box machine learning model can understand the results but cannot see the logic behind it.

Importance of Statistics for Data Science

Probability theory, which is the foundation of statistics, was developed between the 17th and 19th centuries by:



- Thomas Bayes
- Pierre-Simon Laplace
- Carl Gauss

Statistics is an applied branch of science, and its primary objective is to analyze data.

Exploratory Data Analysis (EDA)

It is the process of analyzing and summarizing the characteristics of a dataset to gain insights and identify patterns.

It is the first step of any data science project and a relatively new area of statistics.

Classical statistics is focused on inference.

Conclusions about large populations are made by studying small samples.

The advancement of computers and software helped expand EDA concepts and applications.

Exploratory Data Analysis (EDA)

The following factors have contributed to its growth:

Rapid development of new technology

Access to bigger data

Greater use of quantitative analysis

Method of Data Interpretation

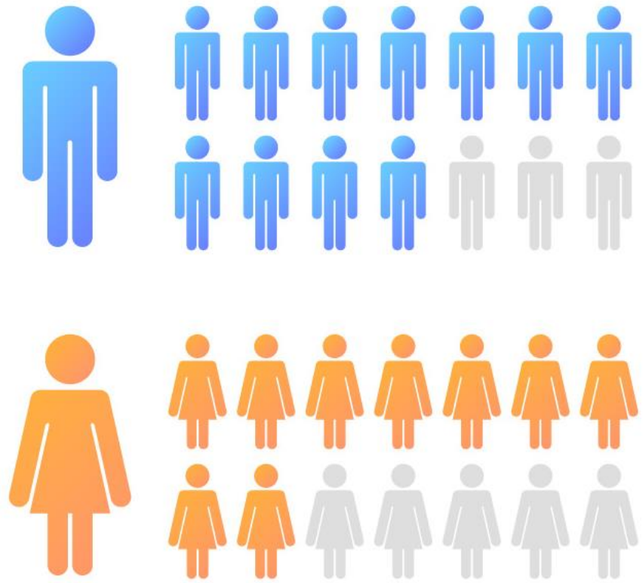
Statistics, as a discipline, performs the following on the data:

A data scientist must mine the raw data, combine domain expertise to find patterns, and use the findings for decision-making in real-world situations.

Common Statistical Terms

Population and Sample

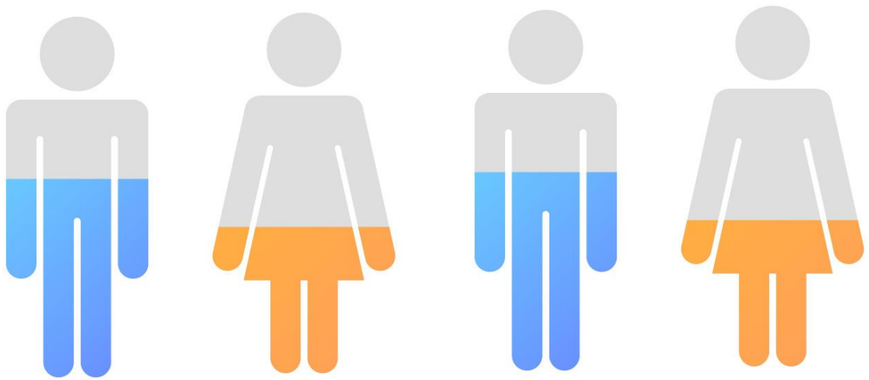
Population is the complete data pool from where a sample is drawn for further analysis.



A sample is a subset of the population.

Measurement and Sample Data

A measurement is a number or attribute calculated for each member of the population or sample.
Example: Height, weight, and age



The measurements of the sample members are collectively called sample data.

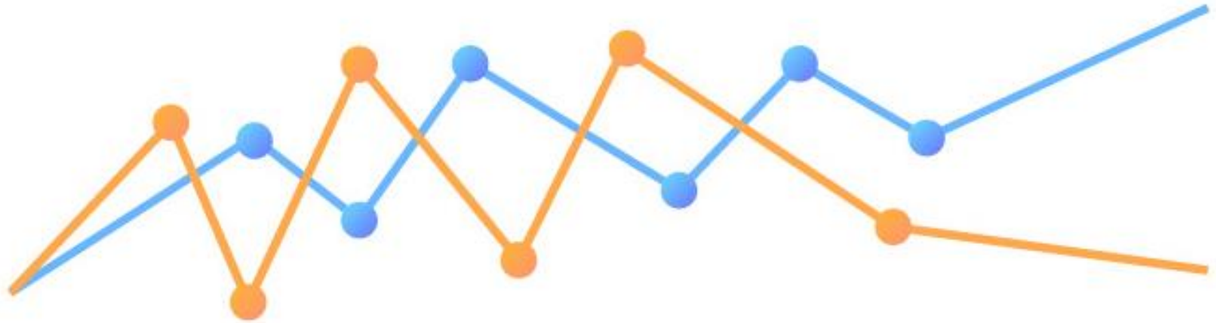
Parameter

It is a characteristic of the population to estimate or test.

Example: Population mean

Statistic

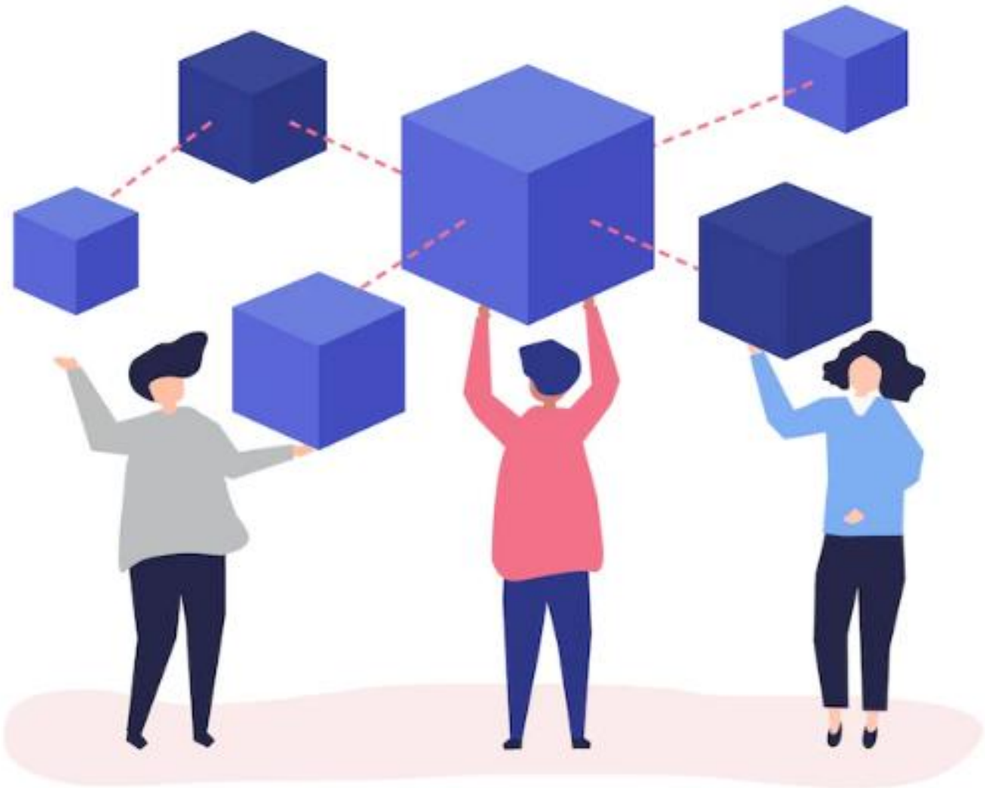
It is a characteristic of the sample that usually helps to estimate or test a population parameter.



Example: Mean

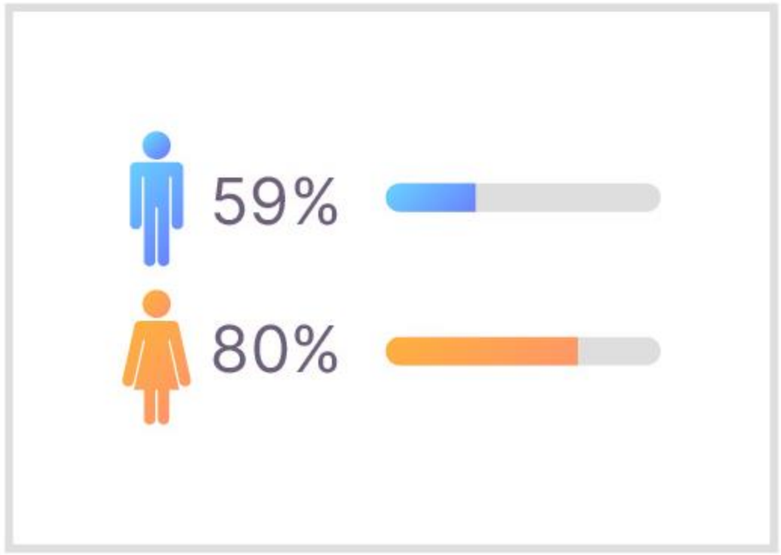
Variable

A variable is something that can take on different values in the data set.



Distribution

The distribution of a variable is the possible number of times an outcome can occur out of a number of trials.



Descriptive Statistics

It is a branch of statistics that involves organizing, displaying, and describing data.

Inferential Statistics

It is a branch of statistics that draws conclusions about the population based on the information obtained from a sample taken from that population.

Qualitative Data

It is the data that is not numerical but categorical.

Example: Labels and attributes

Quantitative Data

It is the data that is numerical and belongs to a certain numerical scale.

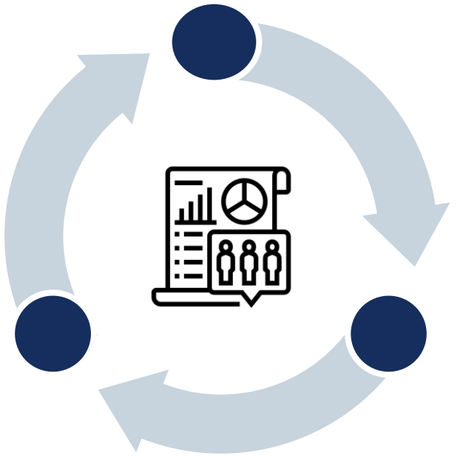
Example: Height measurement

Types of Statistics

Types of Statistics

There are three main categories of statistical theory:

Inferential statistics



Descriptive statistics

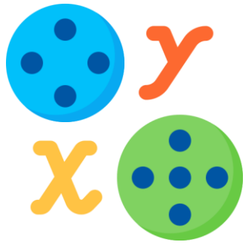
Predictive statistics

Descriptive Statistics

It involves the following methods to describe the various features of data:

Collection

Presentation



Characterization

Inferential Statistics

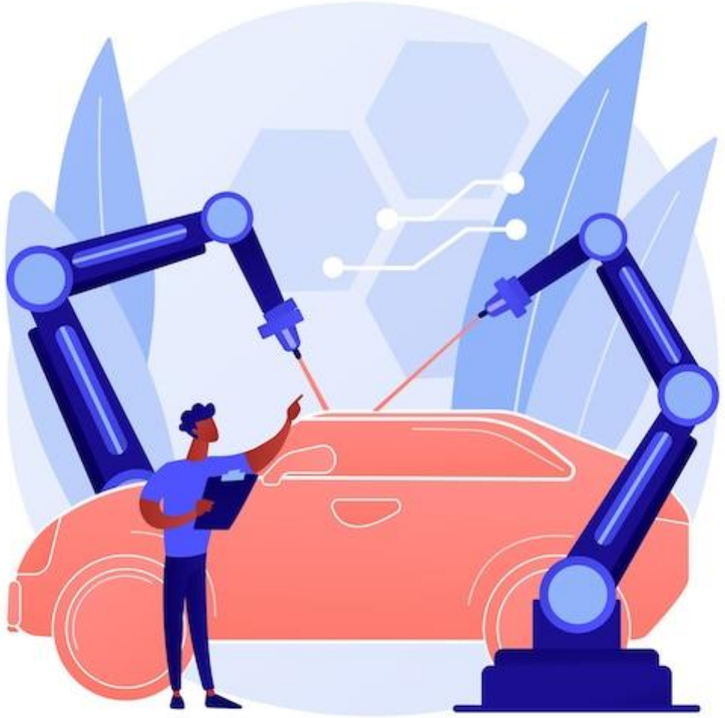
Based only on sample results, inferential statistics:

Enables estimation of a characteristic of a population

Facilitates decision-making regarding a population

Understanding Types of Statistics: An Example

For estimating the number of automobiles manufactured in a month, the entire output is considered as the population.



Automobiles are inspected for quality characteristics like:

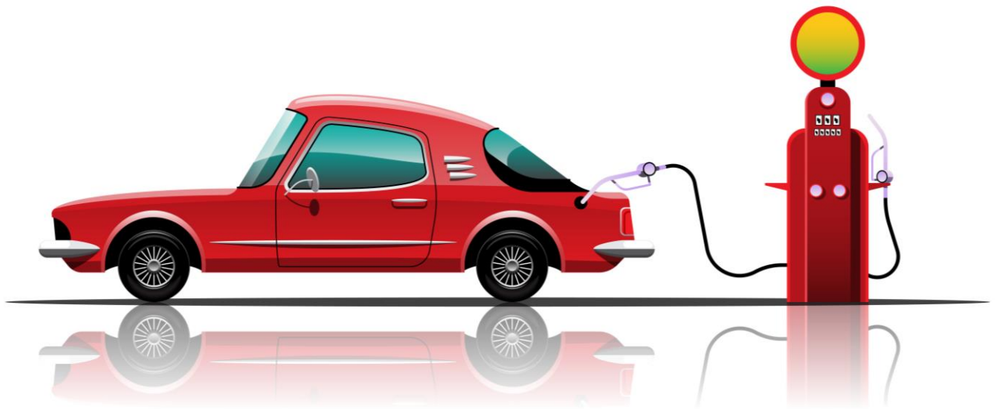
A sample of miles per gallon data was collected from a sample of automobiles.

The parameter represents the average mileage of all cars in the entire population.

The statistic represents the average life of the automobiles in the sampled data.

Understanding Types of Statistics: An Example

If dealing with descriptive statistics, it would include:



- Selection of a sample
- Presentation of the sample data
- Computation of the value of a statistic

Understanding Types of Statistics: An Example

Inferential statistics is used to generalize the population.

Example

From the sample being studied, is the population average at least 23 miles per gallon?

Predictive Statistics

It can be defined as the science of extracting information from data and using it to predict:

- Trends
- Behavior patterns
- Relationships between characteristics

Example

Data on the number of residents in a city recorded over the years could be used to predict the city's future population and estimate the demand for infrastructure in the years to come.

Inter-Relatedness of Data

Sometimes data on two or more characteristics tend to have relationships.

It can be used to predict the values of one characteristic if the values of the others are known.

Example

Predicting the demand for spares in the forthcoming years by evaluating the sales of cars during the past few years

Discussion: Statistics for Data Science

Duration: 10 minutes

- What is statistics, and why is it important?

Answer: Statistics refers to a branch of mathematics dealing with data collection, analysis, interpretation, presentation, and modeling. The understanding of core statistical concepts allows accurate data analysis, informed decision-making, and effective use of statistical techniques across various fields.

- What are the different types of statistics?

Answer: Statistics primarily fall into three categories: descriptive, inferential, and predictive.

Data Categorization and Types of Data

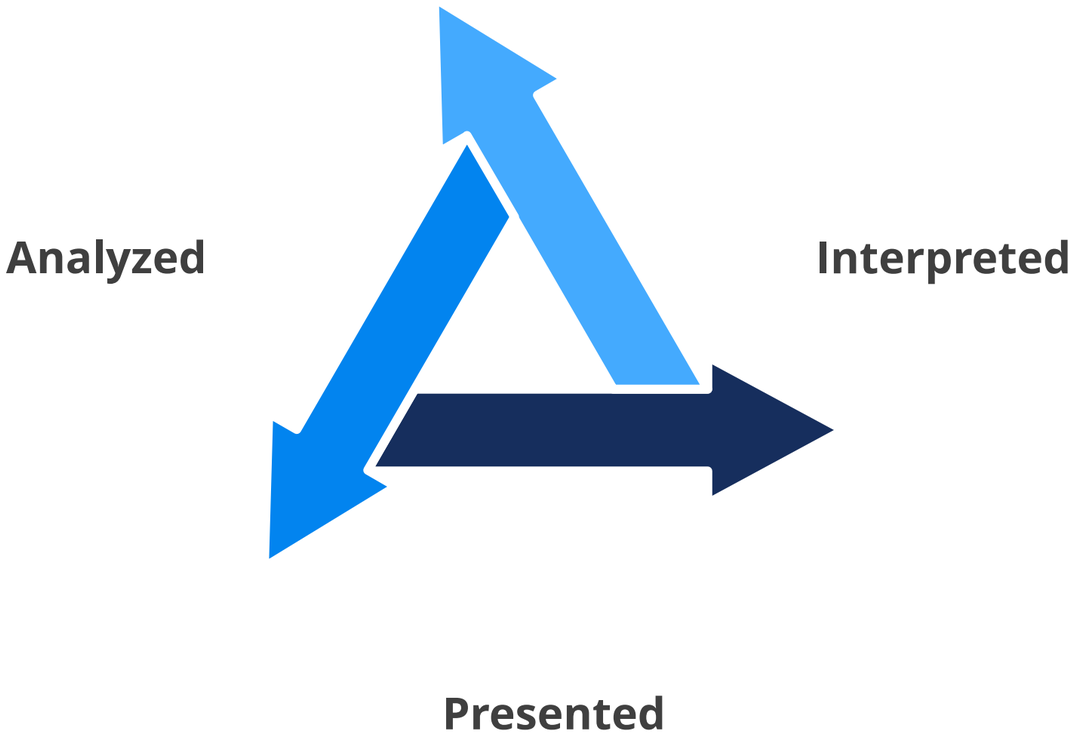
Discussion: Categorization and Types of Data

Duration: 10 minutes

- What is data?
- What types of data exist?

Data

Data is a collection of facts and figures and can be:

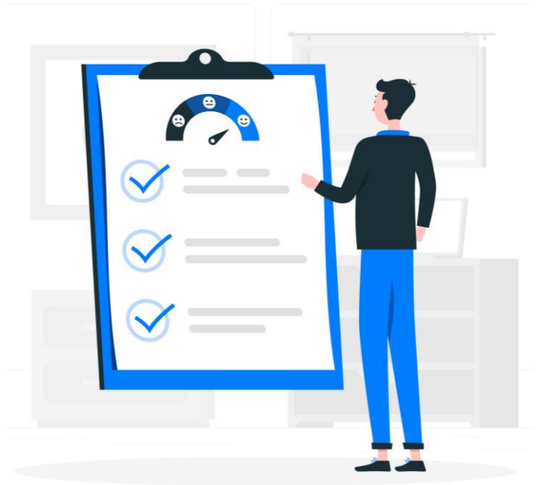


Data Categorization

Data must be categorized before performing statistical analysis to gain useful insights from the varied types of data collected.

Example

Evaluating employee satisfaction through survey



The data collected can be studied by grouping the respondents into different categories, such as:

- Level of seniority
- Functional affiliation
- Work experience
- Location

Data Categorization

Categorization is vital for identifying the possible differences between various categories.

Statistical data can be broadly categorized into:

Qualitative, attributed, or categorical data

Quantitative, variable, or measurable data

Qualitative Data

Data that is stated descriptively is referred to as qualitative data.

Example

- Customer preferences for a particular product brand
- Types of faults in a game of tennis, such as hitting on the net or outside the baseline

It cannot be quantified.

Quantitative Data

Data that is specified numerically through a process of measurement or numerical count is called quantitative data.

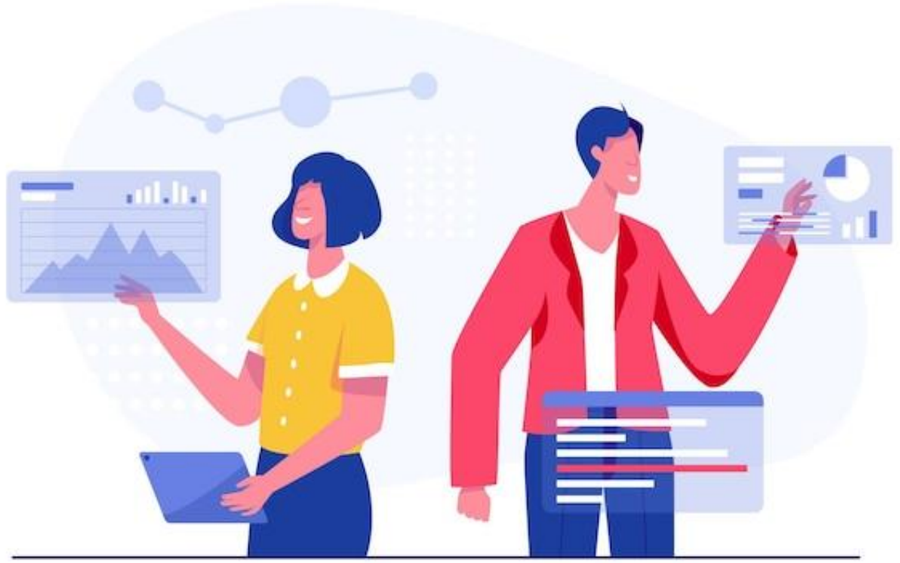
Example

- Diameter of a pole
- Seats occupied in a theater

It can be used in arithmetic operations.

Combination of Quantitative and Qualitative Data

The nature of the characteristic studied determines whether qualitative or quantitative data should be used.



Quantitative data can be used to create qualitative categories or groups based on specific criteria or characteristics for further analysis or interpretation.

Combination of Quantitative and Qualitative Data

Example 1

Diameter of a pole

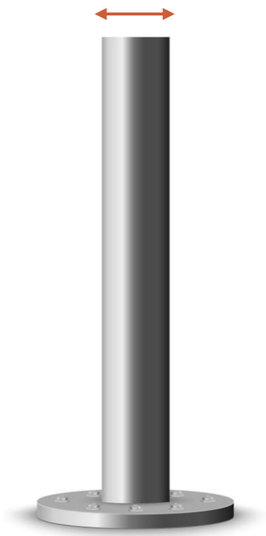
The pole may be inspected using a go or no-go gauge and categorized as within or outside specification limits.

- The go or no-go determination is defined by predefined specification limits or tolerances.
- If the pole's dimensions or attributes fall within the specified limits, it is categorized as a go, indicating that it meets the required standards.
- If the measurements or attributes exceed the specified limits, it is categorized as a no-go, indicating that it does not meet the required standards and requires further action or correction.

Combination of Quantitative and Qualitative Data

Example 1

Diameter of a pole



There are no direct measurements.

The qualitative data is directly collected.

The process of data collection is thus simplified.

Combination of Quantitative and Qualitative Data

Example 2

Scores of students

The scores of students may be used to classify them into several grades such as:

Grade A, Grade B, Grade C, and Grade F.

The marks of each student are calculated to assign a grade.

Quantitative data is more convenient to use while disseminating information.

Discussion: Categorization and Types of Data

Duration: 10 minutes

- What is data?

Answer: Data is a collection of facts and figures and can be analyzed, interpreted, and presented.

- What types of data exist?

Answer: Statistical data can be broadly categorized into:

- Qualitative, attributed, or categorical data
- Quantitative, variable, or measurable data

Levels of Measurement

Levels of Measurement

Levels of measurement demonstrate precisely how variables are recorded.

They are also called scales of measurement.

Understanding the levels of measurement is important to determine the statistical analysis needed to address the problem at hand.

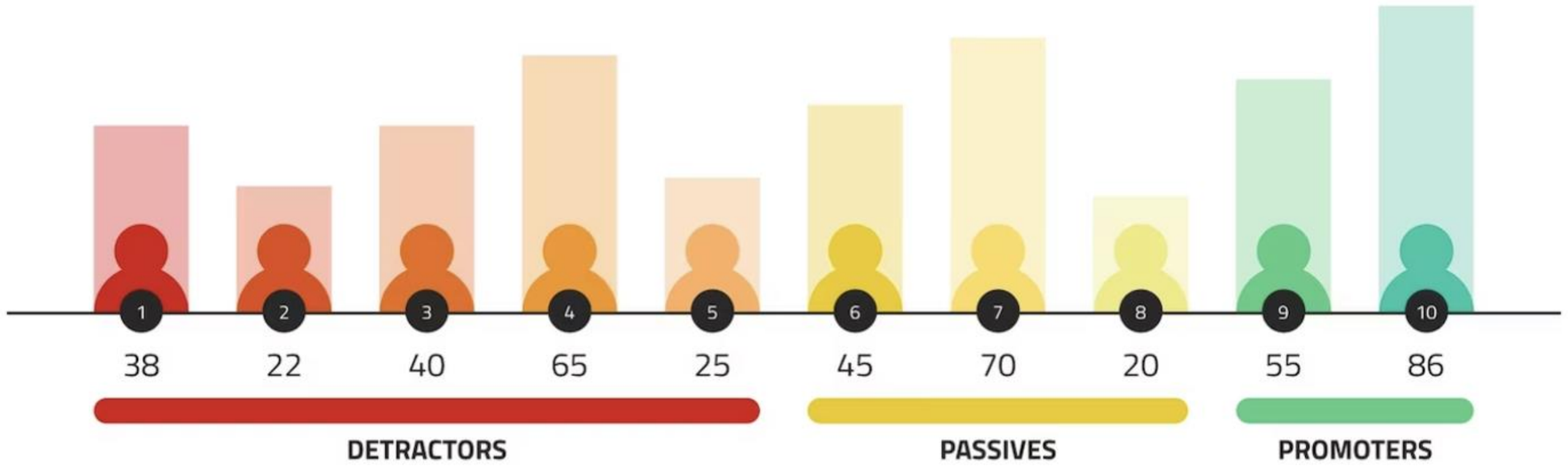
Levels of Measurement

There are four levels of measurement.

- Nominal level
- Ordinal level
- Interval level
- Ratio level

Nominal Level

In the nominal level of measurement, the data is only categorized.



The numerical values assigned to each category can’t be construed as absolute values.

Nominal Level

Example

Assigning numbers to brands when collecting data on cell phone brand preferences

1

Apple

SAMSUNG

2

Samsung

NOKIA

3

Nokia

These numbers here are used only to categorize the data.

In some cases, letters or words may be used, depending on the type of categorization.

Ordinal Level

Here, a range of values is assigned to a category where each category is ranked.

Example

Categorizing students based on their scores

Grade A:

Above 90

Grade B:

Between 80 and 89

Grade C:

Between 70 and 79

Values in each category cannot be distinguished.

Interval Level

In this level, the data is categorized and ranked.

The distances between each interval on the scale are equivalent along the scale, from low to high.

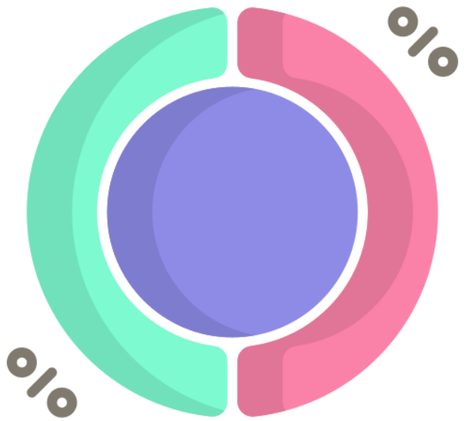
There is no true zero point.

Example

The difference between 100° C and 120° C is equivalent to the difference between 930° C and 950°C. But zero degree centigrade doesn't imply an absence of temperature.

Ratio Level

In this level, in addition to the interval level, there exists an absolute zero value.



An absolute zero here means that the variable is absent.

Example

The Kelvin temperature scale has no negative temperatures.

Measures of Central Tendency: Mean

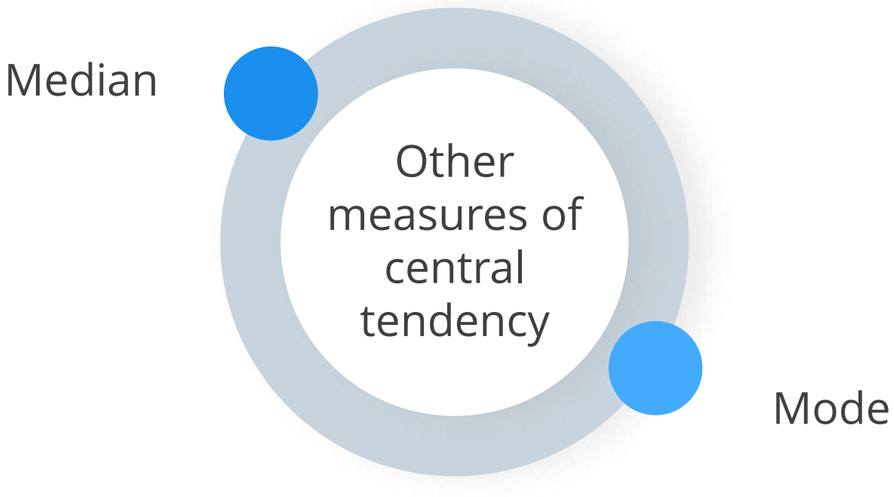
Discussion: Mean, Median, and Mode

Duration: 10 minutes

- What are mean, median, and mode?

Measures of Central Tendency

Measures of central tendency or measures of central location refers to measures that describe the central position in a dataset.



The most common measure of central tendency is the average, or mean.

Mean

The mean is a statistical measure that represents the average value of a set of numbers.

It is calculated for both:

Continuous data

Discrete data

To calculate mean, add all the numbers in the data and divide the result by the number of data points.
Mean is sensitive to outliers and skewed data.

Mean

Example: Consider a set of n numbers given by (x_1, x_2, ……., x_n)

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

Mean of the seven numbers 4, 89, 54, -7, -9, 27 and 5 is given by:

= (4+89+54-7-9+27)/7 = 158/7 = 22.57

Mean

Python code implementation for calculating mean:

```
import statistics

dataset = [2, 4, 6, 8, 10]

mean_value =
statistics.mean(dataset)

print("Mean = ", mean_value)
```

Output:
Mean = 6

- The mean function from the statistics library is utilized by passing the dataset as an argument.
- This function computes the mean of the dataset, which is then stored in the variable mean_value.
- The mean function calculates the mean of the dataset, which in this case is 6.

Measures of Central Tendency: Median

Median

Median is the middle number obtained by arranging data in either ascending or descending order.

When the total number of data points is odd, the exact middle number is the median.

When the total number of data points is even, the average of the two middle numbers is the median.

The median is not sensitive to outliers and skewed data.

Median

Examples:

Dataset (odd)	Dataset (even)
5, 76, 98, 32, 1, -6, 34, 3, -65	5, 76, 98, 32, 1, 99, -6, 34, 3, -65
Ascending order	**Ascending order**
-65, -6, 1, 3, 5, 32, 34, 76, 98	-65, -6 ,1, 3, 5, 32, 34, 76, 98, 99
Median	**Median**
Fifth number = 5	(5+32)/2 = 18.5

Median

Python code implementation for calculating median:

```
import statistics

dataset = [2, 4, 6, 8, 10, 12]

median_value =
statistics.median(dataset)

print("Median = ", median_value)
```

- The median function from the statistics library is utilized by passing the dataset as an argument.
- This function computes the median of the dataset, which is then stored in the variable median_value.
- The median function calculates the median of the dataset, which in this case is 7.

Output:
Median = 7

Measures of Central Tendency: Mode

Mode

Mode is the most frequently occurring data point in the set.

Advantage

Can be calculated for both numerical and categorical data

Disadvantage

May not reflect the center of the distribution well

A few data points at one end of the spectrum may satisfy the definition of mode, but they may be far from the real center of the spectrum.

Mode

Example: To infer the cell phone brand preference from categorical data, data from 100 people is collected.

Preferred brand	No. of people
Nokia	12
Realme	32
Apple	10
Samsung	36
Oppo	4
Vivo	6

The mode of the given data set is Samsung.

Mode

Python code implementation for calculating mode:

```
import statistics

dataset = [2, 4, 4, 6, 8, 4, 10]

mode_value =
statistics.mode(dataset)

print("Mode = ", mode_value)
```

Output:
Mode = 4

- The mode function from the statistics library is utilized by passing the dataset as an argument.
- This function computes the mode of the dataset, which is then stored in the variable mode_value.
- The mode function calculates the mode of the dataset, which in this case is 4.

Discussion: Mean, Median, and Mode

Duration: 10 minutes

- What are mean, medium, and mode?

Answer:

- The mean represents the average value of a set of numbers. It is calculated by adding all the numbers together and then dividing by the total count of numbers.
- The median is the middle number in a set of data points that have been arranged in either ascending or descending order. If there is an even number of data points, the median is the average of the two middle numbers.
- The mode refers to the data point that appears most frequently in a set. If no number repeats, the set does not have a mode.

Measures of Dispersion

Measures of Dispersion

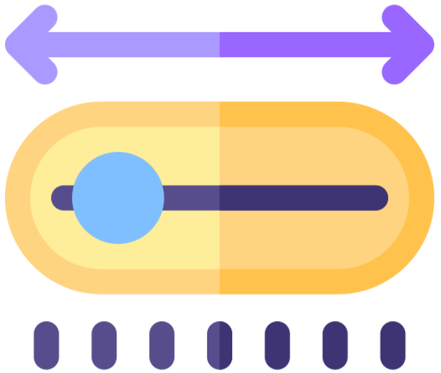
When two separate datasets share the same mean or median, it is difficult to evaluate the level of data variation around the mean.

Measures of dispersion give an idea of data variability around the central point.

The commonly used measures of dispersion are Range, Interquartile Range, and Standard Deviation.

Range

Range is the difference between the largest and the smallest data points in a set.

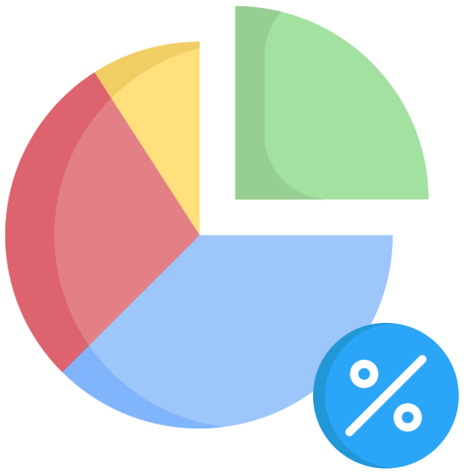


Though easy to calculate, it is sensitive to outliers and doesn't use every data point in the set.

Providing maximum and minimum values in the set makes more sense than range.

Interquartile Range

It is the difference between the 25th and the 75th percentile.



- It describes the middle 50% of the observations, and if they are spaced widely apart, their interquartile range will be large.
- It is useful even if the extreme values are not accurate, as it is insensitive to them.
- It is not amenable to mathematical manipulation.

Interquartile Range: Example

Consider the following dataset where the values are arranged in ascending order:

[10, 15, 20, 25, 30, 35, 40, 50, 70, 100]

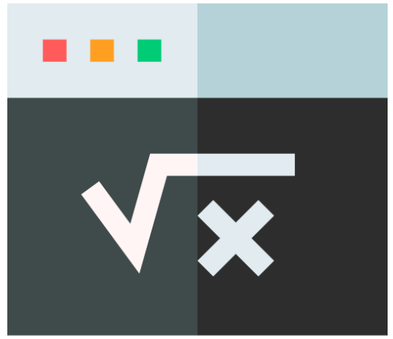
The 25th percentile = average of 2nd and 3rd values = (15 + 20)/2 = 17.5.

The 75th percentile = average of the 7th and 8th values = (40 + 50)/2 = 45.

The interquartile range = 45 - 17.5 = 27.5.

Standard Deviation

Standard deviation (SD), the most popular measure of dispersion, measures the spread of data around the mean.

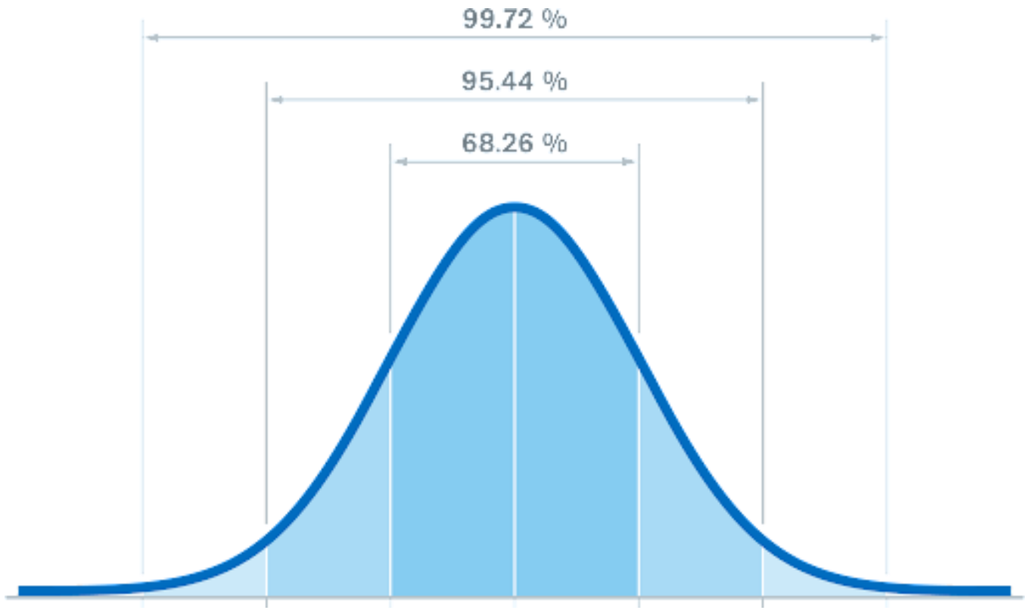


It is defined as the square root of the sum of the squares of the deviation around the mean divided by the number of observations.

$$SD = [(\Sigma_i (x_i - \bar{x})^2 / n]^{1/2} = [\Sigma_i x_i^2 - (\Sigma x)^2/n]^{1/2}$$

Standard Deviation

The advantage of SD is that if the data points are from a normal distribution:



- 68% of the observations lie at a distance of one SD from the mean
- 99.7% between three SDs
- 95% between two SDs

It can also detect skewness but is not an appropriate measure of dispersion for skewed data.

Standard Deviation

Python code implementation for standard deviation:

```
import statistics

data = [1, 2, 3, 4, 5]
std_dev = statistics.stdev(data)

print("Standard Deviation = ", std_dev)
```

- In this example, statistics.stdev() takes a list of numbers as an argument and returns the standard deviation of the data.
- The calculated standard deviation is then printed to the console, which in this case is 1.5811.

Output:
Standard Deviation = 1.5811

Sets

Sets

A set is a well-defined collection of objects. Every member of a set is called an element.

- Two sets are equal if they have the same elements.
- A null set has no elements.
- If every element of set X is present in set Y, then X is a subset of Y.

Sets

A set is usually denoted by capital letters.	Example: X, Y, A, B
The elements of a set are denoted by small letters.	Example: a, b, x, y
A set is represented by enclosing its elements within curly braces.	Example: Set A with elements 2, 6, 4, 9, 12 is written as A = {2, 6, 4, 9, 12}

Sets

A null set is denoted by {} or ø.

A set can also be represented by a rule, such as a set of even numbers.

A = {2, 4, 6,.....}

Using Sets in Probability

A statistical experiment has many possible outcomes.

The set of all outcomes is called the sample space.

A particular outcome or a combination of outcomes is a subset of the sample space called an event.

Set Operations

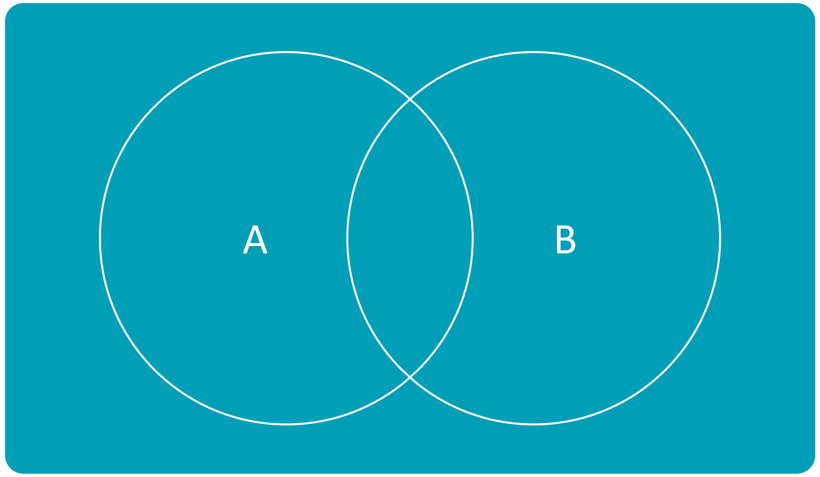
Consider sample space S = {a, c, d, f, g, i, k} and two subsets of S, A = {a, c, d} and B = { d, f, i}.

Union of sets

Set with elements belonging to either or both.

It is denoted by U.

A U B = {a, c, d, f, i}



Set Operations

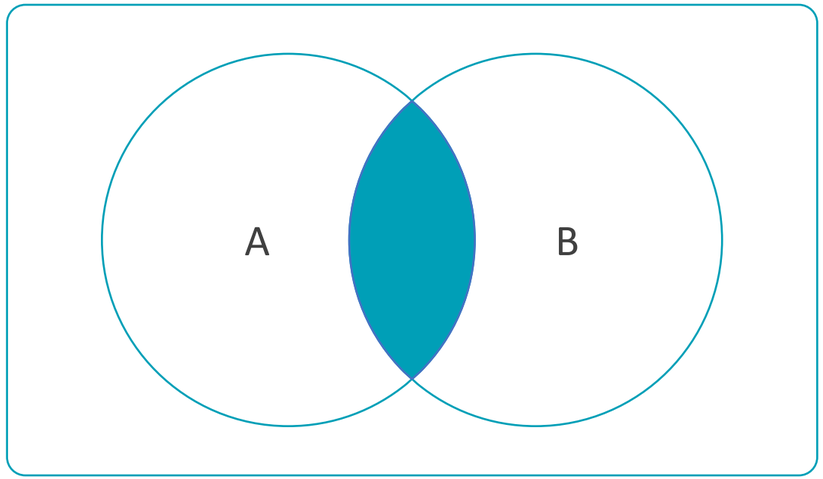
Consider sample space S = {a, c, d, f, g, i, k} and two subsets of S, A = {a, c, d} and B = { d, f, i}.

Intersection of two sets

Set with elements common to both sets

Denoted by ∩

A ∩ B = {d}



Set Operations

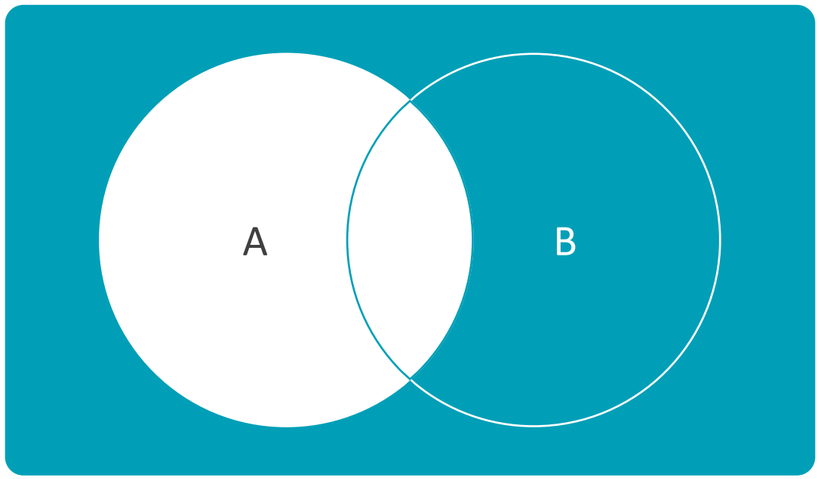
Consider sample space S = {a, c, d, f, g, i, k} and two subsets of S, A = {a, c, d} and B = { d, f, i}.

Complement of a set

Set whose elements are present in the sample space but not in the set

Denoted by the superscript ‘c’ or ^{c}

A^c = {f, g, i, k} and B^c = {a, c, g, k}



Measures of Shape (Skewness)

Measures of Shape

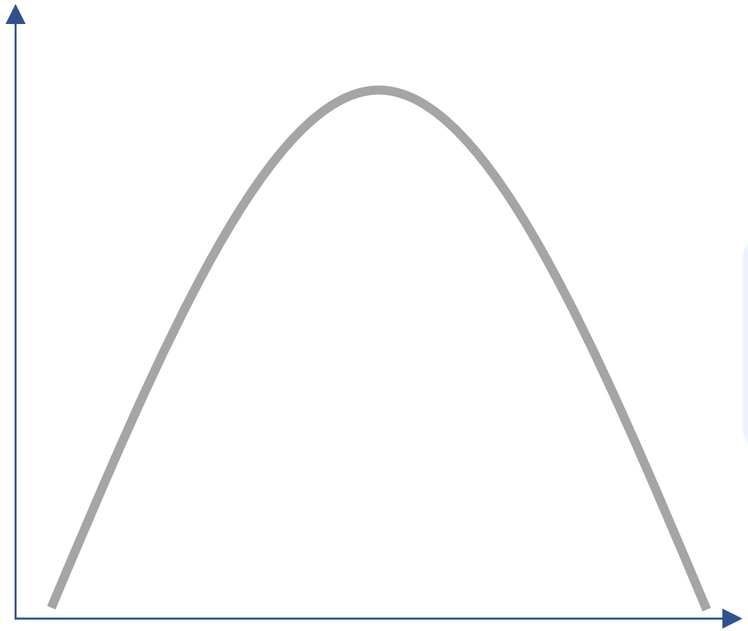
Measures of shape describe the distribution of data in a data set.

A histogram can describe the distribution shape as low and high values on the x-axis for numerical data but not for categorical data.

Although a histogram gives the overall shape, skewness and kurtosis are more precise quantitative measures of shape.

Skewness

Skewness is defined as the amount and direction of deviation from horizontal symmetry.



Any distribution is symmetric if it looks the same to the left and right of the center point.

In many statistical inferences, the distribution needs to be normal or nearly normal.

Skewness

Skewness is important as it helps test for normality.

In a normal distribution, skewness is 0.

In a nearly normal distribution, skewness is close to 0.

Formula for Skewness

For univariate data $x_1, x_2, \ldots\ldots, x_n$ skewness is given by:

$$g1 = (\sum_{i=1}^{N} (Xi - \overline{X})^3 / N) / s^3$$

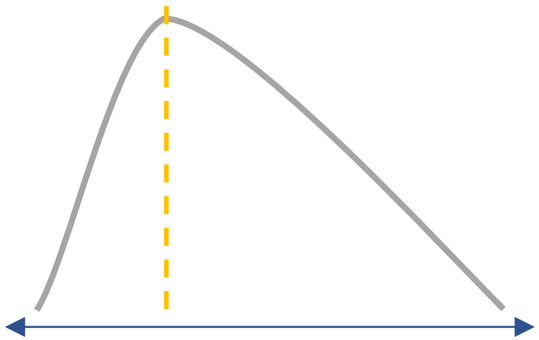
Where $\overline{X}$ is the mean

s is the standard deviation

N is the number of data points

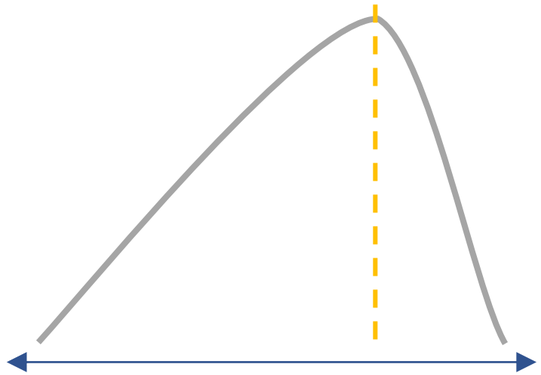
Skewness

Skewness for a normal distribution is 0, and symmetrical data has a near 0 skewness.



Negative skewness

It means the data is skewed left, i.e., the left tail is longer than the right tail.



Positive skewness

It implies it is skewed right, i.e., the right tail is longer than the left tail.

Skewness

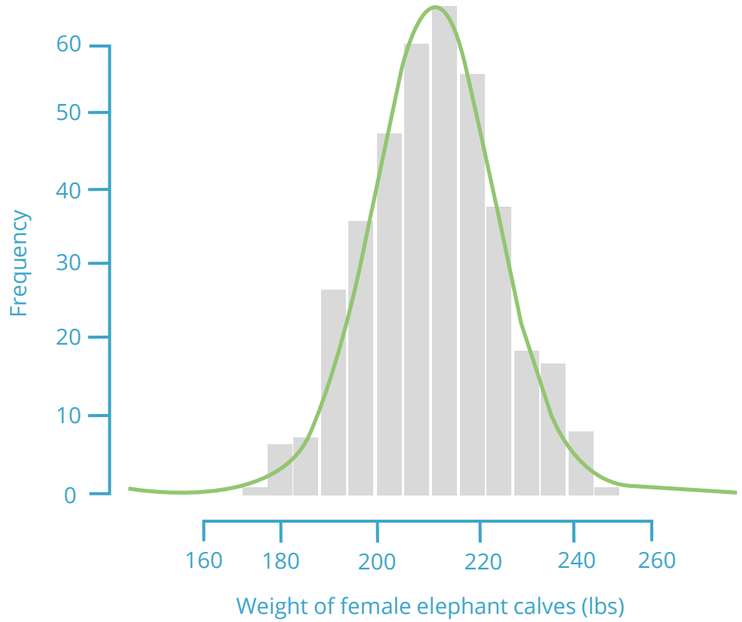
The Python implementation for skewness can be seen on the screen:

```
import numpy as np

from statsmodels.stats.stattools

import roburt_skewness

x = np.array([2,4,5,7,8,9,11,15])

skewness = medcouple(x)
```

Measures of Shape (Kurtosis)

Kurtosis

Kurtosis measures how heavy-tailed or light-tailed the distribution is relative to a normal distribution.



Data with high kurtosis tend to have heavy tails or outliers.

If kurtosis is low, there will be no outliers.

A uniform distribution is an extreme case of low kurtosis.

Formula for Kurtosis

For univariate data x_1, x_2, ..., x_n:

$$\text{Kurtosis} = \left(\sum_{i=1}^{N} (X_i - \overline{X})^4 / N\right) / s^4$$

Where $\overline{X}$ is the mean

s is the standard deviation

N is the number of data points

Formula for Kurtosis

Kurtosis for a normal distribution is 3.

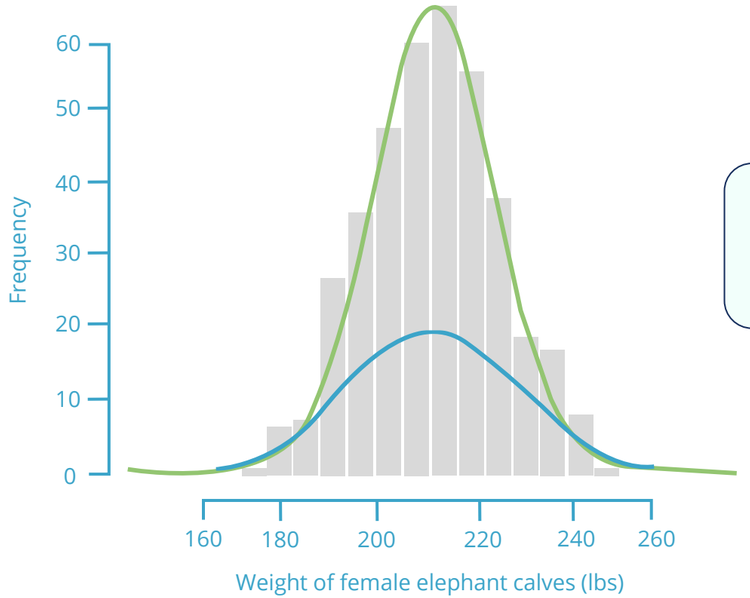
Sometimes the following definition of kurtosis, called excess kurtosis, is used.

$$\text{Kurtosis} = (\sum_{i=1}^{N} (X_i - \overline{X})^4 / N) / s^4 - 3$$

It is used so that the normal distribution has zero kurtosis.

Kurtosis

Positive excess kurtosis is called heavy-tailed distribution.



Negative excess kurtosis is called light-tailed distribution.

Kurtosis

The Python implementation for kurtosis can be seen below:

```
import numpy as np

from statsmodels.stats.stattools

import robust_kurtosis

x = np.array([2,4,5,7,8,9,11,15])

kurtosis  = robust_kurtosis(x)
```

Covariance and Correlation

Covariance and Correlation

Covariance and correlation measure the relationship or dependency between two variables.

Correlation gives both direction and strength of the linear relationship.

Covariance gives the direction of the linear relationship.

Therefore, correlation is a function of covariance.

Correlation values are standardized, while covariance values are not.

Covariance

If E[x] is the expected value or mean of sample x, the covariance of x and y is given by:

$$\begin{aligned} \text{cov}(x,y) &= E[(x - \mu_x)(y - \mu_y)] \\ &= E[xy] - E[x]\,E[y] \\ &= E[xy] - \mu_x \mu_y \end{aligned}$$

Covariance

The Python implementation of the covariance of two variables can be seen below:

```
meanx = sum(x)/float(len(x))

meany = sum(y)/float(len(x))

xpart = [i -meanx for i in x]

ypart = [i - meany for i in y]

numerator = sum([xpart[i]*ypart[i]

for i in range(len(xpart))])

denominator = len(x) -1

covariance = numerator/denominator
```

Here, pure Python is used instead of statsmodels as it makes more intuitive sense.

Correlation

A correlation coefficient is also called the Pearson correlation coefficient.

$$\text{Correlation coefficient of two variables} = \frac{\text{Covariance}}{\text{Product of their individual standard deviations}}$$

Since the standard deviation measures the absolute variability of the distribution, this division ensures that the correlation coefficient scales down to the -1 to +1 range.

Correlation

The formula for the correlation coefficient is:

$$\text{Corr}(x,y) = \text{Cov}(x,y) / S_x S_y$$

$$= E[(x-\mu_x)(y-\mu_y)] / S_x S_y$$

Here, S_x and S_y are standard deviations of x and y, respectively.

The closer it is to -1 or 1, the higher the correlation.

A positive correlation coefficient implies that when one variable increases, the other also increases, and vice versa.

Correlation

The Python code implementation of the correlation coefficient can be seen below:

```
correlation = covariance(x,y) /
(standard_deviation(x) *
standard_deviation(y))
```

Chi-Square Distribution Using Python

Chi-Square Distribution Using Python

The chi-square distribution is a probability distribution used to model the sum of the squares of independent standard normal variables.

Chi-square distribution can be applied in two steps:

1. Square v independent variables with standard normal distributions.
2. Add the results obtained in the previous step.

To calculate the chi-square, take the square of the difference between the observed and expected values and divide it by the expected value.

Chi-Square Distribution Using Python

The formula for the probability density function of the Chi-square distribution is:

$$f(x) = \left(e^{-x/2} x^{(v/2)-1}\right) / \left(2^{v/2}\Gamma(v/2)\right) \quad \text{for } x \geq 0$$

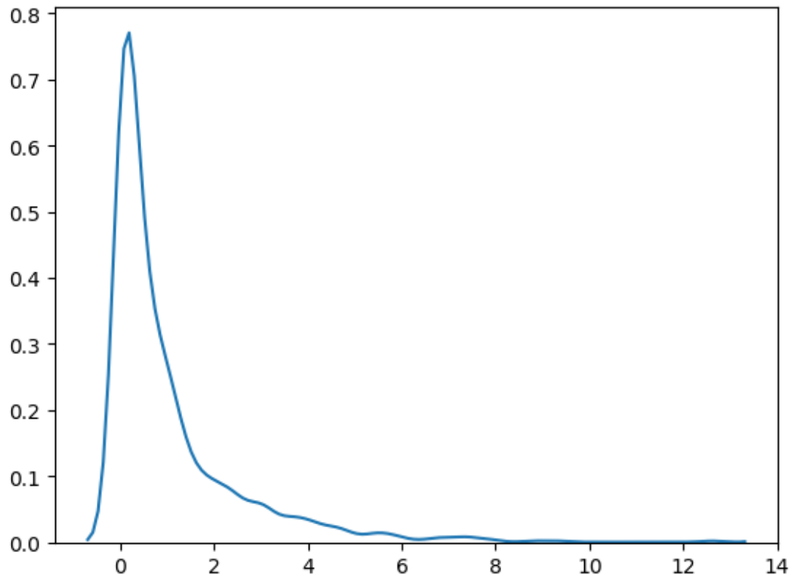
where v is the shape parameter and Γ is the gamma function. Γ is given by

$$\Gamma(a) = \int_0^\infty t^{a-1} e^{-t}\, dt$$

The Chi-square distribution is an asymmetric distribution with a minimum value of 0 and no maximum value.

Chi-Square Distribution Using Python

Its curve reaches a peak to the right of 0 and then gradually declines in height, proportional to the value of the Chi-square.



The curve approaches the x-axis but never quite touches it.

Chi-Square Distribution Using Python

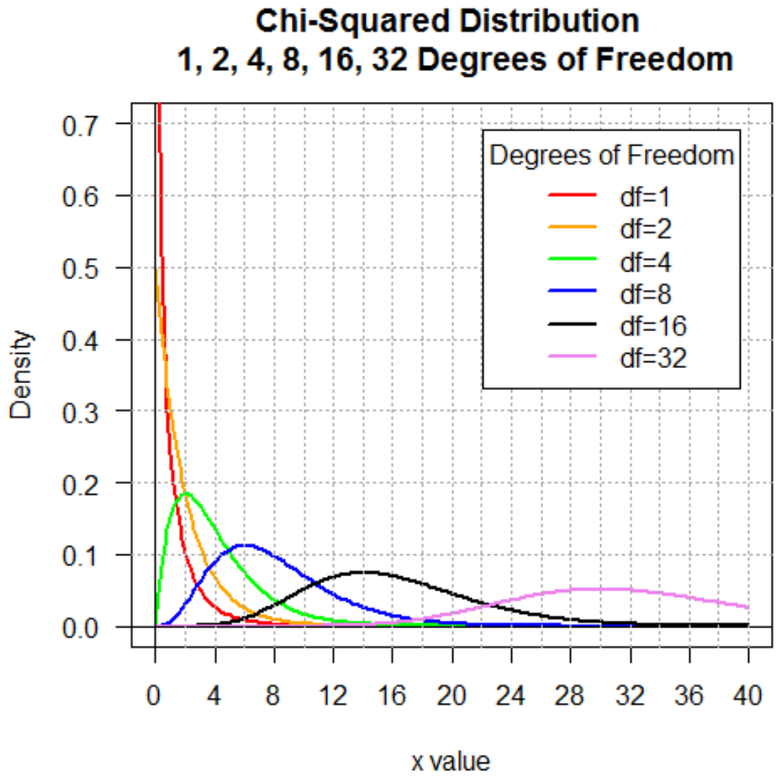
For each degree of freedom, there is a different Chi-square distribution.

Mean = Degrees of freedom

Standard deviation = Twice the degrees of freedom

Chi-Square Distribution Using Python

It is more spread out, with a peak farther to the right for larger than for smaller degrees of freedom.

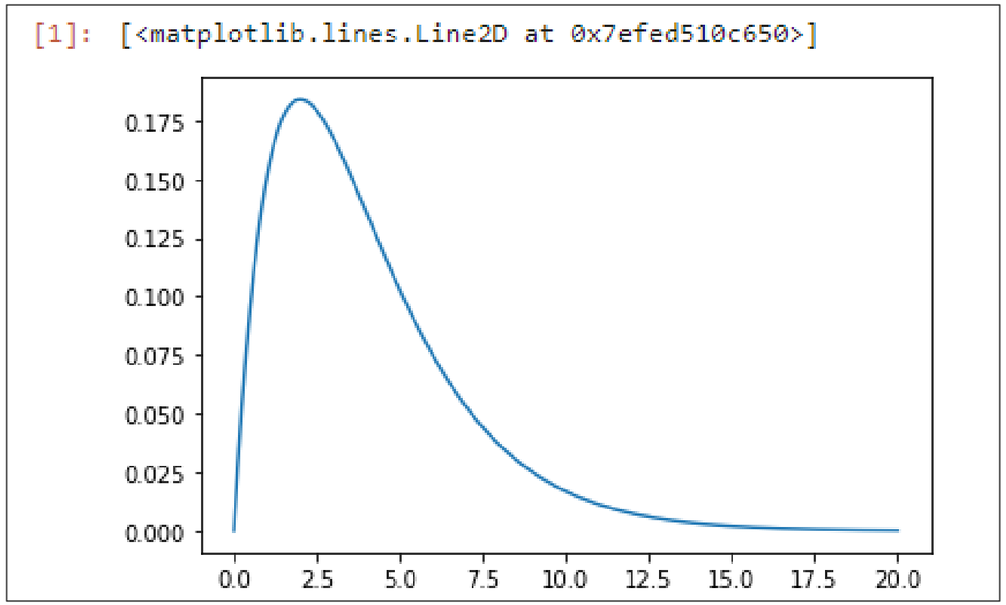


Therefore, for any given level of significance, the critical region begins at a larger Chi-square value, the larger the degree of freedom.

Chi-Square Distribution Using Python

Example: Plot a single chi-square distribution

```
import numpy as np

import matplotlib.pyplot as plt

from scipy.stats import chi2

#x-axis ranges from 0 to 20 with .001 steps

x = np.arange(0, 20, 0.001)

#plot Chi-square distribution with 4 degrees of
freedom

plt.plot(x, chi2.pdf(x, df=4))
```

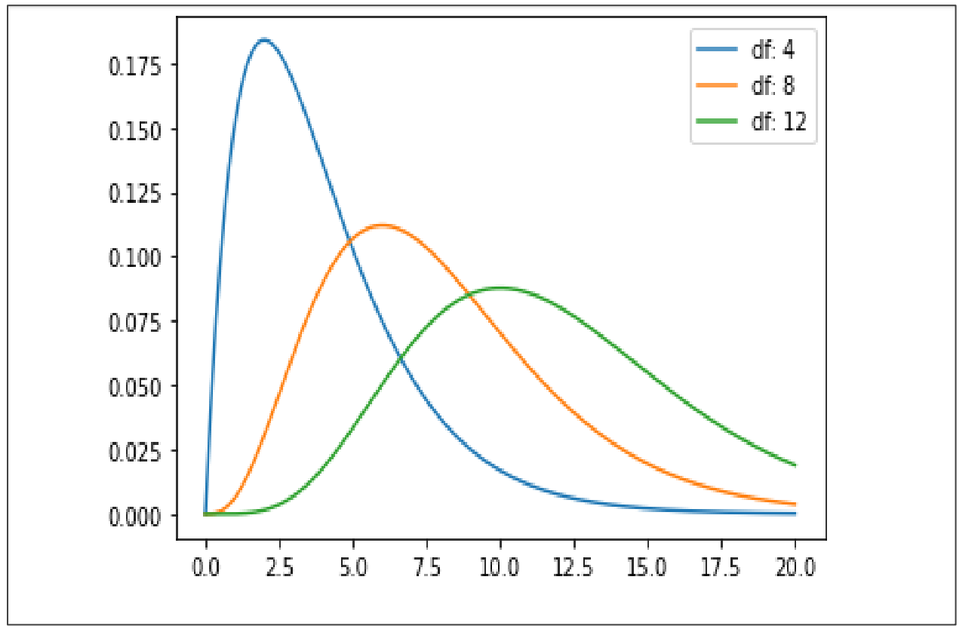


The code and output above demonstrate how to generate a single chi-square distribution curve with four degrees of freedom.

Chi-Square Distribution Using Python

Example: Plot multiple chi-square distributions

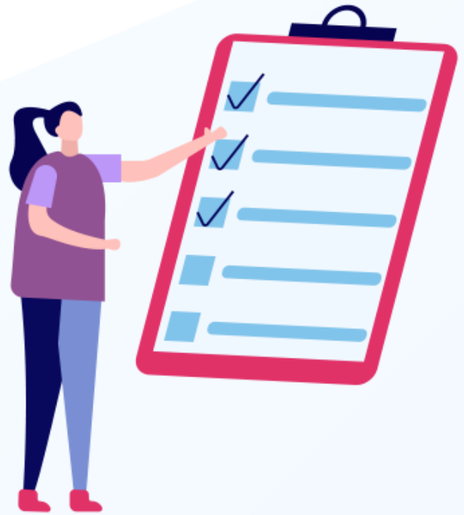
```
import numpy as np

import matplotlib.pyplot as plt

from scipy.stats import chi2

#x-axis ranges from 0 to 20 with .001 steps

x = np.arange(0, 20, 0.001) #define multiple Chi-square distributions

plt.plot(x, chi2.pdf(x, df=4), label='df: 4')

plt.plot(x, chi2.pdf(x, df=8), label='df: 8')

plt.plot(x, chi2.pdf(x, df=12), label='df: 12')  #add legend to plot

plt.legend()
```



The code and output above demonstrate how to plot multiple chi-square distribution curves with varying degrees of freedom.

Key Takeaways

- Data science is a combination of computer science and statistics.
- Statistics is vital for developing machine learning algorithms.
- There are three main categories of statistical theory: descriptive statistics, inferential statistics, and predictive statistics.
- Data is categorized into quantitative and qualitative data.
- There are four levels of measurement: nominal, ordinal, interval, and ratio.
- Mean, median, and mode are the three measures of central tendency.



Key Takeaways

- Measures of dispersion give an idea of data variability around the central point.
- Skewness is the amount and direction of deviation from horizontal symmetry.
- Covariance and correlation measure the relationship or dependency between two variables.
- Chi-square distribution is an asymmetric distribution with a minimum value of 0 and no maximum value.
- A set is a well-defined collection of objects, and every member of a set is called an element.

Knowledge Check

Knowledge Check

1

What is the difference between a parameter and a statistic?

A. A parameter is a statistical measure that describes the characteristic of the sample, while a statistic is a characteristic of the population that we want to estimate or test

B. A parameter is a characteristic of the sample that we want to estimate or test, while a statistic is a statistical measure that describes the characteristic of the population

C. A parameter is a statistical measure that describes the character of the population, while a statistic is a characteristic of the sample that usually helps us estimate or test the population parameter

D. A parameter and a statistic are the same thing

Knowledge Check

1

What is the difference between a parameter and a statistic?

A. A parameter is a statistical measure that describes the characteristic of the sample, while a statistic is a characteristic of the population that we want to estimate or test

B. A parameter is a characteristic of the sample that we want to estimate or test, while a statistic is a statistical measure that describes the characteristic of the population

C. A parameter is a statistical measure that describes the character of the population, while a statistic is a characteristic of the sample that usually helps us estimate or test the population parameter

D. A parameter and a statistic are the same thing

The correct answer is **C**

A parameter is a statistical measure that describes the character of the population to be estimated or tested, such as population mean. A statistic is a characteristic of the sample that usually helps us to estimate or test the population parameter, such as its mean.

Knowledge Check

2

What is the mode in statistics?

A. The middle value in a data set

B. The most frequently occurring data point in a data set

C. The average value in a data set

D. The difference between the largest and smallest data points in a data set

Knowledge Check

2

What is the mode in statistics?

A. The middle value in a data set

B. The most frequently occurring data point in a data set

C. The average value in a data set

D. The difference between the largest and smallest data points in a data set

The correct answer is **B**

Mode is the most frequently occurring data point in a data set.

Knowledge Check

3

What does a positive skewness value indicate?

A. The data is skewed left, i.e., the left tail is longer than the right tail.

B. The data is skewed right, i.e., the right tail is longer than the left tail.

C. The data has a near 0 skewness.

D. The data is symmetric.

Knowledge Check

3

What does a positive skewness value indicate?

A. The data is skewed left, i.e., the left tail is longer than the right tail.

B. The data is skewed right, i.e., the right tail is longer than the left tail.

C. The data has a near 0 skewness.

D. The data is symmetric.

The correct answer is **B**

A positive skewness value indicates that the data is skewed right, i.e., the right tail is longer than the left tail.

Thank You