

Introduction to Python Fundamentals and Statistics

Asst.Prof. Sachin Jaiswal,
Computer Science & Engineering



CHAPTER 2: Python Fundamentals and Statistics



2	Introduction to Python Fundamentals and Statistics: Introduction, Importance of Python, Levels of Data measurement, Central tendency and Dispersion, Distribution of Sample Means, Population and Variance, Confidence interval estimation	15	8
---	---	----	---

Population

In statistics, a population refers to the complete set of items or individuals that share at least one characteristic of interest. It represents the entire group that you want to draw conclusions about. Populations can be finite or infinite.

Examples:

All students in a university.

All households in a city.

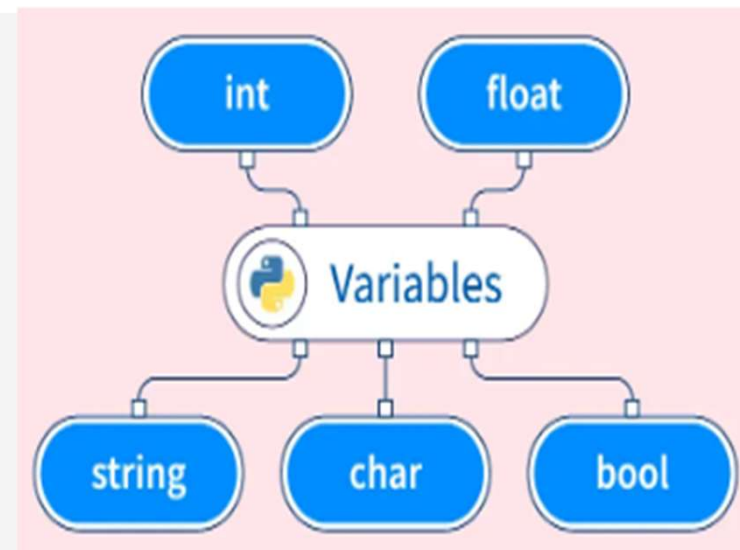
All manufactured products from a factory.

Python Fundamentals

1. Basic Syntax

• **Variables and Data Types:** Variables store data, and Python supports various data types like integers, floats, strings, and booleans.

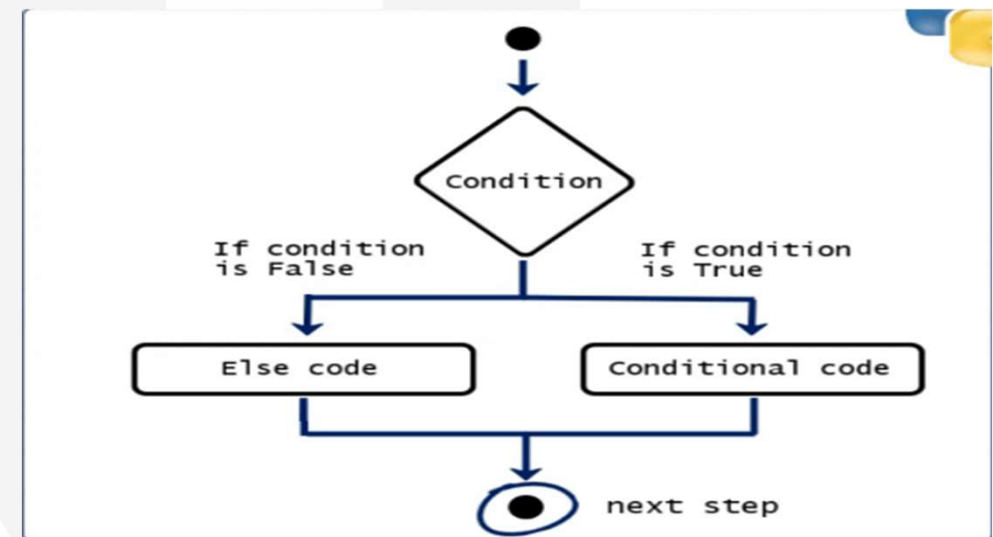
```
x = 5           # Integer
y = 3.14        # Float
name = "Alice"  # String
is_active = True # Boolean
```



2. Control Structures

•**Conditional Statements:** `if`, `elif`, and `else` are used for decision-making.

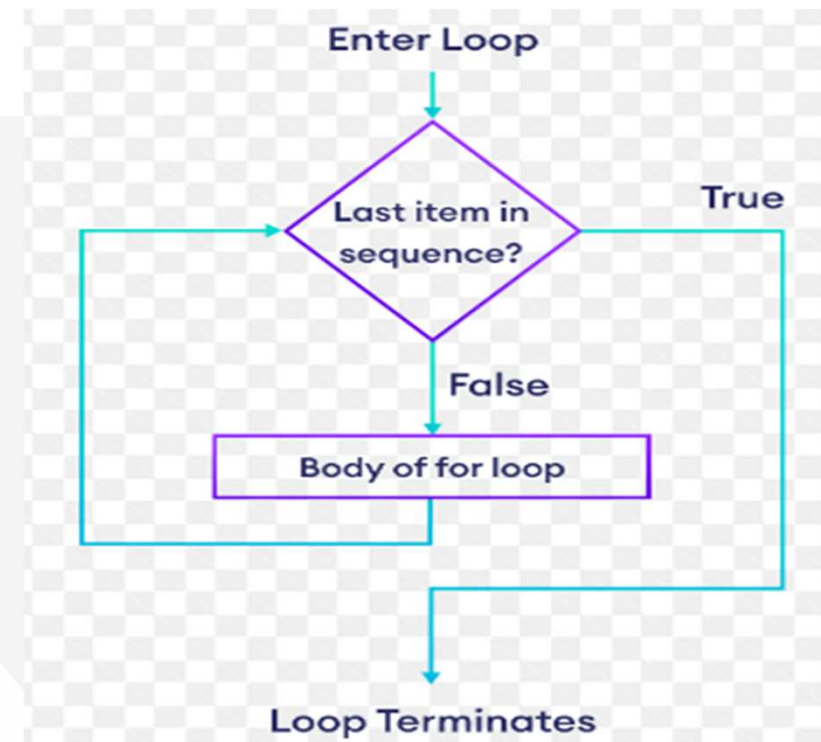
```
if x > 0:  
    print("x is positive")  
elif x == 0:  
    print("x is zero")  
else:  
    print("x is negative")
```



Loops:

for and while loops are used for iteration.

```
for i in range(5):  
    print(i)  
while x > 0:  
    print(x)  
    x -= 1
```



3. Functions:

Python Functions

In Python, the **function** is a block of code defined with a name

- A Function is a block of code that only runs when it is called.
- You can pass data, known as parameters, into a function.
- Functions are used to perform specific actions, and they are also known as methods.
- **Why use Functions?** To reuse code: we define the code once and use it many times.

```
def add(num1, num2):  
    print("Number 1:", num1)  
    print("Number 2:", num1)  
    addition = num1 + num2  
  
    return addition
```

Function Name Parameters

Function Body

Return Value

res = add(2, 4) → Function call
print(res)

Functions are defined using the `def` keyword and are used to encapsulate reusable code blocks.

```
def add(a, b):  
    return a + b
```

```
result = add(3, 4) # result is 7
```


4. Data Structures

Lists: Ordered, mutable collections.

```
fruits = ["apple", "banana", "cherry"]  
fruits.append("date")
```

Tuples: Ordered, immutable collections.

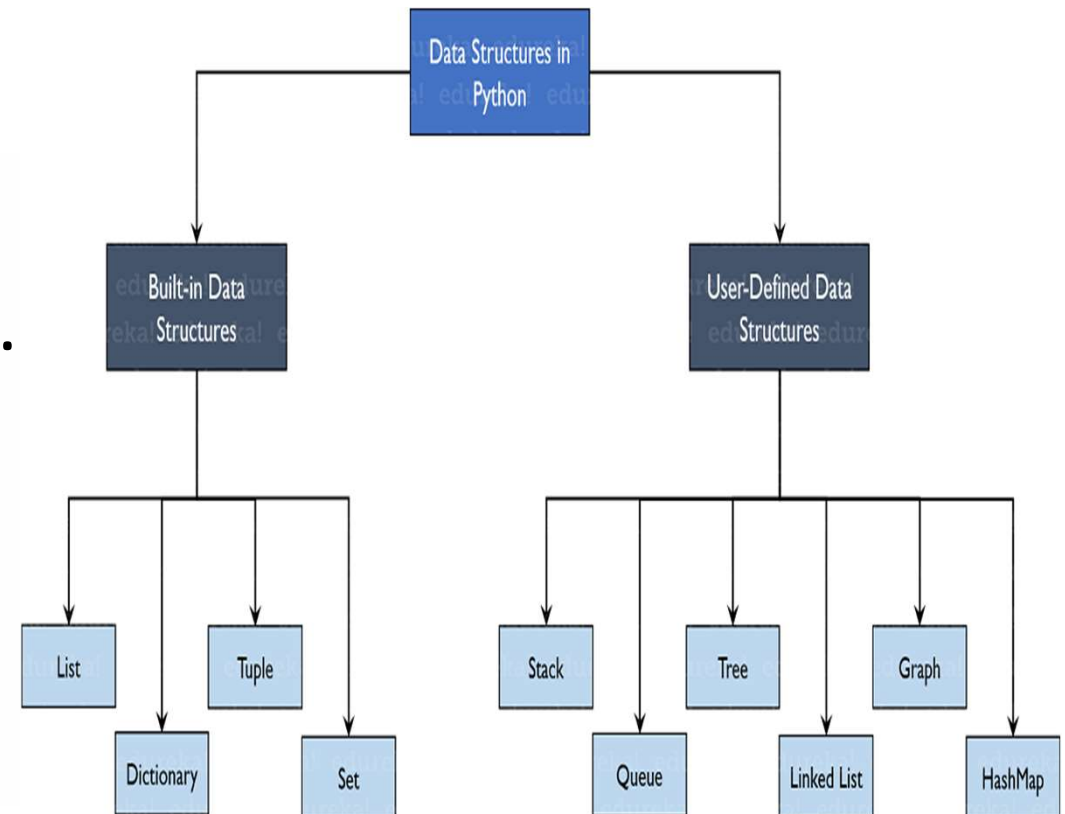
```
point = (3, 4)
```

Dictionaries: Unordered collections of key-value pairs.

```
student = {"name": "Alice", "age": 21}  
age = student["age"]
```

Sets: Unordered collections of unique elements.

```
unique_numbers = {1, 2, 3, 3, 4} # {1, 2, 3, 4}
```



5. Libraries

Python has a rich ecosystem of libraries for various tasks. For example, numpy and pandas are popular for data manipulation and analysis.

Top 10 Python Libraries

 Pandas Data analysis and manipulation	 NumPy Mathematical functions
 Matplotlib Data visualisations	 SeaBorn Data visualisations
 Tensorflow Machine Learning	 Keras Deep Learning
 SciPy Scientific computing	 PyTorch Machine Learning
 Scrapy Web crawling	 SQLModel Interact with SQL databases

 | DATA RUNDOWN



Python

Python Standard Libraries

math

math
statistics
random

File system

os.path
fileinput
gzip
zipfile

Data types

collections
array
datetime
calendar

Text processing

string
re
readline

File formats

configparser
csv

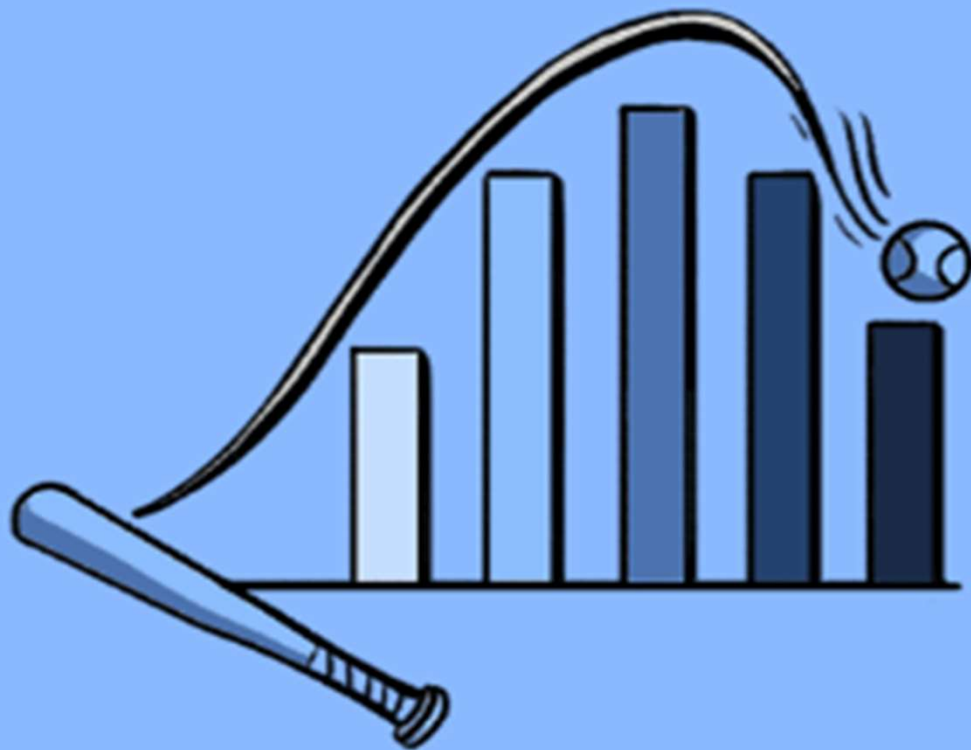
Operating System

platform
os
io

Introduction to Statistics



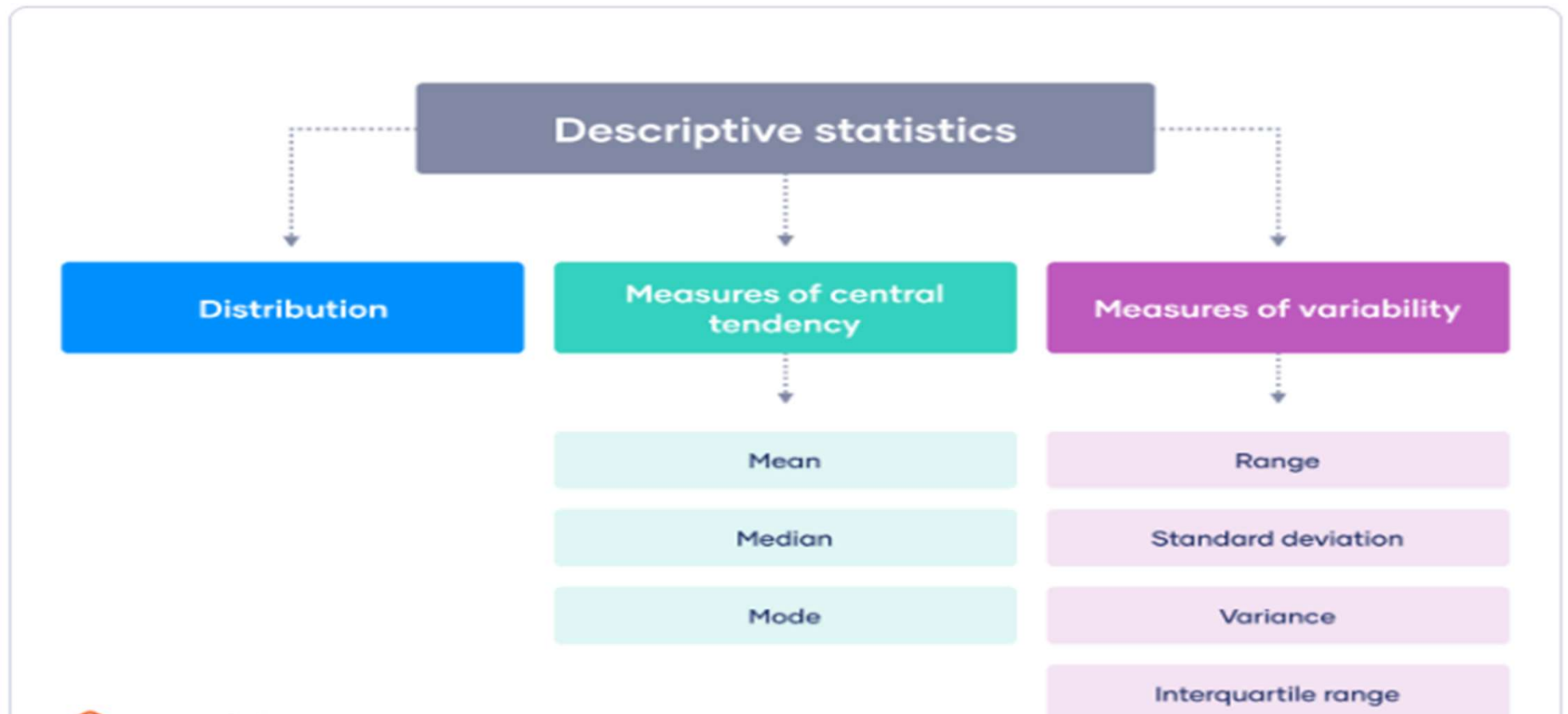
1.Descriptive Statistics



Descriptive Statistics

[di-'skrip-tiv stə-'ti-stiks]

Statistics that summarize or describe features of a data set, such as its central tendency or dispersion.



1. Descriptive Statistics

Mean: The average of a set of numbers.

```
import numpy as np  
data = [1, 2, 3, 4, 5]  
mean = np.mean(data) # 3.0
```

Median: The middle value in a set of numbers.

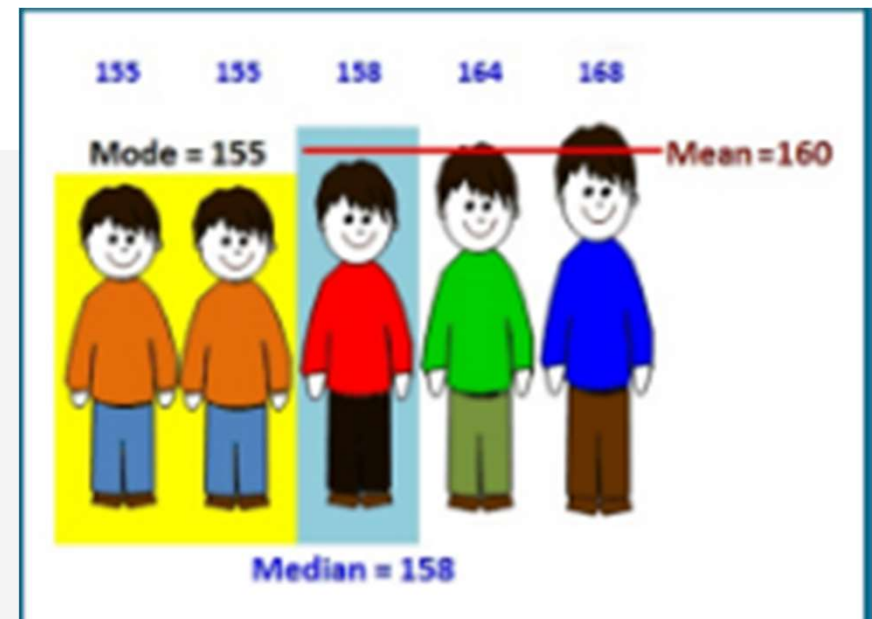
```
median = np.median(data) # 3
```

Mode: The most frequent value in a set of numbers.

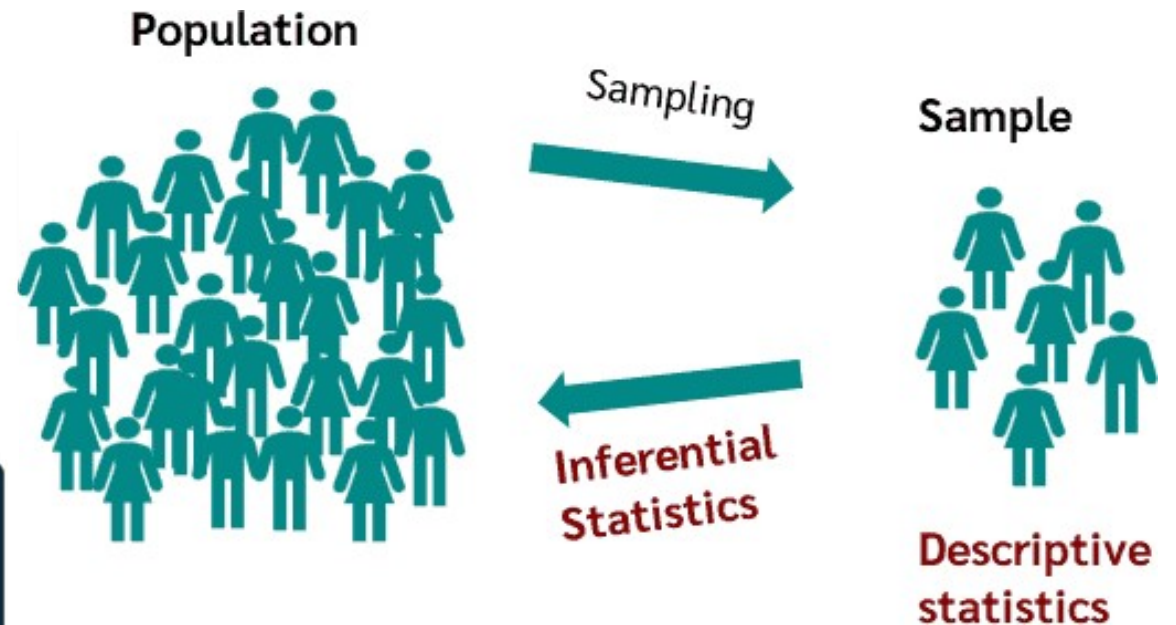
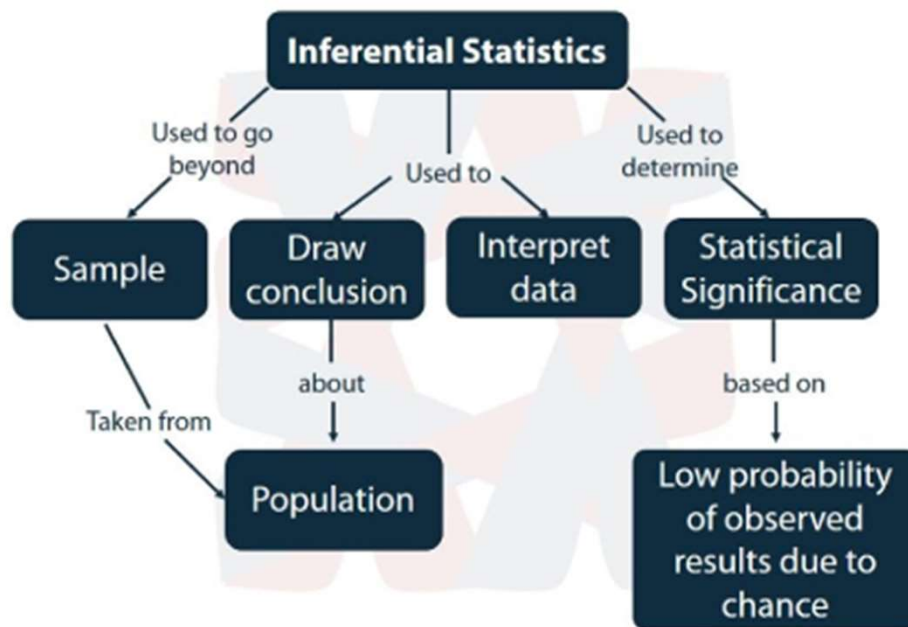
```
from scipy import stats  
mode = stats.mode(data) # 1
```

Standard Deviation: Measures the amount of variation or dispersion in a set of values.

```
std_dev = np.std(data) # 1.414...
```



2. Inferential Statistics



Importance of Python Programming

Python is a high-level, interpreted, interactive, and object-oriented scripting language. Python was designed to be highly readable which uses English keywords frequently whereas other languages use punctuation and it has fewer syntactical constructions than other languages.

It is used in :

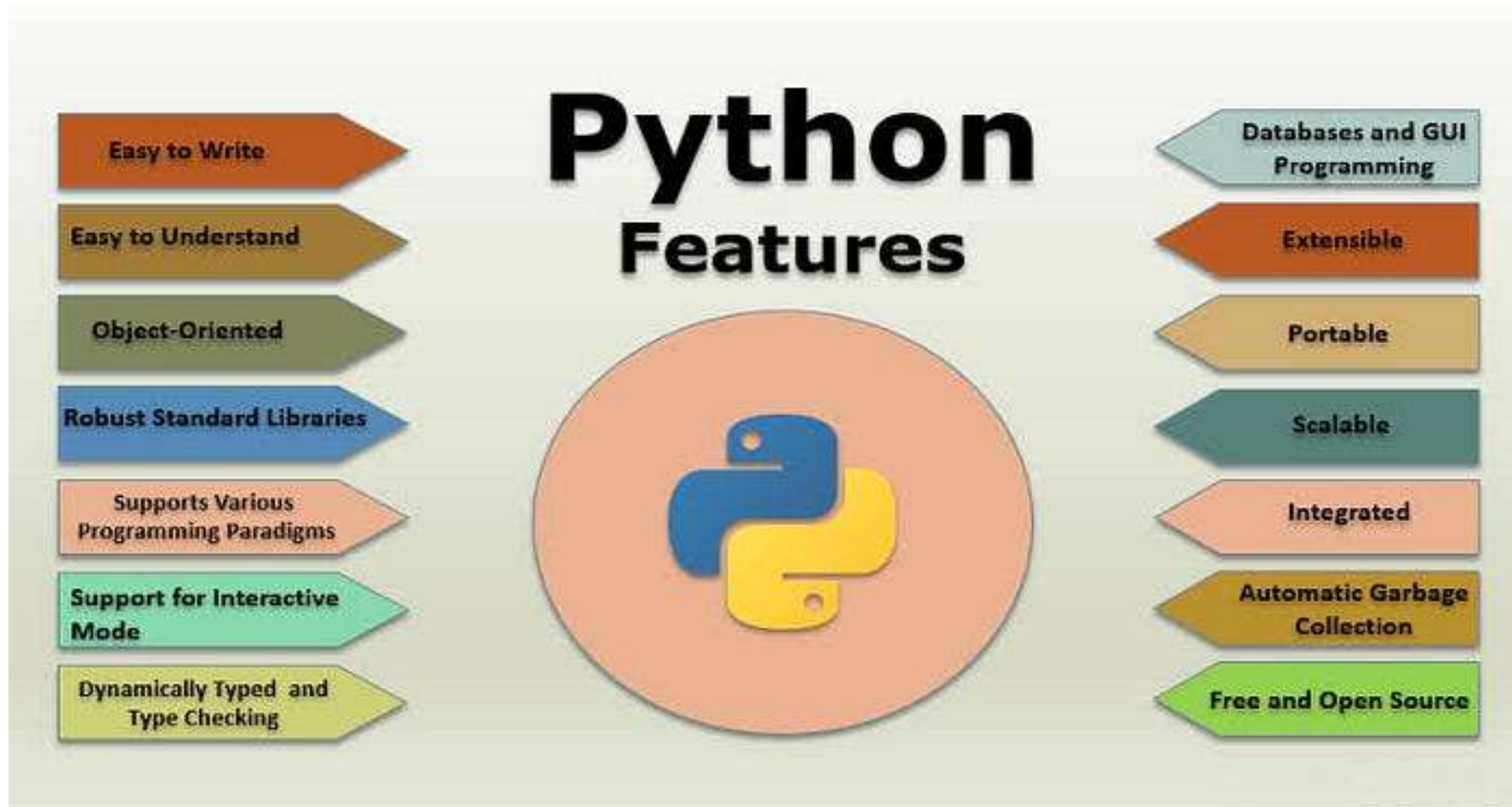
Software Development

Web Development

System Scripting

Mathematics

PYTHON FEATURES



Importance of Python:

Python is a highly versatile and powerful programming language that has gained immense popularity across various domains. Here are some key reasons highlighting the importance of Python:

1. Ease of Learning and Use

- **Simple Syntax:** Python's syntax is clear and readable, making it an excellent choice for beginners. Its design philosophy emphasizes code readability and simplicity.
- **Rapid Development:** The straightforward syntax allows developers to write less code, speeding up the development process.



Importance of Python:

2. Versatility and Flexibility

- **General-Purpose Language:** Python can be used for a wide range of applications, including web development, data analysis, artificial intelligence, scientific computing, and automation.
- **Platform Independence:** Python code can run on any operating system with a compatible interpreter, making it a cross-platform language.

3. Strong Community Support

- **Extensive Libraries and Frameworks:** Python has a vast ecosystem of libraries and frameworks that extend its capabilities. For example:
 - **Web Development:** Django, Flask
 - **Data Analysis:** Pandas, NumPy
 - **Machine Learning:** TensorFlow, PyTorch, scikit-learn
 - **Visualization:** Matplotlib, Seaborn
- **Active Community:** Python has a large and active community that contributes to its continuous improvement and provides support through forums, tutorials, and documentation.



Importance of Python:

4. Relevance in Data Science and AI

- **Data Analysis and Visualization:** Python's libraries like Pandas, NumPy, Matplotlib, and Seaborn make it a preferred language for data manipulation and visualization.
- **Machine Learning and AI:** Python is a leading language in the AI and machine learning community, with powerful libraries such as TensorFlow, Keras, and PyTorch.

5. Industry Adoption

- **Enterprise Use:** Many large corporations (e.g., Google, Facebook, Amazon) use Python for various applications, demonstrating its reliability and scalability.
- **Startups and Academia:** Python's efficiency and versatility make it a popular choice among startups and academic researchers.



Importance of Python:

6. Automation and Scripting

- **Automation:** Python is widely used for automating repetitive tasks, such as file manipulation, web scraping, and task scheduling.
- **Scripting:** Its ease of use makes Python an ideal language for writing scripts to automate workflows.

7. Integration Capabilities

- **Interoperability:** Python can easily integrate with other languages and technologies, such as C/C++, Java, and .NET, facilitating its use in diverse environments.
- **API Interactions:** Python's libraries allow for seamless interaction with various APIs, enhancing its functionality in web and software development.



Importance of Python:

8. Educational Use

- **Teaching and Learning:** Python's simplicity makes it an excellent introductory language for computer science and programming courses.
- **Interactive Learning:** Tools like Jupyter Notebooks provide an interactive environment for learning and experimenting with Python code, especially in data science.

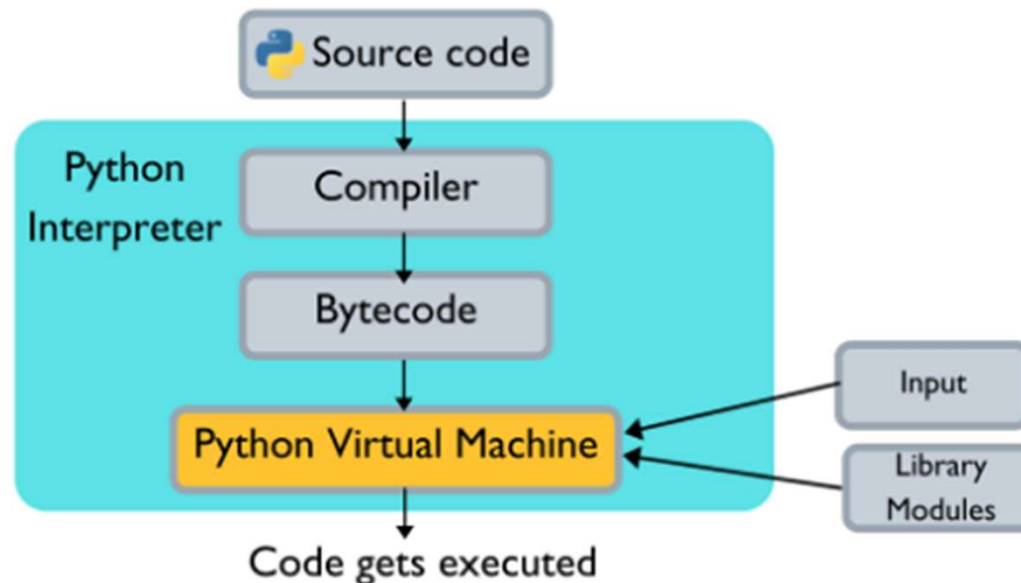
9. Open Source Nature

- **Cost-Effective:** Python is open-source and free to use, which makes it accessible to a wide range of users and organizations without licensing costs.
- **Community Development:** Being open-source, Python benefits from contributions from developers worldwide, ensuring continuous enhancement and adaptation to new technologies.

In summary, Python's importance lies in its simplicity, versatility, extensive support libraries, and strong community backing, making it a powerful tool for a wide array of applications in today's technology-driven world.

Python is Interpreted

It means that each line is processed one by one at runtime by the interpreter and you do not need to compile your program before executing it.



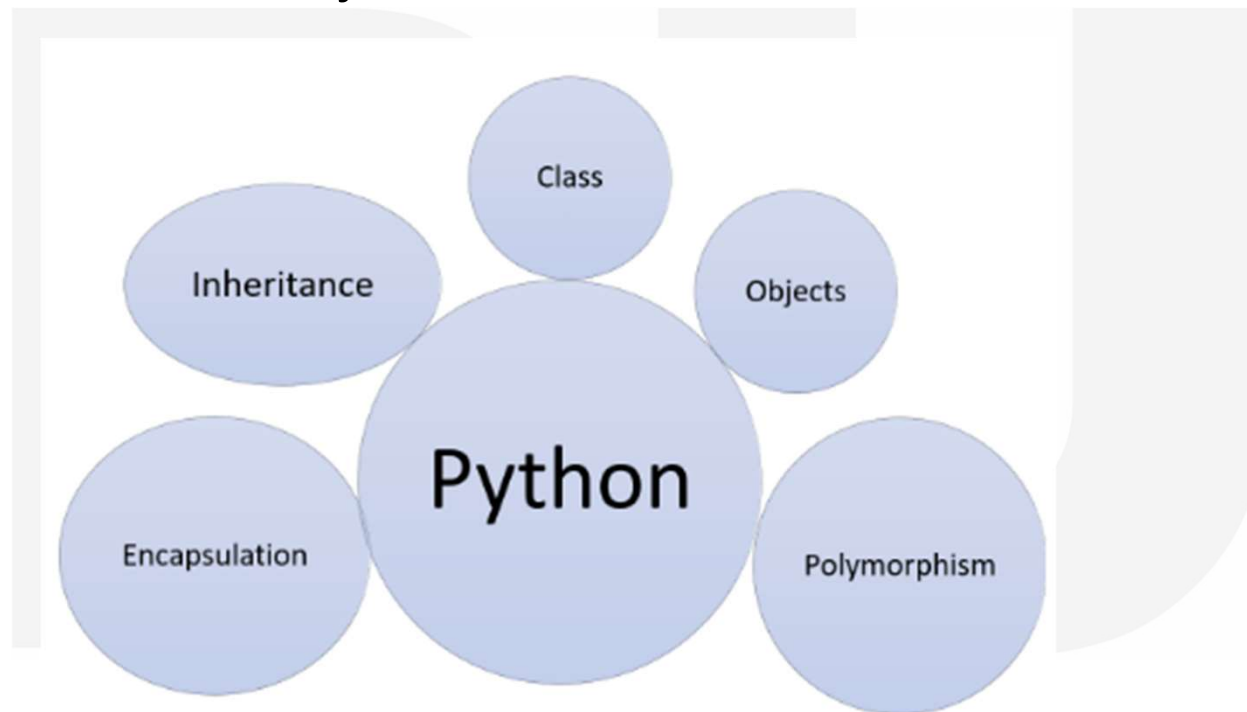
Python is Interactive

It means that you can actually sit at a Python prompt and interact with the interpreter directly, to write and execute your programs.

A screenshot of the Python IDLE Shell 3.10.0 window. The window has a menu bar with 'File', 'Edit', 'Shell', 'Debug', 'Options', 'Window', and 'Help'. The main text area shows the Python 3.10.0 startup message: 'Python 3.10.0 (tags/v3.10.0:b494f59, Oct 4 2021, 19:00:18) [MSC v.1929 64 bit (AMD64)] on win32' followed by 'Type "help", "copyright", "credits" or "license()" for more information.' The prompt '>>>' is visible on the left, and the status bar at the bottom right shows 'Ln: 3 Col: 0'.

Python is Object-Oriented

Python supports the Object-Oriented style or technique of programming that encapsulates code within objects.





Python is Beginner's Language

Python is an excellent language for beginning programmers and facilitates the construction of a wide range of programs ranging from simple text processing to web browsers to games. Python does not have pointers, which is one of the main challenges that many of us have encountered when programming.

Easy-to-maintain

Python's success is that its source code is fairly easy-to-maintain. One reason for that is, it is read and written like a lot of everyday English.



A Broad Standard Library

One of Python's greatest strengths is the bulk of the library, which makes it very portable and cross-platform compatible. Python has libraries for almost everything one can think of.

The **Python** scientific software ecosystem:



Portable

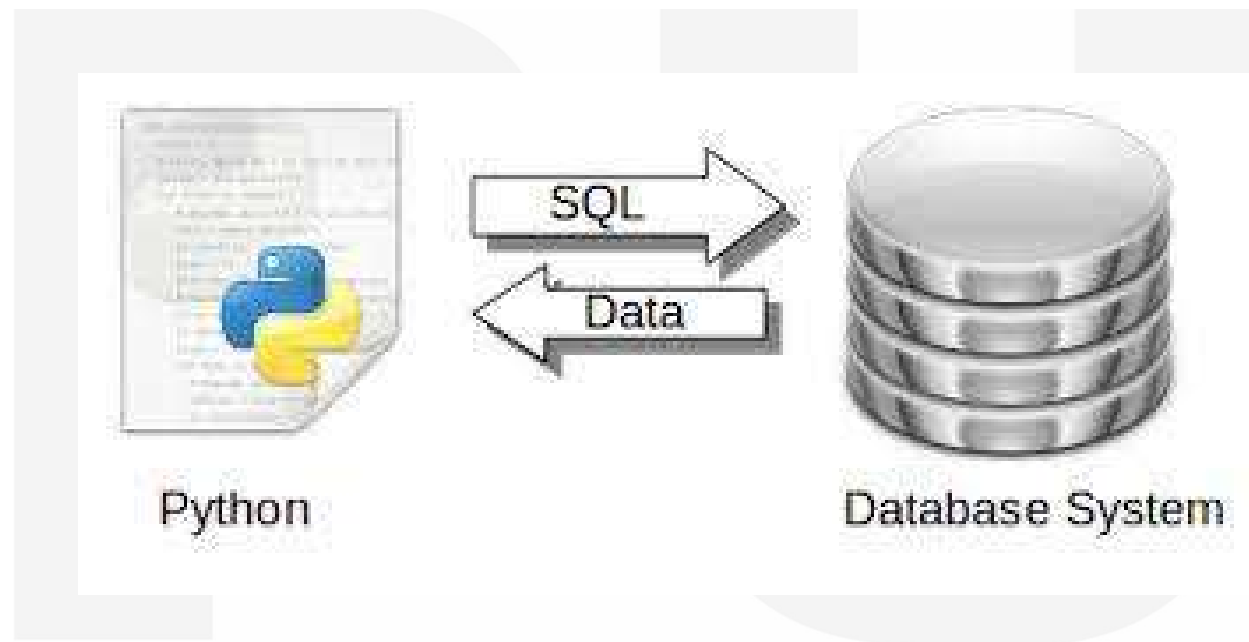
Python can run on a wide variety of hardware platforms and has the same interface on all platforms. You can run the same python program on Windows, Linux, Mac, Raspberry Pi, Mango Pi, Android, etc.

Extendable

You can add low-level modules to the Python interpreter. These modules enable programmers to add to or customize their tools to be more efficient. Generally, we do that using the PIP command.

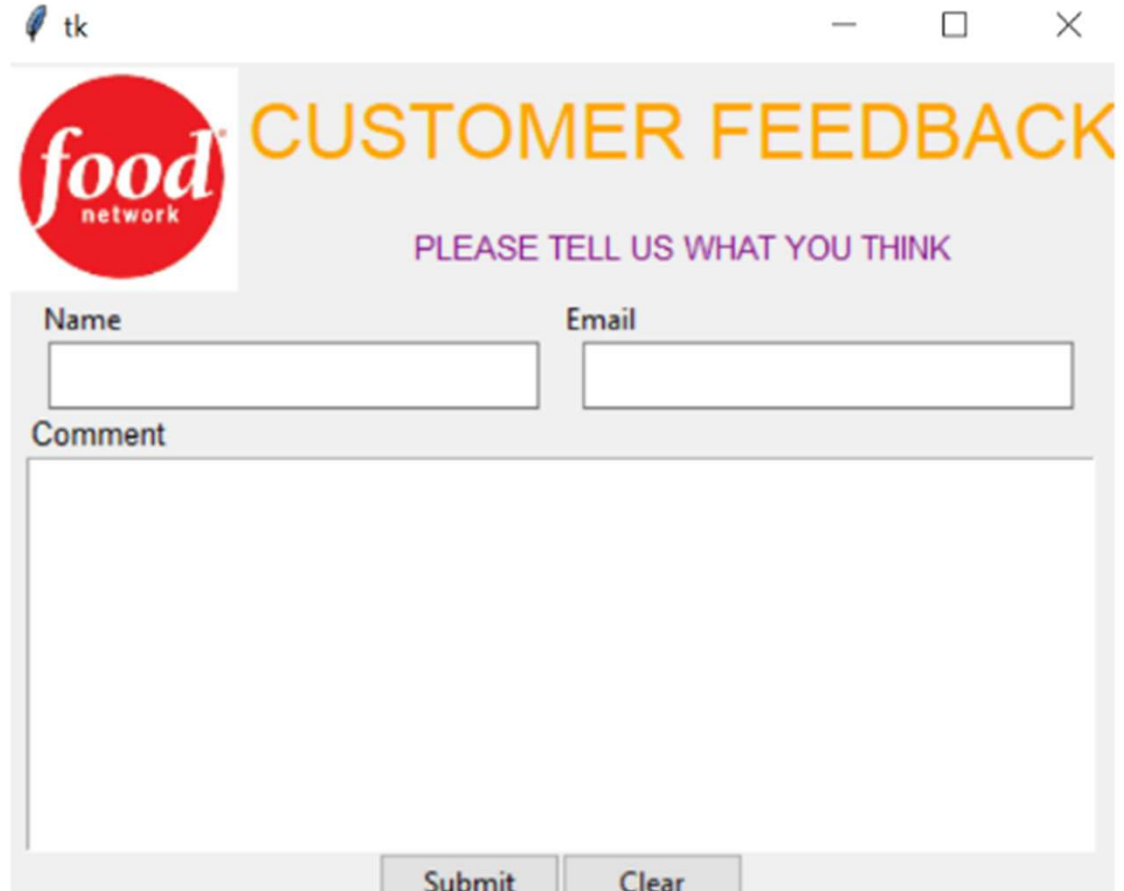
Databases

Python provides interfaces to all major commercial databases. It has packages to communicate with SQL, NoSQL, etc. databases, ranging from MongoDB to MySQL.



GUI Programming

Python supports GUI applications that can be created and ported to many system calls, libraries, and windows systems, such as Windows, Macintosh, and the X Window system of Unix. It has libraries like Tkinter, WxPython, etc.

A screenshot of a Tkinter window titled 'food network CUSTOMER FEEDBACK'. The window has a red circular logo with 'food network' text. Below the logo, the text 'PLEASE TELL US WHAT YOU THINK' is displayed in purple. There are two input fields for 'Name' and 'Email', and a large text area for 'Comment'. At the bottom, there are 'Submit' and 'Clear' buttons.

tk

food network **CUSTOMER FEEDBACK**

PLEASE TELL US WHAT YOU THINK

Name

Email

Comment

Submit Clear

Scalable

Python provides a better structure and support for large programs than shell scripting.

Apart from the above-mentioned features, Python has a big list of good features, few of them are-

Support for functional and structured programming methods as well as OOP.

It can be used as a scripting language or can be compiled to byte-code for building large applications.

Very high-level dynamic data types and supports dynamic type checking.

Supports automatic garbage collection.

It can be easily integrated with C, C++, COM, ActiveX, CORBA, and Java.



Levels of Data measurement:

In statistics and research, levels of data measurement refer to the ways in which data can be categorized, counted, and measured. Understanding these levels helps in choosing the appropriate statistical methods for analysis. There are four main levels of data measurement: nominal, ordinal, interval, and ratio.

1. Nominal Level

- **Definition:** This is the most basic level of measurement, where data are categorized without any order or ranking.
- **Characteristics:**
 - Categories are mutually exclusive and exhaustive.
 - No inherent order or ranking between the categories.
- **Examples:**
 - Gender (male, female)
 - Eye color (blue, green, brown)
 - Types of cuisine (Italian, Chinese, Mexican)



Levels of Data measurement:

2. Ordinal Level

- **Definition:** At this level, data can be categorized and ranked, but the intervals between the ranks are not equal or known.
- **Characteristics:**
 - Categories are mutually exclusive and exhaustive.
 - There is a meaningful order or ranking between categories.
 - The difference between ranks is not quantified.
- **Examples:**
 - Education level (high school, bachelor's, master's, doctorate)
 - Customer satisfaction ratings (very dissatisfied, dissatisfied, neutral, satisfied, very satisfied)
 - Class rankings (first, second, third)

3. Interval Level

- **Definition:** This level involves data that can be ordered and where the difference between values is meaningful and consistent, but there is no true zero point.
- **Characteristics:**
 - Categories are mutually exclusive and exhaustive.
 - There is a meaningful order or ranking between categories.
 - The intervals between values are consistent and measurable.
 - No true zero point (zero does not represent an absence of the attribute).
- **Examples:**
 - Temperature in Celsius or Fahrenheit
 - IQ scores
 - SAT scores

4. Ratio Level

- **Definition:** The highest level of measurement, where data can be categorized, ranked, and have consistent intervals, with a meaningful zero point.
- **Characteristics:**
 - Categories are mutually exclusive and exhaustive.
 - There is a meaningful order or ranking between categories.
 - The intervals between values are consistent and measurable.
 - There is a true zero point, indicating the absence of the attribute.
- **Examples:**
 - Height (in centimeters or inches)
 - Weight (in kilograms or pounds)
 - Age
 - Income

Summary Table

Level	Order	Equal Intervals	True Zero	Examples
Nominal	No	No	No	Gender, eye color, types of cuisine
Ordinal	Yes	No	No	Education level, satisfaction ratings, class rankings
Interval	Yes	Yes	No	Temperature, IQ scores, SAT scores
Ratio	Yes	Yes	Yes	Height, weight, age, income

Understanding these levels of measurement is crucial for selecting the correct statistical techniques and ensuring the validity of your data analysis.

Central tendency and Dispersion:

Central tendency and dispersion are fundamental concepts in statistics that describe the characteristics of a dataset. Central tendency measures the center or typical value of a dataset, while dispersion measures the spread or variability of the data.

Central Tendency

Central tendency provides a single value that represents the center point or typical value of a dataset. The most common measures of central tendency are the mean, median, and mode.

1. Mean

- Definition:** The arithmetic average of all the values in a dataset.

- Calculation:** Sum of all values divided by the number of values.

```
pythonCopy codemean = sum(data) / len(data)
```

- Example:**

```
pythonCopy codedata = [1, 2, 3, 4, 5]mean = sum(data) / len(data) # mean = 3
```

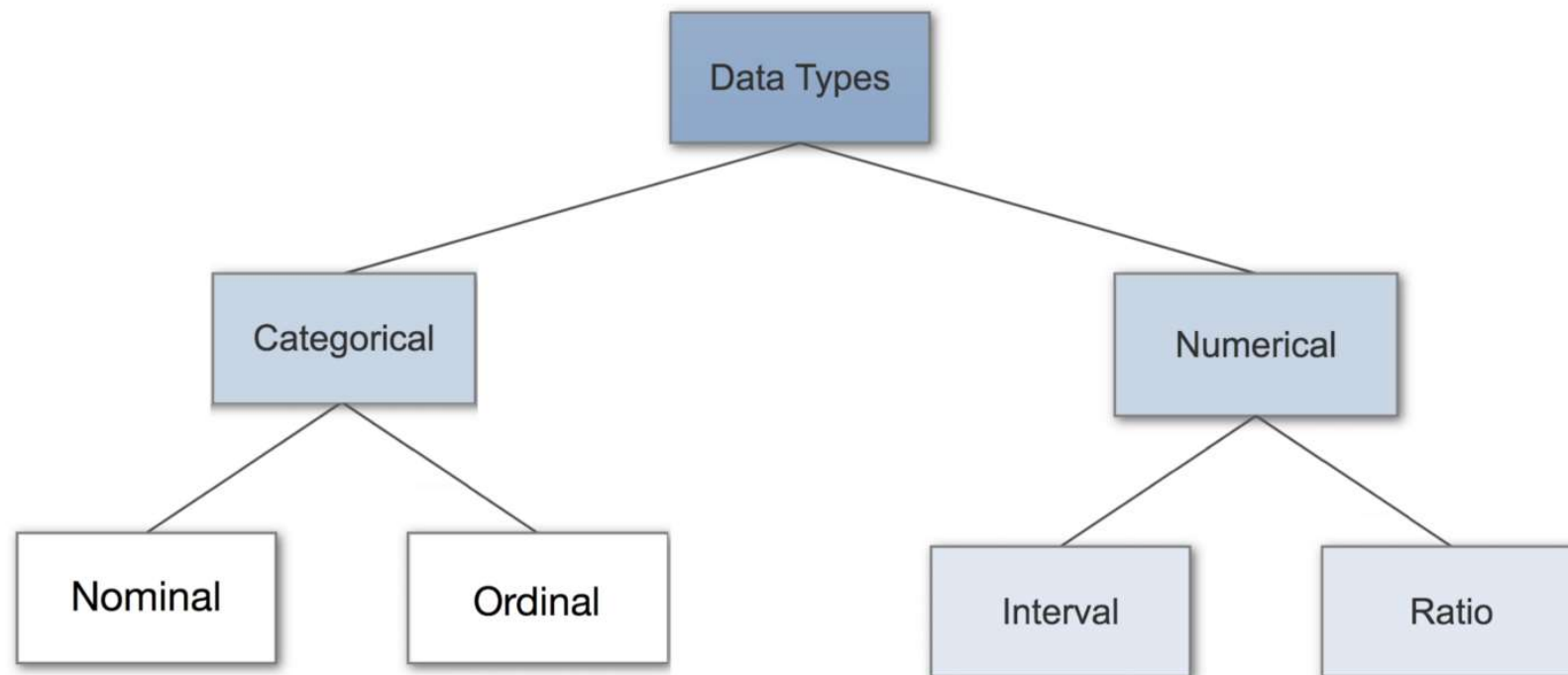
2. Median

•**Definition:** The middle value in a dataset when the values are arranged in ascending or descending order.

•**Calculation:**

- For an odd number of values: Middle value.
- For an even number of values: Average of the two middle values.

level of data measurement?



Categorical data:-

In categorical data we see the data which have a defined category.

For example: Marital Status, Political Party, Eye color.

- Categorical data represents characteristics.
- Therefore it can represent things like a person's gender, language etc.
- Categorical data can also take on numerical values.
- Example: 1 for female and 0 for male. Note that those numbers don't have mathematical meaning.

Nominal

1st Level of **Measurement**

Ordinal

2nd Level of **Measurement**

Interval

3rd Level of **Measurement**

Ratio

4th Level of **Measurement**

Examples of nominal features:-

What is your Gender?

- ☐ Female
- ☐ Male

What languages do you speak?

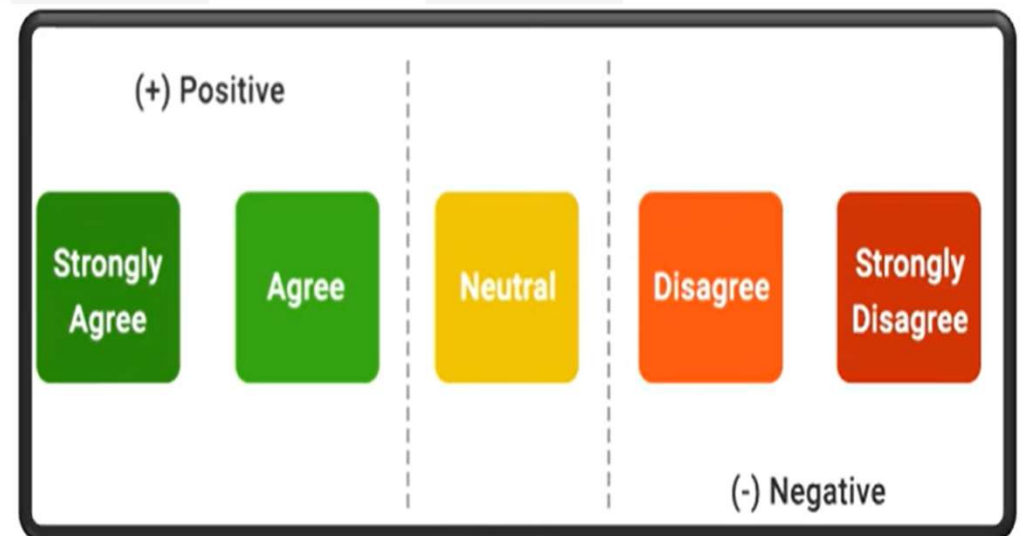
- ☐ Englisch
- ☐ French
- ☐ German
- ☐ Spanish

Ordinal:- An ordinal scale classifies data into distinct categories during which ranking is implied.

- Ordinal' sounds similar to 'Order', which is exactly the purpose of this scale.
- Ordinal' scale is one where the order matters but not the difference between the values.
- Ordinal scale can be presented in Tabular or Graphical form.

For example:

- Faculty rank: Professor, Associate Professor, Assistant Professor
- Students grade: A,B,C,D,E,F.
- Examples:



Discrete Data:-

- We speak of discrete data if its values are distinct and separate. In other words: We speak of discrete data if the data can only take on certain values.

This type of data can't be measured but it can be counted. It basically represents information that can be categorized into a classification. An example is the number of heads in 100 coin flips.

Continuous Data:-

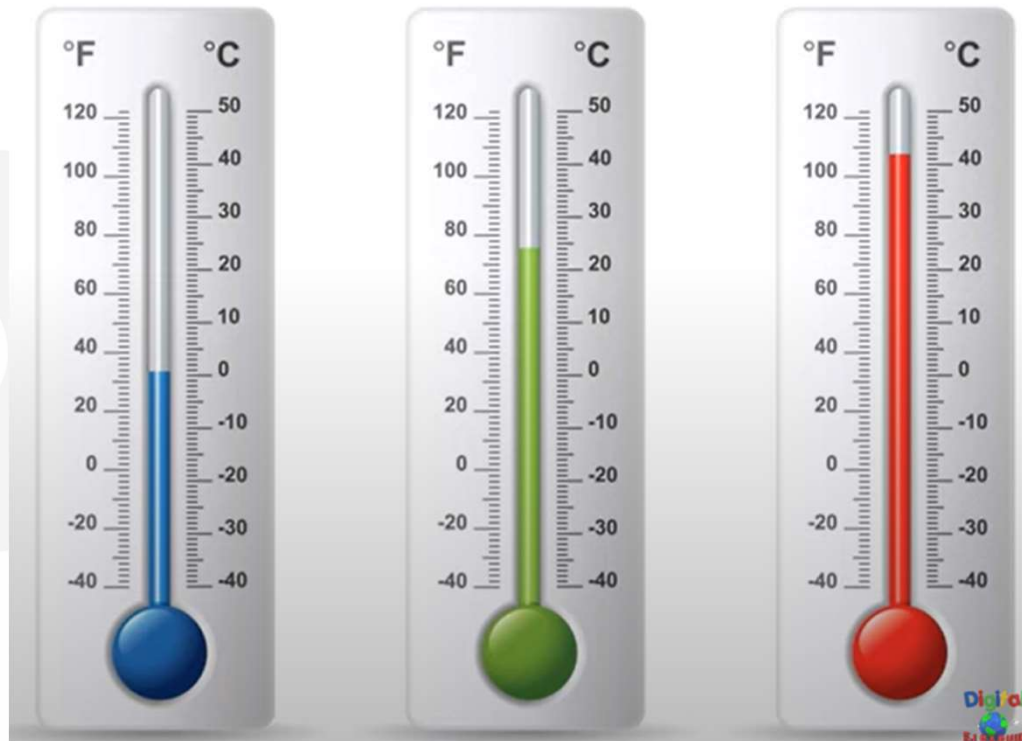
- **Continuous Data** represents measurements and therefore their values can't be counted but they can be measured. An example would be the height of a person, weight, voltage which you can describe by using intervals on the real number line.

Interval Data:-

- Interval itself means “**Space in Between**”.
- Interval scale are numerical values in which we know and the exact differences between the values.
- **Drawback** : No pre-decided starting point or a true zero value.

Temperature in Fahrenheit and Celsius, Years

Interval Data:-



**10 degrees C + 10 degrees C = 20 degrees C.
But 20 degrees C is not as hot as 10 degrees C**

Temperature?

- ☐ - 10
- ☐ -5
- ☐ 0
- ☐ + 5
- ☐ + 10
- ☐ + 15

Ratio Data:-

- Ratio values are also ordered units that have the same difference. Ratio values are the same as interval values, with the difference that they do have an absolute zero. Good examples are height, weight, length etc.
- Ratio scales not only produces order of variables but also make difference between variables.
- Ratio scales has all the properties of Interval Scale and **Absolute Or True Zero**.
- **Examples:**



Nominal :-

Nominal data are recorded as categories. For this reason, nominal data is also known as categorical data.

Note that nominal data that has no order. Therefore if you would change the order of its values, the meaning would not change.

- Nominal scale describes a variable that **do not have a natural order or ranking**.
- Nominal scale is a naming scale, where variables are simply “**named**” or **labeled**, with no specific order.
- Nominal data can be **Qualitative and Quantitative** both.

level of data measurement?

The levels of data measurement, also known as scales of measurement, are fundamental to understanding how data can be analyzed and interpreted.

There are four primary levels:

Nominal:

Description: Categorizes data without any order. The data can only be classified into categories.

Examples: Gender (male, female), Hair color (blonde, brunette, redhead), Types of cuisine (Italian, Chinese, Mexican).

Ordinal:

Description: Categorizes data with a meaningful order, but the intervals between the categories are not necessarily equal.

Examples: Class ranks (1st, 2nd, 3rd), Levels of satisfaction (satisfied, neutral, unsatisfied), Military ranks (private, corporal, sergeant).

Interval:

Description: Measures data with equal intervals between values, but there is no true zero point. The difference between values is meaningful.

Examples: Temperature in Celsius or Fahrenheit, IQ scores, Dates in calendar years.

Ratio:

Description: Measures data with equal intervals and a true zero point, allowing for the comparison of absolute magnitudes.

Examples: Height, Weight, Age, Income, Duration (time).

Understanding these levels is crucial for selecting appropriate statistical methods for analysis, as some methods are only suitable for certain types of data.

Measures of central tendency

Definition

A measure of central tendency (also referred to as measures of centre or central location) is a summary measure that attempts to describe a whole set of data with a single value that represents the middle or centre of its distribution.

There are three main measures of central tendency:

mode

median

Mean

Each of these measures describes a different indication of the typical or central value in the distribution.

Mode

The mode is the most commonly occurring value in a distribution.

Consider this dataset showing the retirement age of 11 people, in whole years:

54, 54, 54, 55, 56, 57, 57, 58, 58, 60, 60

This table shows a simple frequency distribution of the retirement age data.

Frequency distribution table

The most commonly occurring value is 54,
therefore the mode of this distribution is 54 years.

Age	Frequency
54	3
55	1
56	1
57	2
58	2
60	2

Advantage of the mode

The mode has an advantage over the median and the mean as it can be found for both numerical and categorical (non-numerical) data.

Limitations of the Mode

Centre of Distribution:

In some distributions, the mode may not reflect the centre of the distribution accurately.

Example: When the distribution of retirement age is ordered from lowest to highest, the centre is 57 years, but the mode is 54 years.

Distribution: 54, 54, 54, 55, 56, 57, 57, 58, 58, 60, 60

Multiple Modes:

It is possible for a distribution to have more than one mode (bi-modal or multi-modal).

Presence of multiple modes can limit the ability of the mode to describe the centre or typical value of the distribution.

No Mode:

In some cases, particularly with continuous data, the distribution may have no mode if all values are different.

Alternative Measures:

In such cases, it may be better to use the median or mean.

Another option is to group the data into appropriate intervals and find the modal class.

Median

Definition:

The median is the middle value in a distribution when values are arranged in ascending or descending order.

Division of Distribution:

The median divides the distribution in half, with 50% of observations on either side of the median value.

Odd Number of Observations:

In a distribution with an odd number of observations, the median is the middle value.
Example: In the retirement age distribution (11 observations), the median is 57 years.
Distribution: 54, 54, 54, 55, 56, 57, 57, 58, 58, 60, 60

Even Number of Observations:

In a distribution with an even number of observations, the median is the mean of the two middle values.

Example: In the following distribution, the median is the mean of 56 and 57, which equals 56.5 years.

Distribution: 52, 54, 54, 54, 55, 56, 57, 57, 58, 58, 60, 60

Advantage of the median

The median is less affected by outliers and skewed data than the mean and is usually the preferred measure of central tendency when the distribution is not symmetrical.

Limitation of the median

The median cannot be identified for categorical nominal data, as it cannot be logically ordered.

Mean

The mean is the sum of the value of each observation in a dataset divided by the number of observations. This is also known as the arithmetic average.

Looking at the retirement age distribution again:

54, 54, 54, 55, 56, 57, 57, 58, 58, 60, 60

The mean is calculated by adding together all the values ($54+54+54+55+56+57+57+58+58+60+60 = 623$) and dividing by the number of observations (11) which equals 56.6 years.

Advantage of the mean

The mean can be used for both continuous and discrete numeric data.

Limitations of the mean

The mean cannot be calculated for categorical data, as the values cannot be summed.

As the mean includes every value in the distribution the mean is influenced by outliers and skewed distributions.

Another thing about the mean

The population mean is indicated by the Greek symbol μ (pronounced 'mu'). When the mean is calculated on a distribution from a sample it is indicated by the symbol \bar{x} (pronounced X-bar).

Impact of shape of distribution on measures of central tendency

Symmetrical distributions

When a distribution is symmetrical, the mode, median and mean are all in the middle of the distribution.

The following graph shows a larger retirement age dataset with a distribution which is symmetrical. The mode, median and mean all equal 58 years.

Retirement age: Symmetrical distribution



Skewed Distributions

General Characteristics:

In a skewed distribution, the mode is the most commonly occurring value.

The median remains the middle value.

The mean is generally 'pulled' in the direction of the tails.

Preferred Measure of Central Tendency:

In skewed distributions, the median is often preferred over the mean because the mean is not usually in the middle of the distribution.

- **Positive (Right) Skew:**

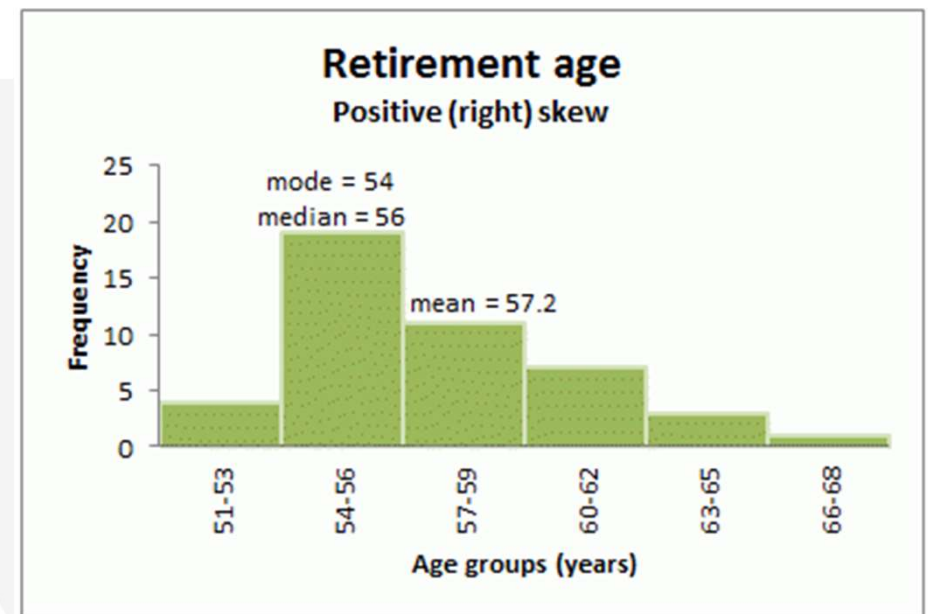
- A distribution is positively or right-skewed when the tail on the right side is longer than the left.
- In a positively skewed distribution, the mean is commonly 'pulled' toward the right tail.
- Most values, including the median, tend to be less than the mean.

- **Example of Right-Skewed Distribution:**

- A larger retirement age data set is right-skewed.
- Data grouped into classes because the variable (retirement age) is continuous.
- Mode: 54 years.
- Modal class: 54-56 years.
- Median: 56 years.
- Mean: 57.2 years.



Retirement age: Positive (right) skew



Negatively (Left) Skewed Distributions

1. Definition:

- A distribution is negatively or left-skewed when the tail on the left side is longer than the right side.

2. Behavior of the Mean:

- In a negatively skewed distribution, the mean is commonly 'pulled' toward the left tail.

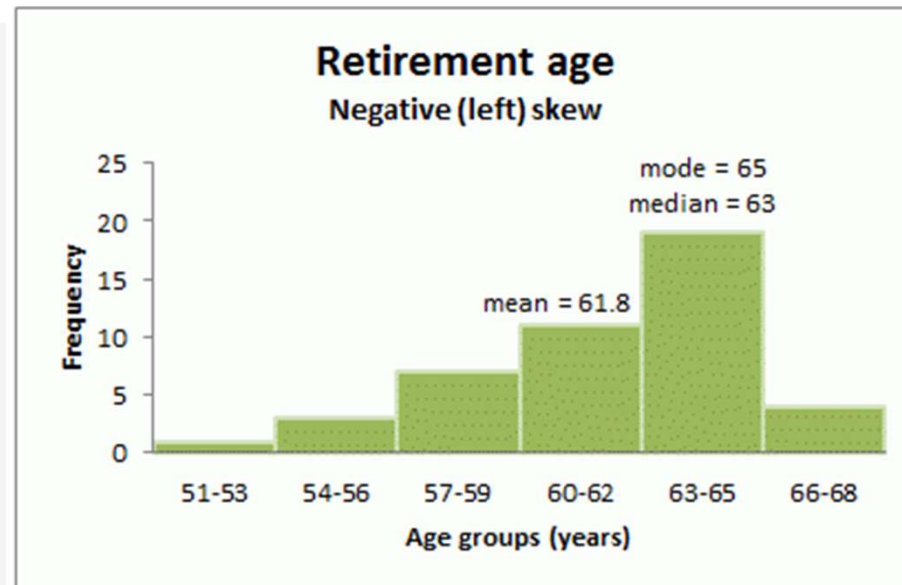
3. Values Relative to the Mean:

- Generally, in a left-skewed distribution, most values, including the median, tend to be greater than the mean. There are exceptions to this rule.

4. Example of Left-Skewed Distribution:

- A larger retirement age dataset is left-skewed.
- Mode: 65 years.
- Modal class: 63-65 years.
- Median: 63 years.
- Mean: 61.8 years.

Retirement age: Negative (left) skew



Outliers' Influence on Measures of Central Tendency

Definition of Outliers:

Outliers are extreme or atypical data values that are notably different from the rest of the data.

Importance of Detecting Outliers:

Detecting outliers is crucial because they can alter the results of data analysis.

Sensitivity to Outliers:

The mean is more sensitive to the existence of outliers compared to the median or mode.

Example with Retirement Age Dataset:

Original dataset: 54, 54, 54, 55, 56, 57, 57, 58, 58, 60, 60.

Modified dataset with outlier: 54, 54, 54, 55, 56, 57, 57, 58, 58, 60, 81.

The median remains unchanged at 57 years.

The mean is influenced by the outlier.

Calculation: $(54+54+54+55+56+57+57+58+58+60+81 = 644)$, divided by 11 = 58.5 years.

The outlier increases the mean value.

Mean as a Measure of Central Tendency:

Despite outliers, the mean can still be appropriate, especially if the rest of the data is normally distributed.

Valid extreme values should not be removed from the dataset.

Regression techniques can help reduce the influence of outliers on the mean value.

Variance

Variance is a measure of the dispersion or spread of a set of values. It quantifies how much the values in a dataset deviate from the mean (average) of the dataset. A higher variance indicates that the data points are more spread out from the mean, while a lower variance indicates that they are closer to the mean.

Formula

For a population of N values X_1, X_2, \dots, X_N with mean μ , the population variance σ^2 is given by:

$$\sigma^2 = \frac{1}{N} \sum_{i=1}^N (X_i - \mu)^2$$

For a sample of n values x_1, x_2, \dots, x_n with sample mean \bar{x} , the sample variance s^2 is given by:

$$s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$$

The sample variance uses $n - 1$ in the denominator instead of n to provide an unbiased estimate of the population variance.

Example Calculation

Let's calculate the population variance and sample variance for a small dataset:

Data: [2, 4, 4, 4, 5, 5, 7, 9]

1. Calculate the Mean:

$$\mu = \bar{x} = \frac{2+4+4+4+5+5+7+9}{8} = 5$$

2. Calculate the Population Variance:

$$\sigma^2 = \frac{1}{8}[(2 - 5)^2 + (4 - 5)^2 + (4 - 5)^2 + (4 - 5)^2 + (5 - 5)^2 + (5 - 5)^2 + (7 - 5)^2 +$$

$$\sigma^2 = \frac{1}{8}[9 + 1 + 1 + 1 + 0 + 0 + 4 + 16]$$

$$\sigma^2 = \frac{1}{8}[32] = 4$$

3. Calculate the Sample Variance:

$$s^2 = \frac{1}{7}[(2 - 5)^2 + (4 - 5)^2 + (4 - 5)^2 + (4 - 5)^2 + (5 - 5)^2 + (5 - 5)^2 + (7 - 5)^2 +$$

$$s^2 = \frac{1}{7}[32] \approx 4.571$$

Key Points

- **Population:** The entire set of items or individuals of interest.
- **Variance:** A measure of how data points differ from the mean, indicating the spread or dispersion in the data.

Population variance is used when you have data for the entire population. Sample variance is used when you have data for only a sample of the population and want to estimate the population variance.



confidence interval?

A confidence interval is a range of values that is used to estimate the true value of a population parameter. It provides an interval within which we are confident that the parameter lies, based on the data from a sample.

Key Points about Confidence Intervals

1. **Point Estimate:** A single value estimate of a parameter (e.g., sample mean).
2. **Margin of Error:** The amount added and subtracted from the point estimate to create the interval.
3. **Confidence Level:** The probability that the confidence interval contains the true parameter value. Common levels are 90%, 95%, and 99%.
4. **Interpretation:** If you have a 95% confidence interval, it means that if you were to take many samples and build a confidence interval from each one, approximately 95% of those intervals would contain the true parameter.

Example

If you have a sample mean of 50, a standard deviation of 10, and a 95% confidence level, you might calculate a confidence interval as follows:

1. **Determine the critical value:** For a 95% confidence level and a normal distribution, the critical value (z-score) is approximately 1.96.
2. **Calculate the margin of error:** Margin of Error = $1.96 \times \left(\frac{10}{\sqrt{n}} \right)$ where n is the sample size.
3. **Construct the interval:** If $n = 100$, the margin of error is $1.96 \times 1 = 1.96$. The confidence interval is 50 ± 1.96 , or (48.04, 51.96).



Importance

Confidence intervals are essential in statistics because they provide a range for estimating unknown parameters and convey the uncertainty inherent in the sampling process. They are widely used in various fields, including science, engineering, economics, and social sciences, to make informed decisions based on sample data.

