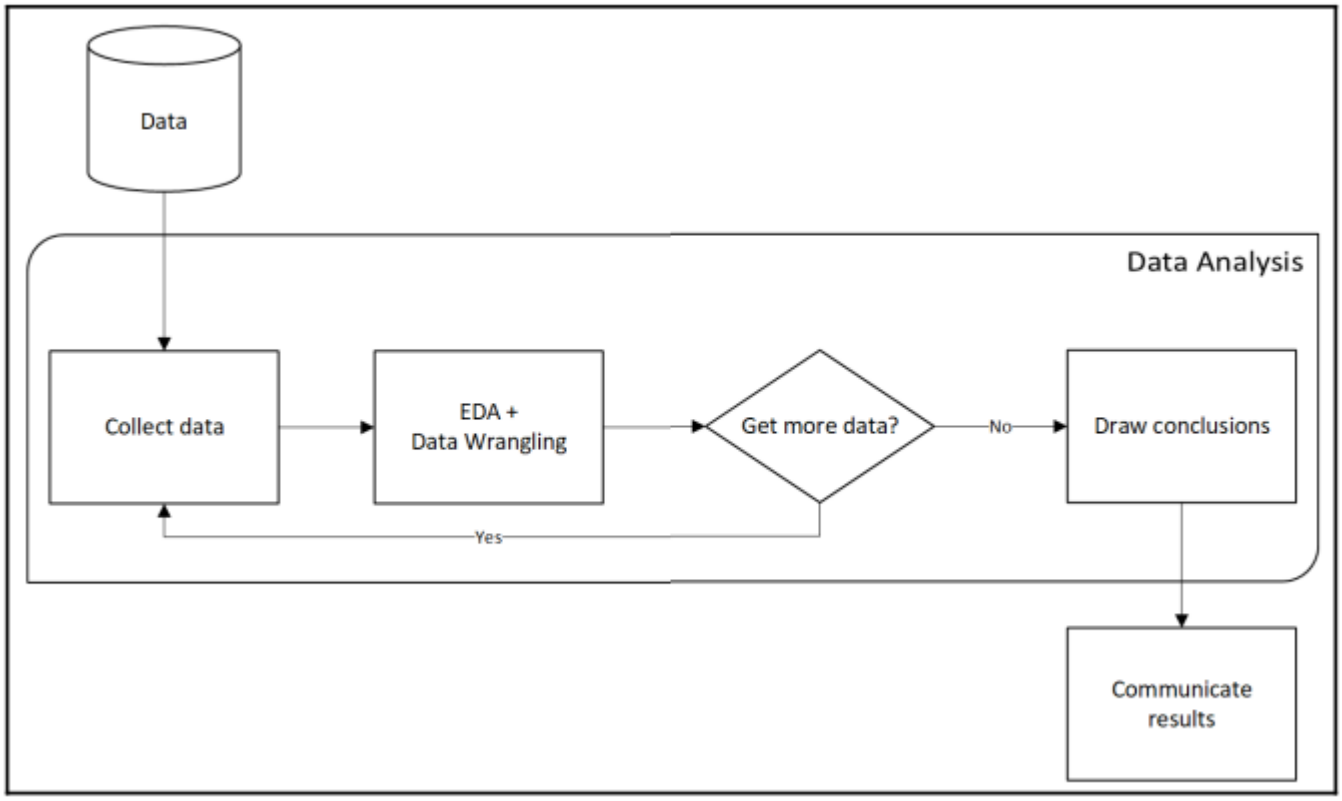
Fundamentals of data analysis

Data analysis is a highly iterative process involving collection, preparation (wrangling), exploratory data analysis (EDA), and drawing conclusions.



Fundamentals of data analysis involve a set of principles, techniques, and processes used to extract meaningful insights from data. Whether you are dealing with small datasets or big data, the following key concepts and steps are essential for effective data analysis:

1. **Data Collection:**
   Data collection is the natural first step for any data analysis.Start by gathering relevant data from various sources, such as databases, surveys, sensors, or external datasets.
   - **Web scraping** to extract data from a website's HTML (often with Python packages such as **selenium, requests, scrapy, and beautifulsou**p)
   - **Application Programming Interfaces (APIs)** for web services from which we can collect data with the requests package
   - **Databases** (data can be extracted with SQL or another database-querying language)
   - **Internet resources** that provide data for download, such as government websites or Yahoo! Finance
   - Log files

2. **Data wrangling (Data Cleaning and Preprocessing):**
   Data wrangling is the process of preparing the data and getting it into a format that can be used for analysis .
   - Ensure data quality by addressing missing values, outliers, and inconsistencies.

- Transform and preprocess data as needed, including normalization, scaling, and encoding categorical variables.

The following are some issues we may encounter with our data:

- **Human errors:** Data is recorded (or even collected) incorrectly, such as putting 100 instead of 1000, or typos. In addition, there may be multiple versions of the same entry recorded, such as New York City, NYC, and nyc
- **Computer error:** Perhaps we weren't recording entries for a while (missing data)
- **Unexpected values:** Maybe whoever was recording the data decided to use ? for a missing value in a numeric column, so now all the entries in the column will be treated as text instead of numeric values
- **Incomplete information:** Think of a survey with optional questions; not everyone will answer them, so we have missing data, but not due to computer or human error
- **Resolution:** The data may have been collected per second, while we need hourly data for our analysis
- **Relevance of the fields:** Often, data is collected or generated as a product of some process rather than explicitly for our analysis. In order to get it to a usable state, we will have to clean it up
- **Format of the data:** The data may be recorded in a format that isn't conducive to analysis, which will require that we reshape it.
- **Misconfigurations in data-recording process:** Data coming from sources such as misconfigured trackers and/or webhooks may be missing fields or passing them in the wrong order.

3. **Exploratory Data Analysis (EDA):**During EDA, we use visualisations and summary statistics to get a better understanding of the data.
   Use **descriptive statistics**, **data visualisation techniques** (histograms, scatter plots, box plots, etc.), and **summary statistics** to understand the dataset's characteristics and relationships between variables.

   In the workflow diagram we saw earlier, EDA and data wrangling shared a box. This is because they are closely tied:

   - Data needs to be prepped before EDA.
   - Visualisations that are created during EDA may indicate the need for additional data cleaning.
   - Data wrangling uses summary statistics to look for potential data issues, while EDA uses them to understand the data. Improper cleaning will distort the findings when we're conducting EDA. In addition, data wrangling skills will be required to get summary statistics across subsets of the data.

   When calculating summary statistics, we must keep the **types of data** we collected in mind.

   ## Types of data:

   In data analysis, data can be categorised into different types based on their nature and characteristics.The main types of data in data analysis are:

   1. **Nominal Data:**

- Nominal data represents categories or labels with no inherent order or ranking. It's used to classify items into distinct groups.
- Examples: Colors (red, blue, green), gender (male, female, non-binary), types of animals (dog, cat, bird).

2. **Ordinal Data:**
   - Ordinal data also represents categories, but these categories have a specific order or ranking. The intervals between values are not necessarily equal.
   - Examples: Education levels (high school, bachelor's, master's, Ph.D.), customer satisfaction ratings (poor, fair, good, excellent), socioeconomic status (low, middle, high).
3. **Interval Data:**
   - Interval data represents values with a consistent interval or difference between them. It has no true zero point, meaning that zero does not indicate the absence of the attribute being measured.
   - Examples: Temperature in Celsius or Fahrenheit, IQ scores, calendar dates (months, years).
4. **Ratio Data:**
   - Ratio data, like interval data, has a consistent interval between values, but it also has a meaningful zero point. A zero in ratio data indicates the absence of the attribute being measured.
   - Examples: Height, weight, income, age, time (measured in seconds), number of items purchased.
5. **Continuous Data:**
   - Continuous data can take any value within a given range and can have an infinite number of possible values. It is often represented with real numbers.
   - Examples: Temperature, height, weight, income, time.
6. **Discrete Data:**
   - Discrete data can only take specific, distinct values and often consists of whole numbers.
   - Examples: Number of people in a household, number of cars in a parking lot, the count of customer complaints.
7. **Categorical Data:**
   - Categorical data includes nominal and ordinal data types. It represents categories or groups.
   - Examples: Eye color (nominal), education level (ordinal), car make and model (nominal).
8. **Numeric Data:**
   - Numeric data includes interval and ratio data types. It represents data that can be measured on a numerical scale.
   - Examples: Temperature (interval), height (ratio), income (ratio).
9. **Text Data:**
   - Text data consists of unstructured textual information. Analysing text data often involves natural language processing (NLP) techniques.
   - Examples: Text documents, customer reviews, social media posts, emails.
10. **Time Series Data:**
    - Time series data consists of observations recorded at successive time intervals. It's commonly used for analysing trends and patterns over time.
    - Examples: Stock prices, temperature readings, website traffic data.

11. **Geospatial Data:**
    - Geospatial data includes information related to geographical locations and coordinates.
    - Examples: GPS coordinates, maps, geographic information system (GIS) data.

4. Drawing conclusions

After we have collected the data for our analysis, cleaned it up, and performed some thorough EDA, it is time to draw conclusions. This is where we summarise our findings from EDA and decide the next steps:

- Did we notice any patterns or relationships when visualising the data?
- Does it look like we can make accurate predictions from our data? Does it
- make sense to move to modelling the data?
- Do we need to collect new data points?
- How is the data distributed?
- Does the data help us answer the questions we have or give insight into the problem we are investigating?
- Do we need to collect new or additional data?

Statistical foundations for Data Analysis:

Statistical foundations are essential for data analysis, as statistics provide the tools and techniques necessary to make sense of data, draw meaningful conclusions, and make data-driven decisions. Two broad categories of statistics are **descriptive and inferential statistics.**

Descriptive Statistics:

- Descriptive statistics help summarize and describe the main features of a dataset. Common measures include mean, median, mode, variance, standard deviation, range, and percentiles.
- Descriptive statistics provide an initial understanding of data distribution and central tendencies.

Inferential Statistics:

- Inferential statistics are used to make inferences or predictions about a population based on a sample of data. Common techniques include hypothesis testing and confidence intervals.
- Hypothesis testing helps assess whether observed differences or relationships in the data are statistically significant.

Descriptive statistics :

our discussion of descriptive statistics with univariate statistics; univariate simply means that these statistics are calculated from one **(uni) variable**.

Descriptive statistics are used to describe and/or summarise the data we are working with. We can start our summarization of the data with a measure of central tendency, which describes where most of the data is centred around, and a measure of **spread**

or **dispersion**, which indicates how far apart values are.

An **outlier** is an observation or data point that significantly deviates from the rest of the data in a dataset. In other words, it's a data point that is markedly different from the majority of the other data points. Outliers can occur in various types of data, including numerical, categorical, and even time-series data. Outliers are sometimes also referred to as **anomalies**.

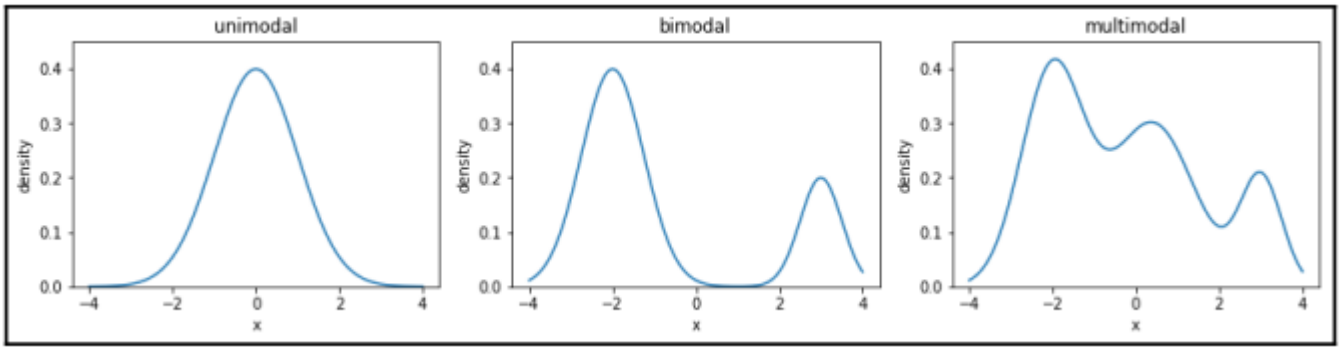
Measures of central tendency
Measures of central tendency describe the centre of our distribution of data. There are three common statistics that are used as measures of centre: **mean, median, and mode**. Each has its own strengths, depending on the data we are working with.

Mean: the most common statistic for summarising data is the average, or mean.The sample mean is calculated by summing all the values and dividing by the count of values;
"One important thing to note about the mean is that it is very sensitive to outliers".

Median : In cases where we suspect outliers to be present in our data, we may want to use the median as our measure of central tendency. Unlike the mean, **the median is robust to outliers**.

The median represents the 50th percentile of our data; this means that 50% of the values are greater than the median and 50% are less than the median.

Mode: The **mode** is the most common value in the data.



In practice, this isn't as useful as it would seem, but we will often hear things like the distribution is bimodal or multimodal (as opposed to unimodal) in cases where the distribution has two or more most popular values. This doesn't necessarily mean that each of them occurred the same amount of times, but, rather, they are more common than the other values by a significant amount. As shown in the following plots, a unimodal distribution has only one mode (at 0), a bimodal distribution has two (at -2 and 3), and a multimodal distribution has many (at -2, 0.4, and 3):

Sampling

There's an important thing to remember before we attempt any analysis: our sample must be a random sample that is representative of the population. This means that the data must be sampled without bias .

Measures of spread

In data analysis and visualisation, measures of spread, also known as measures of dispersion, provide valuable insights into the variability or spread of data points within a dataset. These measures help to understand how data points are distributed around the central tendency (mean, median) and provide important information about the data's variability and potential outliers. Here are some common measures of spread in data analysis and visualisation:

Range:

Definition: The range is the simplest measure of spread and represents the difference between the maximum and minimum values in the dataset.
Formula: Range = Max Value - Min Value
Use: It gives a rough idea of the spread but is sensitive to outliers.

Variance:

Definition: The variance is calculated as the average squared distance from the mean. Variance measures how each data point differs from the mean squared. It provides an average of the squared differences from the mean.
Formula: Variance = $\Sigma(x_i - \overline{x})^2 / N$ where $\overline{x}$ is the mean and N is the number of data points.
Use: Variance quantifies the overall spread of the data but is sensitive to extreme values.

Standard Deviation:

Definition: The variance gives us a statistic with squared units. The standard deviation is the square root of the variance. It provides a measure of the average deviation from the mean.
Formula: Standard Deviation = √Variance
Use: It is a commonly used measure of spread and provides a more interpretable value compared to variance.

Coefficient of Variation (CV): **Definition**: Ratio of the standard deviation to the mean. The CV is a relative measure of spread, expressed as a percentage. It helps compare the spread of datasets with different units or scales.

Formula: CV = (Standard Deviation / Mean)
Use: It helps assess the relative variability of data and is useful for comparing datasets with different magnitudes.

Percentiles:

Definition: Percentiles indicate the value below which a given percentage of data falls. For example, the 25th percentile is the value below which 25% of the data falls.
Use: Percentiles help understand data distribution and identify extreme values.

Quartiles:

Definition: Quartiles divide a dataset into four equal parts, with Q1, Q2 (median), and Q3 marking the 25th, 50th, and 75th percentiles, respectively.
Use: Quartiles are essential for calculating the IQR(Interquartile Range) and identifying potential outliers.

Percentiles and quartiles are both quantiles—values that divide data into equal groups each containing the same percentage of the total data; percentiles give this in 100 parts, while quartiles give it in four (25%, 50%, 75%, and 100%).

Interquartile Range (IQR):

Definition: The IQR is a robust measure of spread that describes the spread of the middle 50% of the data.
Formula: IQR = Q3 (Third Quartile) - Q1 (First Quartile)
Use: It is less affected by outliers compared to the range and provides a better understanding of the central spread.

Box-and-Whisker Plot (Box Plot):

Definition: A box plot visually represents the median, quartiles (Q1 and Q3), and potential outliers in the data distribution.
Use: It provides a graphical summary of the data's central tendency and spread.

Histogram:

Definition: A histogram is a graphical representation of data distribution. It shows the frequency of data within predefined bins or intervals.
Use: Histograms provide insights into the shape and spread of data.
Measures of spread are crucial for understanding data variability, making informed decisions, and identifying anomalies or outliers in datasets. When combined with measures of central tendency (e.g., mean, median), they provide a comprehensive view of data distribution.

Correlation and Covariance

Correlation and Covariance are both measures used in statistics to describe the relationship between two or more variables. They provide insights into how changes in one variable are related to changes in another.They are often used in various fields, including finance, economics, and data analysis, to understand how two variables are related to each other.

Let's delve into each of these concepts in detail, along with examples.

Covariance:

Covariance measures the degree to which two variables change together.Specifically, it quantifies the extent to which deviations from the mean (average) of one variable correspond with deviations from the mean of another variable. It indicates whether an increase in one variable corresponds to an increase or decrease in another.

The formula for covariance between two variables X and Y is:

$$Cov(X, Y) = \frac{1}{n-1} \sum_{i=1}^{n} (Xi - \overline{X})(Yi - \overline{Y})$$

Where:

- Cov(X,Y) is the covariance between X and Y.
- n is the number of data points.
- Xi and Yi are individual data points of X and Y, respectively.
- $\overline{X}$ and $\overline{Y}$ are the means (averages) of X and Y, respectively.

Interpretation of covariance:

- If Cov(X,Y) > 0, it implies a positive relationship. An increase in X is associated with an increase in Y, and vice versa.

- If Cov(X,Y)<0, it indicates a negative relationship. An increase in X corresponds to a decrease in Y, and vice versa.

- If Cov(X,Y)=0, it suggests no linear relationship between X and Y. However, it's important to note that zero covariance does not necessarily mean there's no relationship; it means there's no linear relationship.

Example of Covariance:

Let's say you have data on the number of hours students spend studying for an exam (X) and their exam scores (Y). Here are some hypothetical data points:

X (hours)	Y (scores)
10	85
5	70
8	78
12	92
7	72

To calculate the covariance between hours of study (X) and exam scores (Y):

Calculate the mean of X($\overline{X}$) and Y($\overline{Y}$):-

$$\overline{X} = \frac{10+5+8+12+7}{5} = 8.4\ hours$$

$$\overline{Y} = \frac{85+70+78+92+72}{5} = 79\ seconds$$

Use the covariance formula to calculate Cov(X, Y):

$$Cov(X,Y) = \frac{1}{4}[(10-8.4)(85-79) + (5-8.4)(70-79) + (8-8.4)(78-79) + (12-8.4)(92-79) + (7-8.4)(72-79)]$$

Cov(X,Y) ≈24.3

The **positive covariance (24.3)** indicates that there is a positive relationship between hours of study and exam scores. As students spend more hours studying, their exam scores tend to be higher on average.

Correlation:

Correlation, on the other hand, is a standardised measure of the strength and direction of the linear relationship between two variables. It is expressed as a value between -1 and 1.

The formula for Pearson's correlation coefficient (r), which measures linear correlation, is:

$$r = \frac{Cov(X,Y)}{\sigma_x\ \sigma_y}$$

Where:

- r is the correlation coefficient.
- Cov(X,Y) is the covariance between X and Y.
- σ_x and σ_y are the standard deviations of X and Y, respectively.

Interpretation of correlation:

- r = 1 indicates a perfect positive linear relationship.
- r = −1 indicates a perfect negative linear relationship.
- r = 0 indicates no linear relationship.

Example of Correlation:

Continuing with the study hours (X) and exam scores (Y) example, if we calculate the correlation coefficient:

$\sigma_x \approx 2.415$ hours

$\sigma_y \approx 8.184$ scores

Using the previously calculated covariance (17.6), we can calculate the correlation coefficient (r):

$$r = \frac{Cov(X, Y)}{\sigma_x \sigma_y} \qquad r = \frac{24.3}{2.415*8.184} \approx \mathbf{1.225}$$

The correlation coefficient (r ≈ 1.225) is close to 1, indicating a strong positive linear relationship between hours of study and exam scores. This suggests that as students spend more hours studying, their exam scores tend to increase linearly.

In summary, covariance measures the direction of the relationship between two variables, while correlation quantifies both the direction and strength of the linear relationship, providing a standardised measure that can be compared across different datasets.

1. Find correlation and covariance of following data
   Hours Studied (X): [3, 5, 2, 8, 6]
   Exam Scores (Y): [60, 75, 50, 85, 70]

2. You have two datasets, P and Q, with the following summary statistics:
   P: Mean = 12, Standard Deviation = 4
   Q: Mean = 18, Standard Deviation = 6
   Covariance between P and Q = 24
   Calculate the correlation coefficient between P and Q.

3. You are given two datasets, X and Y, with the following covariance matrices:

Cov(X, X) = 16
Cov(Y, Y) = 25
Cov(X, Y) = -10
Calculate the correlation coefficient between X and Y.

4. Consider another dataset with values for variables A and B:
   A: [10, 12, 14, 16, 18]
   B: [20, 22, 24, 26, 28]
   Calculate the covariance between A and B.
   Calculate the correlation coefficient between A and B.

5. Suppose you have a dataset with the following values for variables X and Y:
   X: [2, 4, 6, 8, 10]
   Y: [5, 7, 8, 12, 15]
   Calculate the covariance between X and Y.
   Calculate the correlation coefficient between X and Y.

Statistical Hypothesis Generation and Testing

Statistical hypothesis generation and testing are fundamental concepts in the field of statistics and research. These processes help researchers make inferences about populations based on sample data and assess the significance of their findings. Here's a detailed explanation of both concepts:

Hypothesis Generation:

- **Definition:** Hypothesis generation is the first step in the scientific method, where researchers formulate a clear and testable statement or assumption about a population or phenomenon of interest.
- **Purpose:** The primary goal of hypothesis generation is to propose a specific idea or theory about the relationship between variables, which can then be tested through empirical research.
- **Components:** A hypothesis typically consists of two parts:
  Null Hypothesis (H0): This is a statement that there is no effect, no difference, or no relationship between variables. It represents the status quo or the default assumption.
  Alternative Hypothesis (Ha): This is a statement that contradicts the null hypothesis and suggests that there is an effect, a difference, or a relationship between variables.
- **Example:** Suppose you want to investigate whether a new drug is more effective than an existing one. Your hypotheses might be:
  Null Hypothesis (H0): The new drug is equally effective as the existing drug.

Alternative Hypothesis (Ha): The new drug is more effective than the existing drug.

Hypothesis Testing:

Definition: Hypothesis testing is a statistical method used to determine whether the evidence from a sample supports the null hypothesis or favours the alternative hypothesis.

Process:

- **Data Collection:** Collect a sample of data relevant to the research question.
- **Assumption Check:** Ensure that the data meet the assumptions of the chosen statistical test (e.g., normal distribution, independence).
- **Choosing a Significance Level (α):** Decide on the level of significance (α), which represents the probability of making a Type I error (rejecting the null hypothesis when it's true).
- **Select a Statistical Test:** Choose an appropriate statistical test based on the type of data and research question (e.g., t-test, chi-squared test, ANOVA).
- **Calculate the Test Statistic:** Compute a test statistic that summarises the relationship between the data and the null hypothesis.
- **Determine the Critical Region:** Identify a critical region or critical value(s) that defines the boundary for rejecting the null hypothesis based on the chosen α.
- **Compare the Test Statistic and Critical Value(s):** If the test statistic falls within the critical region, reject the null hypothesis; otherwise, fail to reject it.
- **Draw a Conclusion:** Based on the comparison, make a decision about the null hypothesis, stating whether there is enough evidence to support the alternative hypothesis.
- **Report Results:** Communicate the results, including the test statistic, p-value (probability of observing the data under the null hypothesis), and the decision made.
  Example: Using the drug effectiveness scenario, if the p-value is less than α (e.g., $p < 0.05$), you would reject the null hypothesis and conclude that there is enough evidence to suggest that the new drug is more effective than the existing one.

Hypothesis testing is a critical tool in scientific research, allowing researchers to make informed decisions based on empirical evidence while considering the possibility of **Type I** and **Type II** errors. It provides a structured and objective way to evaluate hypotheses and draw conclusions about the real-world phenomena being studied.