Inferential Statistics

Imagine you own a juice company and you've created a new juice flavour. You want to know if customers prefer this new flavour over the current best-seller.

However, instead of asking **every customer** (which would take too long and be too expensive), you decide to test it with a smaller group of people (a **sample**).

Step 1: Hypothesis Testing

- **Population**: All your customers who might buy the juice.
- **Sample**: A small group of 100 customers who will try both flavours.

You ask the group of 100 customers to rate both juices. Then, you compare the ratings to see which juice is more popular.

- **Null Hypothesis (H_0)**: There is no difference in preference between the new juice flavour and the current best-seller.
- **Alternative Hypothesis (H_1)**: The new juice flavour is preferred by customers more than the current best-seller.

Step 2: Conduct the Test

You analyse the data and calculate the **p-value**, which tells you the probability of getting the results you did, assuming that the null hypothesis is true. If the p-value is very small (for example, less than 0.05), you reject the null hypothesis and say that the new juice is likely preferred by customers.

Step 3: Confidence Interval

You also calculate a **confidence interval** to give you a range of possible preference differences (e.g., the new juice is between 5% and 15% more popular).

Conclusion:

Based on your **inferential statistics** (hypothesis test and confidence interval), you conclude whether the new juice flavour is likely to be more popular across all customers, not just the sample.

Probability

Imagine you are planning a picnic and checking the weather forecast. The forecast says there is a 70% chance of rain tomorrow. Now, you want to know the probability that it will rain tomorrow **and** that it will also be windy.

Step 1: Basic Probability

- **Event A**: It rains tomorrow.
- **Event B**: It is windy tomorrow.

The **Basic Probability** of rain (Event A) is given as 70% or 0.7, meaning there is a 70% chance of rain. If the wind condition is also known to happen independently 40% of the time (Event B), you can calculate the probability of both events happening using basic probability.

- **Formula**: $P(A \cap B) = P(A) * P(B) = 0.7 * 0.4 = 0.28$

This means there is a 28% chance that both it will rain **and** it will be windy tomorrow.

Step 2: Conditional Probability

Now, suppose you learn that tomorrow is likely to be cloudy. This information changes the probability of it raining, since you know rain is more likely when it's cloudy.

- **Conditional Probability**: The probability that it rains tomorrow **given** it is cloudy.

This is a **conditional probability** ($P(A | B)$), and it would tell you how the chance of rain changes given the new condition (cloudy weather).

- **Formula**: $P(A | B) = P(A \cap B) / P(B)$
- If you know that the chance of rain increases to 80% when it's cloudy, you can update your understanding of the likelihood of rain.

Step 3: Bayes' Theorem

If you get new information, such as a weather expert confirming that there's an 85% chance of rain if it’s cloudy and windy, you can use **Bayes' Theorem** to update your probability.

- **Bayes' Theorem**: $P(A | B) = (P(B | A) * P(A)) / P(B)$

Frequency Distribution
Imagine you are a teacher who has graded an exam for 50 students. You want to understand how the students' marks are distributed across different score ranges to identify patterns.

Step 1: Collecting Data

The scores of the 50 students might look like this:

50, 55, 60, 70, 85, 90, 45, 80, 70, 75, 85, 60, 50, 95, 60, 80, 90, 75, 85, 95, ...

Step 2: Creating a Frequency Table

A frequency distribution organises this data by grouping the scores into intervals or "bins," and counting how many students fall into each bin. Let's say you decide to group the scores into ranges like 40-50, 51-60, 61-70, etc.

Score Range	Frequency (Number of Students)
40-50	4
51-60	8
61-70	6
71-80	7
81-90	10
91-100	5

This table shows how many students scored in each range.

Step 3: Creating a Histogram (Visualising the Distribution)

You can use a histogram to visualise this frequency distribution. The x-axis will represent the score ranges, and the y-axis will represent the frequency (how many students scored in each range).

The heights of the bars in the histogram will show how many students fall into each score range.

Step 4: Analysing the Data

From the frequency table, you can analyse the distribution of students' scores:

- The highest frequency is in the 81-90 range, meaning most students scored between 81 and 90.
- You can see that fewer students scored in the lowest and highest ranges (40-50 and 91-100).

Descriptive Statistics

Imagine a school wants to understand the performance of students in a maths exam. They collect the scores of 100 students. Using descriptive statistics, they can summarise the data to make it easier to interpret. For example:

- Mean (Average): The average score of all students helps the school understand how well students performed overall. If the average score is 75 out of 100, the school knows the general level of performance.
- Median: If the exam scores are arranged in order from lowest to highest, the median represents the middle score. If there are 100 students, the median score would be the score of the 50th student. This helps understand the central tendency of scores, especially if there are extreme high or low scores that might skew the mean.
- Mode: The mode is the most common score. For example, if 15 students scored 80 and no other score is repeated more often, then 80 is the mode.
- Range: The difference between the highest and lowest scores. If the highest score is 95 and the lowest is 50, the range is 45. This tells us how spread out the scores are.
- Standard Deviation: This tells the school how spread out the scores are around the mean. If most students score close to the average, the standard deviation will be low, indicating consistency. If there is a lot of variation in scores, the standard deviation will be high, suggesting some students are performing much better or worse than others.

In this case, descriptive statistics help the school quickly understand how students are performing on the exam without having to look at each individual score.