HyperionDev

Statistics

Task

Visit our website

Introduction

Machine learning leverages statistical methods to uncover patterns and trends within data. By analysing these patterns, we can build models that can predict future outcomes or make classifications on new data. Statistics and probability form the foundation of machine learning, and a strong understanding of these concepts is essential for anyone interested in this field. This lesson will provide you with a solid foundation in these areas, giving you the tools to unlock the power of your data.

Statistics vs probability

Statistics and probability are two closely related fields that both use data to answer questions, but they do so in slightly different ways:

Statistics involves analysing the frequency of past events to draw inferences about a larger population. Suppose we are building a recommendation system for an online bookstore. We collect data on book ratings from 1,000 users. By calculating the average rating (mean) and assessing the spread (standard deviation) of these ratings, we can infer the overall preferences of all users and recommend books accordingly.

Probability deals with predicting the likelihood of future events based on existing known data. Consider a medical diagnosis application. Given a dataset of patient symptoms and corresponding diagnoses, we can estimate the probability of a new patient having a specific disease based on observed symptoms. This informs the diagnostic process and helps prioritise further tests.

Descriptive statistics

For this task, we will be focusing on three main classes of statistical descriptions:

1. Measures of centrality
2. Measures of spread
3. Frequency distributions

Measures of centrality

Centrality in data typically describes where most of the data is. This is, admittedly, a very vague definition. That is why there are a few different measures of centrality. We will go through the three most common measures: mean, median, and mode.

Mean

For most people, the mean or average is the easiest measure of centrality to grasp. It is simply the sum of all values divided evenly by the number of values. The main property of the mean is that if all values included were the mean value, it would add up to the same total. The formula for the mean is:

$$M = \frac{\sum_i^n X_i}{n}$$

Let's look at a practical example. You are a HyperionDev graduate living in Durban, and you have your first job interview in Pietermaritzburg. The drive there is about 80 km in total, and you have to be there in one hour. Logically speaking, you must maintain an average speed of 80km/h. This should be easy enough, right?

But there is a snag, the road is terrible today! Lots of accidents and trucks are on the road, slowing you down. You have been monitoring your speed for the trip, and you have made the following observations:

- Your journey starts smoothly, and you travel at 120 km/h for 20 minutes.
- Suddenly, there is an accident with traffic backed up. As a result, you travel at 40km/h for 5 minutes.
- You pass the accident, but now it seems there are a few trucks on the road, travelling at a steady 60 km/h. This carries on for about 10 minutes.
- The trucks start thinning out, and now you see you're travelling at 80 km/h. It looks like this will be the speed you are travelling at for the rest of the trip.

Now that you are maintaining 80 km/h, you will only make it to the interview on time if your first 45 minutes of the trip averaged 80 km/h. The question is whether you managed a suitable mean speed initially to make it to the interview.

To work this out, let's calculate your average speed for those 45 minutes:

$$M = \frac{\sum_i^n X_i}{n} = \frac{(120 \times 20) + (40 \times 5) + (60 \times 10)}{45} = 71.11$$

The calculation shows that your average speed for the first 45 minutes of the trip was only 71.11 km/h. The real lesson here is to leave early to avoid being late!

Median

The median is a slightly less commonly known measure of centrality. The median is just the middle-most value in a sequence of numbers. This provides a good measure of what is typical in a set of values. Unlike the mean, this value is unaffected by extreme outliers in the data or by differences in the distribution of the data.

The formula for the median is more of an algorithm:

- First, you arrange your data from lowest to highest. We will use X to represent the array containing this sorted data,
- you find the total number of data points, N, and
- then, you find $X_{\frac{N+1}{2}}$ if N is odd and $\frac{X_{\frac{N}{2}} + X_{\frac{N}{2}+1}}{2}$ if N is even.

For example, if you have three values in X you aim to find the second value in this set. If you have eight values in X, you will aim to find the mean of the fourth and fifth values.

Now, let's look at a practical example. You have been going to school for a year now. You have taken a few different subjects, some of which you absolutely love and some of which you really detest. You want to figure out what your "typical" mark (out of 100) is, so you can gauge what type of student you are.

- You absolutely loved Computer Science at school. As a result, you got a 95.
- Your Maths classes were interesting, but some of the equations were confusing. You got a 70 for this.
- Natural Science was so confusing! So many Latin names! You only got a 45 for this.
- Geography was cool, especially learning about population statistics! You achieved 73 for this.
- Accounting was like Maths, but not as fun. You got a middling 65 for this.
- You did not realise you were taking History until the final exams. You only got 15 for this subject!
- In Physical Sciences, Physics was really fun! The Chemistry part made no sense though... 55 for this subject.
- English was a snooze fest! Shakespeare made no sense. And you are quite sure that the poet was simply describing the colour of the sky, and not trying to reflect some underlying melodramatic theme. 52 for this subject.
- Your additional language was fairly boring, but thank the stars that there was no poetry. You scraped a decent 68 for this subject.

Keeping all of this in mind, how do we find the "typical" score? Step one says that we need to arrange everything in ascending order. Easy enough: [15, 45, 52, 55, 65, 68, 70, 73, 95].

Now, to find the middle value: easy enough, there are nine total values. The middle value in this would therefore be the fifth value: 65. Not too shabby!

Oops, there was one subject that we forgot:

- Life Orientation. It is too easy to forget that this subject exists in the first place. It was easy enough, though: you scored a solid 80 for it.

Okay, so now to recalculate. Our new list of values is: [15, 45, 52, 55, 65, 68, 70, 73, 80, 95].

Now, there are ten values: this means we need to take the mean of the fifth and sixth values. The mean of 68 and 70 is 66.5. Nice!

Mode

The mode can be easily explained: it is simply whichever number appears the most frequently in the data. Like the median, this is a good measure of the "typical" value of something. However, the mode is better suited for working with **categorical data**. (Think about why).

Let's take a look at an example. You are a new data analyst at Disney and have been assigned the ever-important role of figuring out the next Marvel movie. You were asked to send out a survey asking people which superhero they want to see next on-screen. You receive the following responses:

```
[Spider-man, Iron Man, Black Panther, Thor, Black Panther, Black Panther, Shang-Chi, Captain America, Spider-man, Dr Strange, Spider-man, Spider-man, Black Widow, Black Panther, Spider-man]
```

Looks like people really loved the last Spider-man movie. Now, let's find the mode of this data. Count the number of occurrences in each type of category. The mode is the category with the highest count. It looks like Spider-man is our mode.

Superhero	**Count**
Black Panther	4
Black Widow	1
Captain America	1
Dr Strange	1
Iron Man	1
Shang-Chi	1
Spider-man	5

Code can spare you the pain of counting:

```
import statistics

# Survey responses from participants
survey_responses = [
    "Spider-man", "Iron Man", "Black Panther", "Thor",
    "Black Panther", "Black Panther", "Shang-Chi",
    "Captain America", "Spider-man", "Dr Strange",
    "Spider-man", "Spider-man", "Black Widow",
    "Black Panther", "Spider-man"
]

# Calculate the mode (most common superhero)
mode_superhero = statistics.mode(survey_responses)

print(f"The most popular superhero is: {mode_superhero}")
```

Measures of spread

The measures of spread describe how “close together” the data is.

Variance

One common term you will encounter in data science is variance. Variance is an absolute measure of how “spread out” the data is. For example, in **PCA (Principal Component Analysis)**, you will encounter the term “explained variance”. This particular term is useful for the task of PCA, where you want to use as few features as possible to “explain” most of the variance in the data. This variance is where the useful information is.

In statistics, variance can be used to explain how useful a statistic the mean is. The formula for variance is:

$$S^2 = \frac{\sum_i^n (x_i - \bar{x})^2}{n-1}$$

We use $\bar{x}$ to represent the mean of all values in x.

Let's use an example of variance in real life: stock markets. You have made it to Wall Street. You are managing many varied stock portfolios. You are approached by a new investor who wants to find a low-risk stock in the coding bootcamp industry. You consider two stocks:

- DaphnisDev: their last few stock prices were [128, 146, 112, 153]. This gives a mean stock price of 134.75.
- TethysDev: their last few stock prices were [163, 52, 208, 128]. This gives a mean stock price of 137.75.

So I guess we go with TethysDev, right? Not so fast: we were asked for a low-risk investment. Both stocks seem to be on a generally upward trajectory, so how do we guarantee stability? Let's look at how much each of these stocks deviates from the mean: the variance of the stock.

Putting DaphnisDev through the equation for variance, we get about 340.92. Putting TethysDev through this same equation, we get 4340.25. This makes it obvious: we go with DaphnisDev, as it tends to vary less from the mean.

Standard deviation

Standard deviation is just the square root of the variance. If we already have the variance, and the standard deviation is just a one-to-one mapping of this, how is it useful to us? Well, standard deviation is commonly used when defining outliers. Typically, when data lies one standard deviation away from the rest of the data, this is considered "unusual". Depending on what you aim to do with the data, you can identify points that are x standard deviations from the mean as an outlier. This data can either be removed or changed appropriately.

The formula for standard deviation is a simple one: it is just the square root of the deviation. Or, more mathematically:

$$S = \sqrt{\frac{\sum_i^n (x_i - \bar{x})^2}{n-1}}$$

Let's look at an example of standard deviation in sports. A.B. de Villiers is a famous South African cricketer. You work in the sports industry, and you are part of an analysis team examining his performances over the past year or so. You discover that at a few points in the year, poor A.B. stubbed his toe quite badly against the trophy cabinet that holds his many awards (talk about suffering from success!)

This, unfortunately, affected his performance in some random cricket matches. These are considered outliers and shouldn't be factored into his overall performance. You get the following dataset of his total runs achieved in certain matches: [42, 29, 39, 53, 12, 52, 46, 22, 48]. This gives a mean of 38.11.

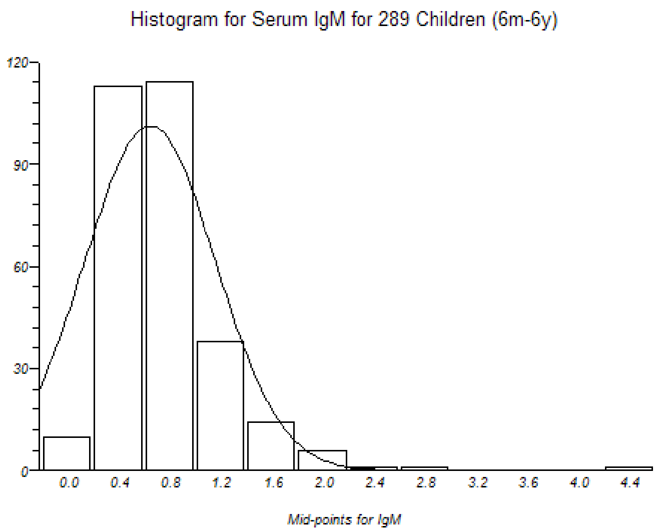
You are told that anything more than one standard deviation from the mean is considered an outlier. Plugging all of our values into the formula for variance, we get 201.86. Therefore, the standard deviation of this data is 14.2. This means that anything less than 23.91 (38.11 - 14.2) and anything more than 52.31 (38.11 + 14.2) is considered an outlier. It seems that 12 and 22 are our outliers here, so he most likely stubbed his toe before the matches in which he achieved these run totals.

Frequency distributions

Frequency distributions provide a snapshot of how often each value occurs in a dataset. They allow us to understand the distribution of data, revealing patterns and insights. They can be expressed as counts, percentages, or frequencies.

Histograms are a common way to visualise frequency distributions. They display the frequency of values within predefined intervals (bins). Bar charts, which are similar to histograms, can also be used to visualise the frequency of different labels for categorical data.

For example, the following histogram shows the frequency of different concentrations of IgM antibodies in blood in a sample of children aged six months to six years.



Histogram for Log (base 10): (StatsDirect, n.d.)

The histogram shows how frequently different concentration levels appear. In this example, the concentration bins with the highest bars 0.4 and 0.8 are the ones that represent the most frequently observed antibody levels.

Analysis

Analysis of frequency distributions gives you insight into the data so that you can choose appropriate statistical methods and make informed decisions.

Armed with insights from frequency distributions, we can:

- Choose appropriate statistical methods (e.g., t-tests, ANOVA, regression).
- Make informed decisions based on data patterns.

Here are the four key aspects of analysing frequency distributions using histograms (Tamplin, 2023):

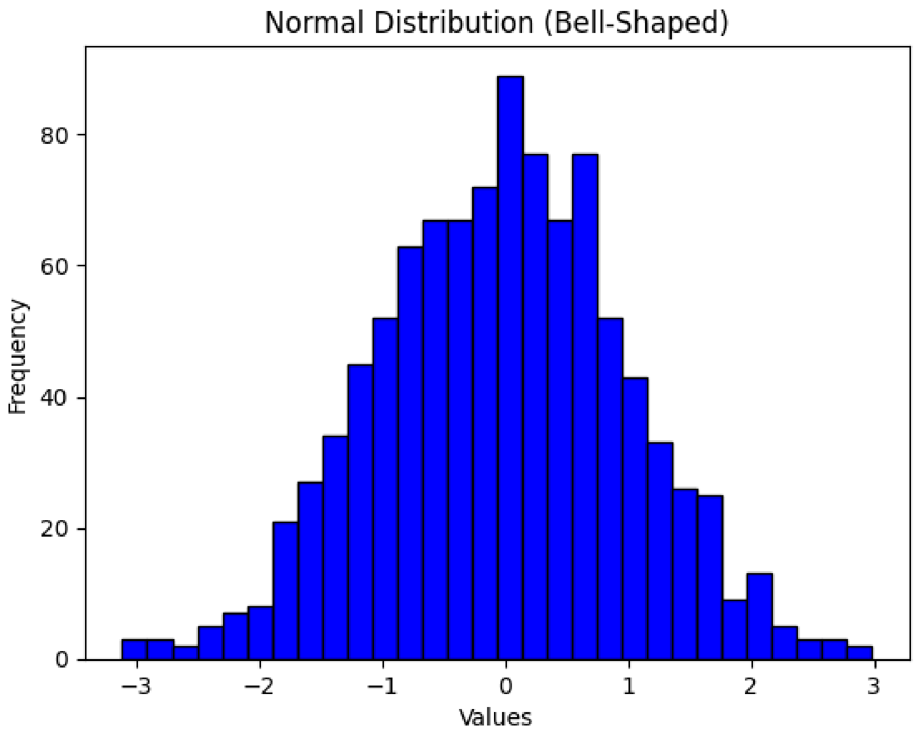
1. Shape: The shape of a histogram can reveal the distribution pattern of the data, such as whether it is symmetrical, skewed, unimodal, bimodal, or multimodal.
2. Centre: The centre of a histogram indicates the median of the dataset, which helps identify the 'typical' value within the data.
3. Spread: This aspect refers to the range of the data, showing how widely the data points are distributed.
4. Outliers: Outliers are data points that significantly deviate from the rest, which may indicate variability or errors in the data.

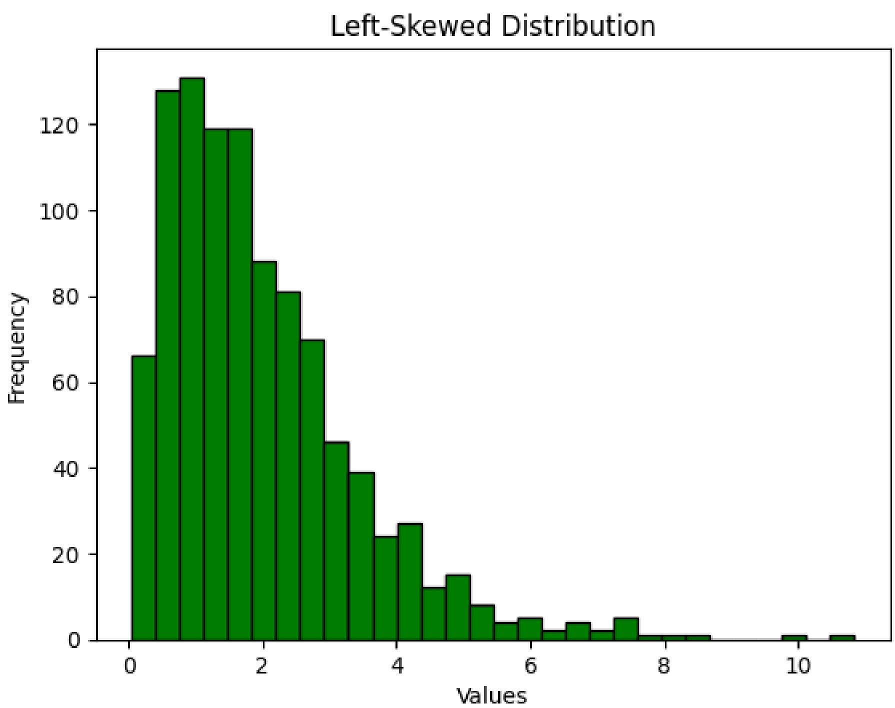
Shape

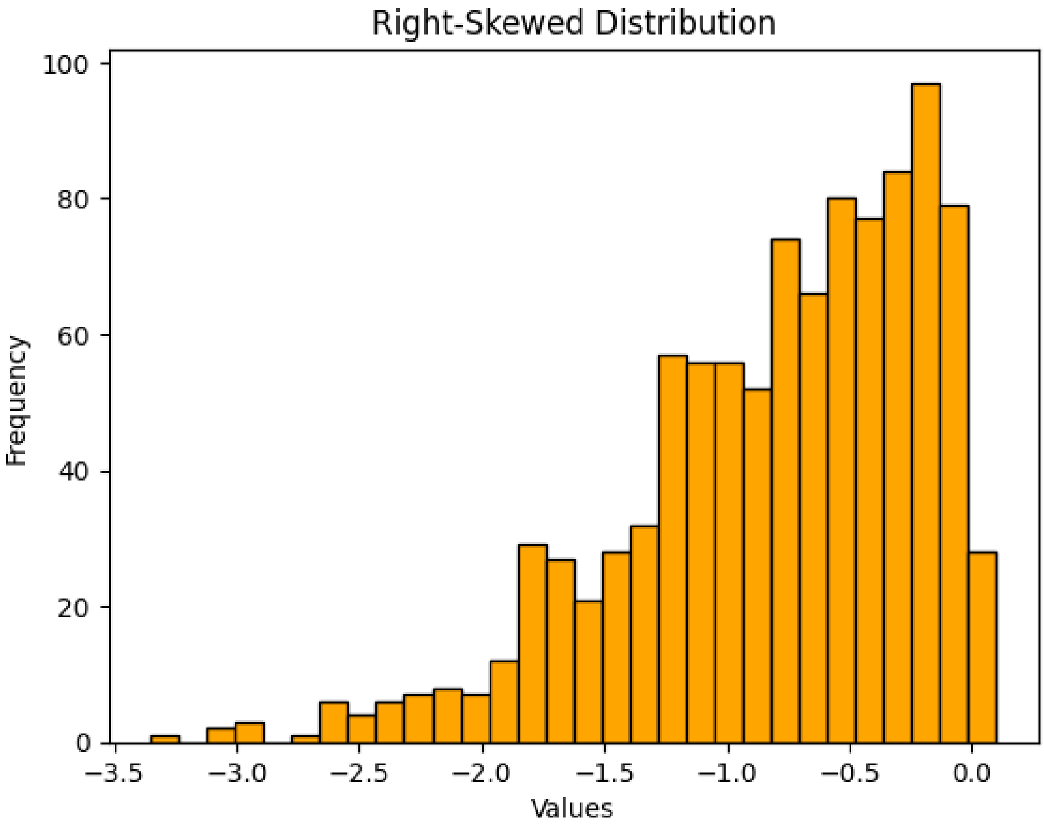
The shape of the histogram reveals the distribution pattern of data. The three main shapes are normal, skewed, and uniform. Additionally, you may consider modality or the number of peaks.

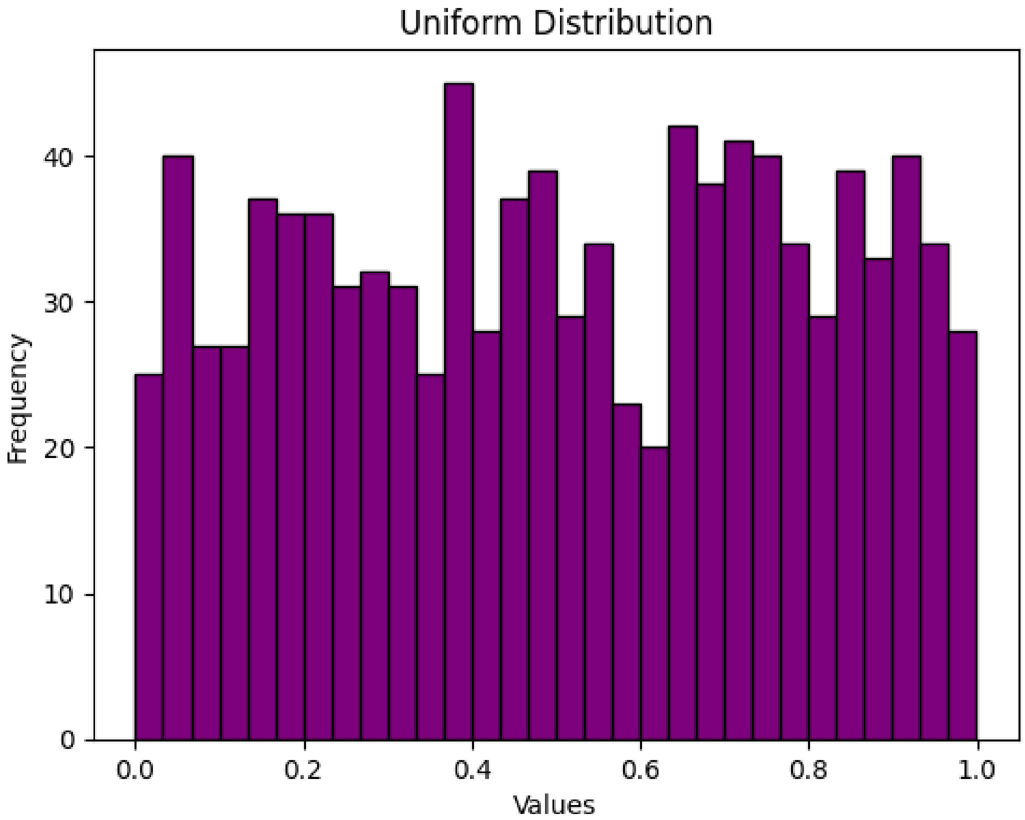
We might observe a normal distribution (bell-shaped curve) if the concentrations are evenly spread. Alternatively, a skewed distribution (left or right) could indicate a specific trend (higher or lower concentrations).

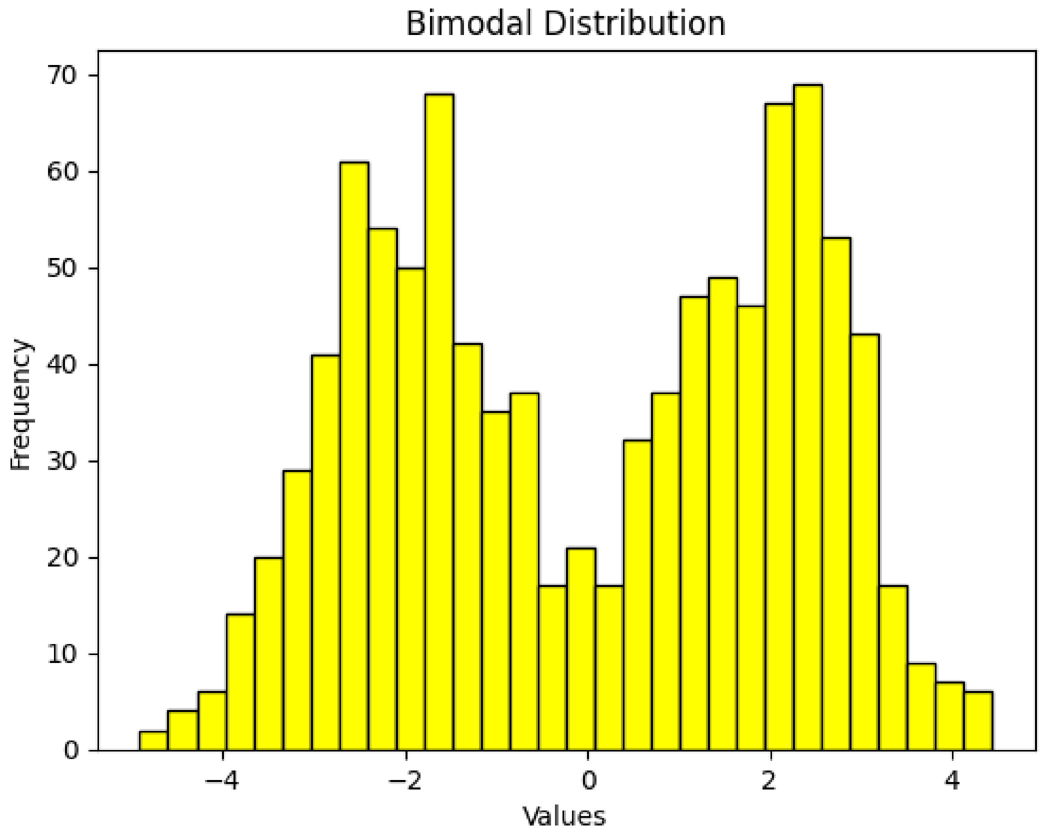
A uniform distribution would mean equal frequencies across all concentration levels. Modality refers to the number of peaks (high points) in a distribution. Multiple peaks indicates several distinct groups.



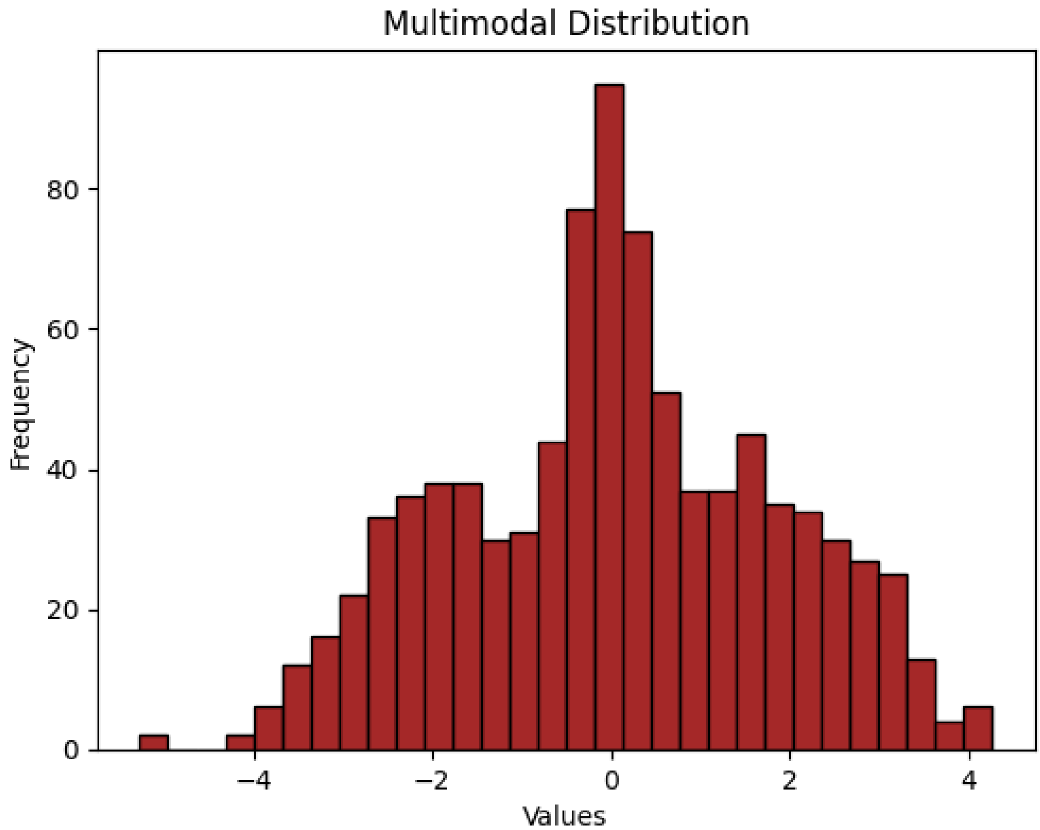



Right-Skewed Distribution
Frequency
Values
0
20
40
60
80
100
−3.5
−3.0
−2.5
−2.0
−1.5
−1.0
−0.5
0.0


Uniform Distribution
Frequency
Values
0
10
20
30
40
0.0
0.2
0.4
0.6
0.8
1.0



Bimodal Distribution
Frequency
Values
70
60
50
40
30
20
10
0
−4
−2
0
2
4

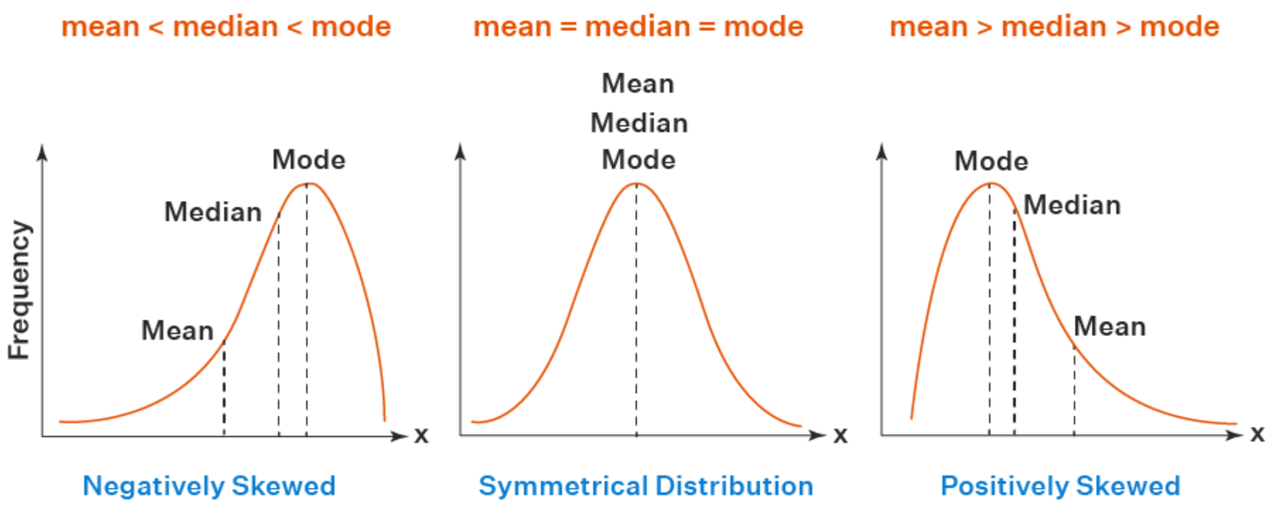


Multimodal Distribution
Frequency
Values
80
60
40
20
0
−4
−2
0
2
4

From our IgM Antibody Concentration Histogram earlier the histogram displays a bimodal shape, characterised by two distinct peaks at concentration levels 0.4 and 0.8. These prominent peaks correspond to the most frequently observed antibody levels in the sample. The bimodal shape suggests the presence of two distinct subgroups within the population, each exhibiting different antibody concentration patterns.

Centre

The centre of a histogram represents the median of the dataset. It identifies the typical value within the data and provides insight into the central tendency of the data distribution.



Empirical relation between mean, median, and mode (CueMath)

Spread

The spread refers to the range of the data. It shows how widely the data points are distributed across the histogram. A narrow spread indicates less variability, while a wide spread indicates a greater dispersion.

Outliers

Outliers are data points that significantly deviate from the rest of the data. They may indicate variability or errors in the dataset. Identifying outliers is essential for understanding extreme values and potential anomalies.

Next, we will focus on probability, a measure of the likelihood of an event occurring.

Spot check 1

Let's see what you can remember from this section.

1. What is the purpose of descriptive statistics?
2. What are the three main types of descriptive statistics?
3. What does a histogram represent?
4. How do you calculate the mean and median of a dataset?

Probability

We will focus on conditional probability and Bayes theorem in the sections that follow. Conditional probability focuses on the likelihood of an event occurring given that another event has already happened. Bayes' theorem, on the other hand, uses prior information and new evidence to update our beliefs about the probability of an event. While both concepts involve probability, Bayes' theorem provides a framework for revising our understanding of probabilities based on new data.

Conditional probability

Sometimes, we need to refine our calculations of probability depending on certain events. Conditional probability takes a look at two different but connected events and provides the probability of one event occurring given that the other event has happened.

For two connected events A and B, the formula for conditional probability is:

$$P(A|B) = \frac{P(AandB)}{P(B)}.$$

$P(A|B)$ is read as the probability of A given B. This is just maths talk for the probability of A assuming that B is true.

Let's take a look at an example: **the weather forecast**.

Imagine, you have Googled the weather forecast for your area. You see that there is a 40% chance of high wind and rain forecast for the day. However, you see that there is a 60% chance of high wind. This means that there is a possibility that there might just be high winds without any rain.

Later on, you find yourself taking your afternoon jog after a long and rewarding day of learning about statistics and probability. Suddenly, you feel the wind start to pick up. This means that the 60% chance of high wind has come true. You remember now that there was also a 40% chance of high wind and rain today. This is bad! It might start raining on you during your well-earned jog. Now that you know that there is high wind, what are the chances that you are about to get rained on?

Well, let's put it in terms of our equation above. Let's say that event A is the rain coming, and B is the weather having very high winds. We want to calculate the probability of A now that we know that B has happened. In other words, we are looking for **P(A|B)**. What do we know? We know that there is a 60% chance of high winds (i.e. **P(B)**). We also know that there is a 40% chance that there will be high winds and rain (i.e. **P(A and B)**). That means we only have to divide 40% by 60% to give us a 66.67% chance it will rain, now that we know that it is windy.

Bayes' theorem

Bayesian statistics is a very important concept in the field of machine learning and AI. Bayes' theorem enables AI systems to model and reason about uncertainty, making it indispensable for decision-making, pattern recognition, and predictive modelling. It helps us make informed decisions by combining prior knowledge with observed evidence. Whether in medical diagnosis, spam filtering, or other applications, Bayes' theorem plays a crucial role in AI and machine learning.

Because P(A and B) = P(B|A) P(A), the theorem states that:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}.$$

Groundbreaking, right? Well, this means nothing if we do not do something with it. So let's calculate the chances you are sick if you take a COVID-19 test.

Let's say you wake up one morning in 2020 with some suspicious symptoms. Nothing too bad, just a headache and a bit of a cough. There's lots of dust around, so it could just be the dust, right? Right? Hmm, you are not so sure. Let's take a look at your chances of having COVID-19!

At the moment 6.5% of the South African population are currently sick with COVID-19. So that means the chance you have of being infected is 6.5%, right? Yes, you are correct in saying that. Easy? Well, there is a lot of uncertainty in that answer. This cough is starting to persist. It is either some very serious dust or you could be legitimately sick. Best to be safe; let's rather get you tested.

You take an at-home antigen test to see if you are infected. While these are typically a bit inaccurate, it will at least give us a bit more of a clue as to whether you are infected. You use the cotton swab, put it in the testing receptacle, and wait ten minutes. That dreaded second line on the testing kit appears, which signifies that you have tested positive. Oh, the stress! You're going to have to go straight home and live on a diet of nothing but instant soup, painkillers, and sleep. But wait, how much can this test be trusted? After all, you have heard that these tests are not fully accurate. Let's do some digging into what happened.

First, let's take some time to identify what event A and event B should be in our formula. Because we know that we have tested positive, let's assume that event B is the event of us getting a positive test result (possibly inaccurately). Therefore, event A is the event where we are genuinely positive. We want to calculate P(A|B), the probability of us being genuinely infected given the fact that we tested positive.

We just need three values to calculate this probability: P(A), P(B), and P(B|A). How do we get these values? Let's consider the values one by one.

P(A) is the overall probability that we are positive. Funny enough, we already know this: 6.5% of the South African population is positive. This is our P(A).

P(B|A), read as the probability of B given A, simply means the probability that we test positive given that we are already infected. In other words, if we are infected, how do our odds of testing positive change? According to **this article**, we see that these tests correctly identify 96.2% of positive cases when the tests are taken within three days of symptom onset. In other words, positive cases are correctly identified as positive 96.7% of the time. Our P(B|A) is, therefore, 96.7%.

P(B) is simply the overall probability that we test positive at all. According to the study conducted, 7% of all tests given to participants tested positive, whether they were genuinely infected or not. Therefore, the overall probability of us testing positive, P(B), is 7%.

Now we have everything we need to calculate the probability that we are genuinely infected:

$$P(A|B) = \frac{(0.967)(0.065)}{(0.07)} = 0.897$$

That means the probability of being genuinely infected given the fact that we tested positive is 89.7%. Even with the uncertainty of these tests, that's still a high probability. Best to play it safe and stay at home!

Visual representation of Bayes' theorem

Explore the **logic and intuition behind Bayes' theorem** and see how geometry can help us update our beliefs systematically based on new evidence. By using relatable examples from psychology and geometry, this **video** demystifies this fundamental formula in probability theory.

Personalised movie recommendations: Suppose you are building a movie recommendation engine. Your goal is to suggest relevant movies to users based on their viewing history and preferences. Bayes' theorem can assist in this process. You will need the following elements and probabilities:

Consider the following events:

A: User enjoys a specific movie genre (e.g., action, romance, sci-fi).

B: User watches a movie from that genre.

Prior probability (P(A)):

You collect data on user preferences. Let's say 20% of users enjoy action movies (P(A) = 0.2).

Likelihood (P(B|A)):

You analyse viewing patterns. Among users who like action movies, 70% watch an action movie (P(B|A) = 0.7).

Overall evidence (P(B)):

Calculate the overall probability of watching any movie (regardless of genre).

Assume 50% of users watch a movie (P(B) = 0.5).

Apply Bayes' theorem:

- $P(A|B) = \dfrac{P(B|A)P(A)}{P(B)}$
- $P(A|B) = \dfrac{0.7 \cdot 0.2}{0.5} = 0.28$

The result of 0.28 (or 28%) is the likelihood that a user who watched a movie is genuinely interested in action movies.

This probability accounts for both the prior belief (prevalence of action movie fans) and the observed evidence (watching a movie).

In practical terms, if a user recently watched a film, there's a 28% chance they specifically enjoy action movies. The recommendation system can use this information to suggest relevant action films to enhance user satisfaction.

Spot check 2

Let's see what you can remember from this section.

1. What is conditional probability, and how does it differ from regular probability?
2. Question: In the weather forecast example, what does P(A|B) represent, and how do we calculate it?
3. Explain the significance of Bayes' theorem in AI and machine learning.Instructions

You will not be writing any code for this task but rather contemplating the concepts covered in this task. You will not have to do any calculations; we just want to see that you can apply these concepts to real life.

Practical task 1

Create a new text file called **statistics** that you will submit as a txt or pdf file. Then, write your answers to the following questions:

1. State whether the mean, median, or mode would be useful in the following scenarios:
   - You are doing population statistics. You are asked to give an estimate of the typical income of a single person in the country. There is one snag: wealth distribution is out of whack, and 10% of the population holds 70% of the nation's wealth.
   - You are running a restaurant, and you are reviewing your menu. You have a list of all orders over the last six months. You are trying to find out which item you should keep based on what customers seem to like the most.
   - You have been buying electricity once a month for the first six months of the year. You are trying to budget your electricity for the rest of the year and therefore need to estimate how much you will spend for the remainder of the year.

- You work in healthcare insurance. You are asked to provide an estimate of the typical amount of money spent on healthcare. This is taking into account the fact that there are a few people who spend a large amount of money on medical healthcare due to major issues.

2. State whether you would use variance or standard deviation to inform the following decisions:

    - You are choosing a new Internet provider. You find two providers with the same mean speed, but you want to have a more stable connection. You get a list of all reported speeds over the last month and are trying to find the provider that doesn't move too much from the mean value.
    - You are going on holiday to Mauritius. You need to find a shuttle from the airport to your hotel, but you are worried about being overcharged or undercharged (being undercharged might mean that you get unreliable transport). You get a list of all available shuttle service prices and need to find out which services, if any, are overcharging or undercharging.

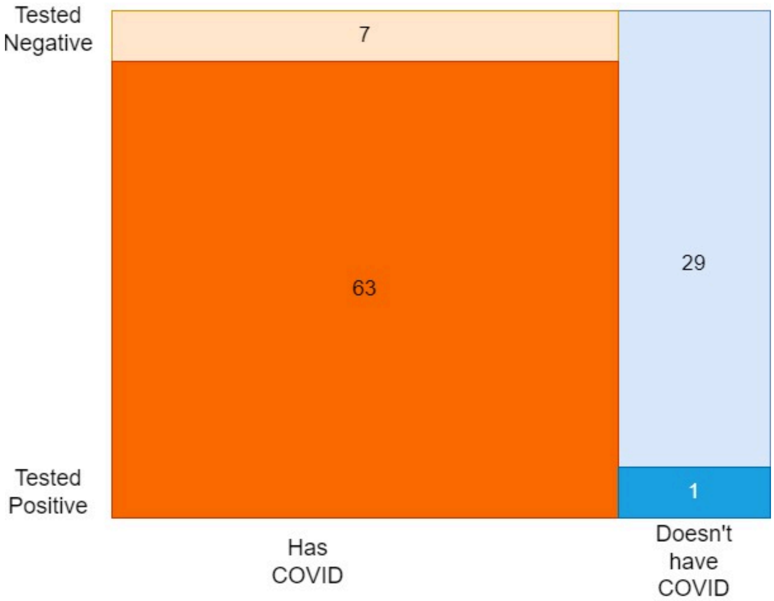
Practical task 2

Create a file called **conditional** that you will submit as a txt or pdf file. Then, write your answers for the following questions:

1. Given the values of **P(B)** and **P(A and B**) in the following scenarios. You are welcome to calculate **P(A|B)** if you choose to do so:

    - You work for a risk analysis insurer. You have read that this year, out of all drivers on the road, 5% have had accidents under the age of 25. You have also read that 10% of all drivers are under the age of 25. A new client approaches you and states that their age is 22. You want to calculate the chance that this driver has had an accident this year based on their age.
    - Your friend told you that they would buy you lunch if you can flip a coin and have it land on heads twice. You flip it the first time, and it lands on heads. What are your chances now of it landing on heads again?
    - You were always told that knowing Maths helps you to achieve 80% in Computer Science. You read some statistics showing that 30% of all Computer Science graduates took Maths and achieved 80%. Overall, 60% of all Computer Science graduates took Maths. Considering you took Maths, what are your chances of achieving 80%?

2. Consider a mock study about COVID diagnosis with a total of 100 participants. A visual representation of participants and their COVID and diagnostic test result status is given below.



The two orange areas show the total number of people who have COVID, and the two blue areas show the total number of people who do not have COVID. The two darkly-coloured areas at the bottom show the people who tested positive for COVID. The two lightly-coloured areas show the people who tested negative for COVID.

- Using this diagram, state the following:
    - H: our hypothesis
    - E: our evidence
- Then, give the values for the following:
    - P(H)
    - P(E|H)
    - P(E)
    - P(H|E)

Important: Be sure to upload all files required for the task submission inside your task folder and then click "Request review" on your dashboard.

Spot check 1 answers

1. What is the purpose of descriptive statistics?

    It summarises and describes data, allowing us to understand its central tendency, variability, and distribution.

2. What are the three main types of descriptive statistics?

    Distribution: the frequency of each value.

    Central Tendency: Involves averages (e.g., mean, median, mode).

    Variability or dispersion: Measures how spread out the values are (e.g., range, variance, standard deviation).

3. What does a histogram represent?

    A histogram displays the frequency of values within predefined intervals (bins). Each bar represents the frequency of values falling into a specific bin.

4. How do you calculate the mean and median of a dataset?

    Mean: Add up all the values and divide by the total number of values.

    Median: Arrange the values in ascending order and find the middle value (or the average of the two middle values if there's an even number of values).

Spot check 2 answers

1. What is conditional probability, and how does it differ from regular probability?

    Conditional probability focuses on the likelihood of one event occurring given that another event has already happened. It considers the context or condition provided by the second event. Regular probability, on the other hand, deals with the overall likelihood of an event without any specific conditions.

2. Question: In the weather forecast example, what does P(A|B) represent, and how do we calculate it?

In the weather forecast scenario, P(A|B) represents the probability of rain (event A) given that there are high winds (event B). To calculate it, divide the probability of both high winds and rain (P(A and B)) by the probability of high winds (P(B)).

3. Explain the significance of Bayes' theorem in AI and machine learning.

   Bayes' theorem combines prior knowledge (P(A)) with observed evidence (P(B|A)) to update our beliefs (P(A|B)). In AI and machine learning, it helps model uncertainty, aids decision-making, and enhances predictive modelling by incorporating new information.

Share your thoughts

Please take some time to complete this short feedback **form** to help us ensure we provide you with the best possible learning experience.

Reference list

IEEE. (1993). IEEE Standards Collection: Software Engineering. IEEE Standard 610.12-1990.

Cuemath (n.d.) Relation Between Mean Median and Mode. **https://www.cuemath.com/data/relation-between-mean-median-and-mode/**

Harmon, A., Chang, C., Salcedo, N., Sena, B. F., Herrera, B. B., Bosch, I., & Holberger, L. E. (2021b). Validation of an At-Home direct antigen rapid test for COVID-19. JAMA Network Open, 4(8), e2126931. **https://doi.org/10.1001/jamanetworkopen.2021.26931**

StatsDirect. (n.d.). Histogram (Frequency Distribution Plot). **https://www.statsdirect.com/help/graphics/histogram.htm**

Tamplin, T. (2023, August 5). Histograms | Definition, Characteristics, and How to Interpret. Finance Strategists. **https://www.financestrategists.com/wealth-management/fundamental-vs-technical-analysis/histograms/**