Hyperiondev

TASK

Statistics

Visit our website

Introduction

WELCOME TO THE STATISTICS TASK!

Machine Learning is essentially a complicated form of statistics. We aim to model data using statistics and predict data using probability. Knowing and understanding the basics of statistics and probability is essential to understanding how data science works. This task gives you an introduction to these topics and so gives you a good groundwork for understanding your data better!

WHAT IS THE DIFFERENCE BETWEEN STATISTICS AND PROBABILITY?

All data follows a statistical trend. Knowing and understanding the statistical trend can help us create a model that makes predictions. Consider a linear relationship between the X and Y values. This linear relationship is something statistical. Sometimes, getting a good understanding of the statistics behind the data is like being a detective. If your data is skewed left or right, this tells you something important about the data.

Probability is a simple concept at the surface level. If a computer can compute the probability of something, the prediction that the computer makes will inevitably be whatever the highest probability is. All machine learning models work with generating probabilities to some degree.

STATISTICS

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

1. Measures of centrality
2. Measures of spread

MEASURES OF CENTRALITY

Centrality in data typically describes where most of the data is. This is, admittedly, a very vague definition. That's 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 useful practical example of mean: **travelling from Durban to Pietermaritzburg**.

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've been monitoring your speed for the trip, and you've 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 at 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 traffic!)

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**
- 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.

Let's look at an example: you have three values in **X**. Following this algorithm, 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: **school marks**.

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 got a solid **73** for this.
- Accounting was like Maths, but not as fun. You got a middling **65** for this.
- You didn't 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're 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. Darn, it's 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: **movies**.

You have somehow managed to become the new CEO of 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. There is only one way to find the mode: count the number of occurrences in each type of category, and the mode is the category with the highest count. Sparing you the pain of counting, 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

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. The big time, working in the Big Apple. You are managing many varied stock portfolios. You are approached by a new investor who wants to find a nice, low-risk stock in the coding bootcamp industry. You consider two stocks:

- HyperionDev: 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 HyperionDev 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 HyperionDev, 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 at all? 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 weirdly 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: **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 data set 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.

PROBABILITY

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**.

You have read the weather forecast or (if you're anywhere under the age of 50) probably 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. It is a simple expansion of the basic formula for conditional probability.

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 don't do something with it. So let's do something with it: take a COVID-19 test and calculate the actual chances you are sick.

Let's say you wake up one morning 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're not so sure. Let's take a look at your chances of having COVID-19!

At the time of writing, 6.5% of the South African population is 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's 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'll 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 3 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!

Note from our team about the task

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

Compulsory 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:

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.

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.

Compulsory Task 2

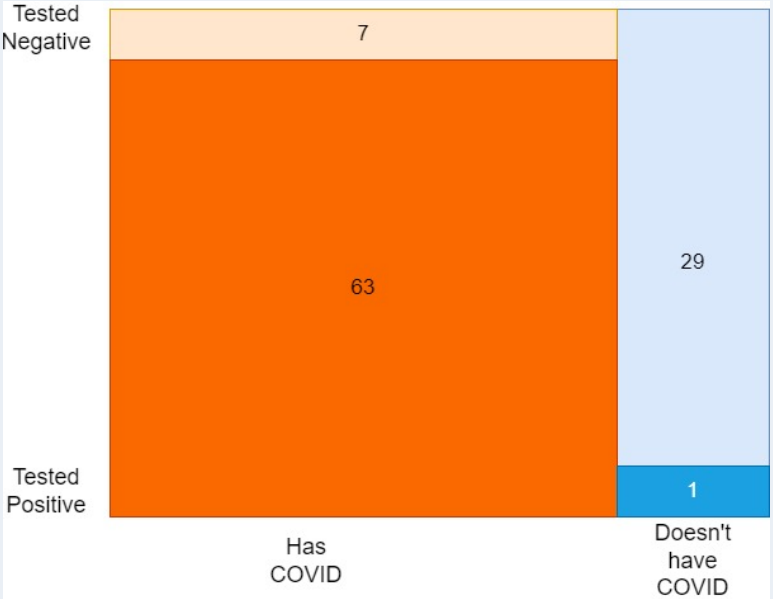
Create a file called **conditional** that you will submit as a txt or pdf file. Then, write your answers for the following questions:

Give 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%?

Watch **this video** on visualising Bayes' Theorem and understanding it visually. We recreated one of these diagrams, which you can see below:



This is a mock study created using a total of 100 participants. The two orange areas show the total number of people who *actually* have COVID, and the two blue areas show the total number of people who *don't actually* 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, and information learned from the video, 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)**

Rate us
Share your thoughts

HyperionDev strives to provide internationally-excellent course content that helps you achieve your learning outcomes.

Think that the content of this task, or this course as a whole, can be improved, or think we've done a good job?

Click here to share your thoughts anonymously.

REFERENCES

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