Maths 101 Handbook

Fundamental Maths Concepts for Data Science

Grid Paper • 8.3 x 11.7

Duke University: Data Science Math Skills

Personal Notes from Coursera

DISCLAIMER

ABOUT:

Personal notes converted into a short Handbook, which are supplemental to the original online course published by Coursera titled “Data Science Maths Skills”, on behalf of Duke University.

ORIGINAL AUTHOR(S):

 Listed author(s) in the order they appear on Coursera: Daniel Egger, Paul Bendich.

NOTE:

You cannot use this short Handbook in place of the original reading materials. Obtaining and or reading this short Handbook does not qualify you as having completed the original online course.

SUGGESTIONS:

For best results, first sign-up to Coursera and complete the original online course created by Duke University listed above. Once you have completed the course, you are then welcome to refer to this short Handbook as a refresher.

- Maths 101 Handbook Series -

https://github.com/NellieAT

2021

Sets

A set is a collection of any number of elements or things,
notation of '$\in$' means 'is an element of'
while '$\notin$' means 'is not an element of'

The cardinality (size) of a set, A, is the number of elements in that set, notation of '|A|' means 'the size of set A', so in this example: $|A| = 5$, there are 5 elements in that set A.

$A = \{1, 2, -3, 7\}$ $B = \{2, 8, -3, 10\}$ $C = \{5, 10\}$

$= \{2, -3\}$ ← $A \cap B$ intersect $B \cap C \rightarrow = \{10\}$

$A \cup B$ ↖ union $B \cup C$ ↘

$= \{1, 2, -3, 7, 8, 10\}$ $= \{2, 5, 8, -3, 10\}$

Note: if there are no common elements then use '$\emptyset$' to notify that the set is empty where $|\emptyset| = 0$, so the cardinality of the empty set is zero.

In general terms, the intersect and union between two sets can be written as:

$A \cap B = \{x : X \in A \text{ and } X \in B\} = \{2, -3\}$

$A \cup B = \{x : X \in A \text{ or } X \in B\} = \{1, 2, -3, 7, 8, 10\}$

This general notation is useful for complicated situations as it defines the conditions of a set, without listing them all out.

Sets (Example I)

Medical testing example, VBS, very bad syndrome where C is the set of people in the clinical trial, so sick and healthy people can be classified as:

$S = \{x \in C : x \text{ has VBS}\}$
$H = \{x \in C : x \text{ does not have VBS}\}$

where $C = S \cap H$, as the total population of the trial are either sick or healthy, they cannot be both, $S \cap H = \emptyset$, where S intersect H is an empty set.

$P = \{x \in C : x \text{ VBS positive}\}$
$N = \{x \in C : x \text{ VBS negative}\}$

where $P \cup N = C$, which is everyone, while $P \cap N$ is noone. In an ideal world $S = P$, sick people test positive, and $H = N$. But in life that rarely happens.

$S \cap P$ = true positive $H \cap P$ = false positive
$H \cap N$ = true negative $S \cap N$ = false negative

$\frac{|S|}{|C|}$ = proportion of people in clinical trial with VBS
$\frac{|H|}{|C|}$ = proportion of people in trial without VBS
} should equal 1.

Studying the cardinality of these sets can provide new interesting insights. The following types of sets are used often in machine learning:

Sets (Example II)

For our medical example, we can look at some relationships:

$\frac{|S \cap P|}{|S|}$ = true positive rate

$\frac{|H \cap N|}{|H|}$ = true negative rate

} want these as big as possible, where they tend to 1

$\frac{|H \cap P|}{|H|}$ = false positive rate

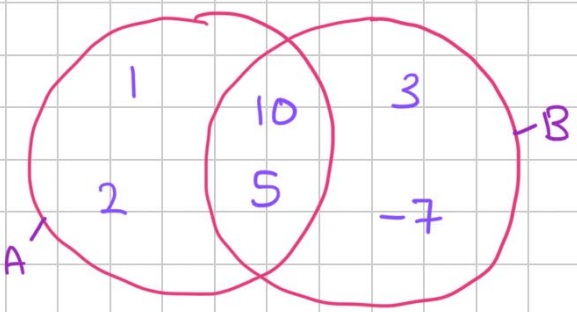
$\frac{|S \cap N|}{|S|}$ = false negative rate

} want these as small as possible, where they tend to 0

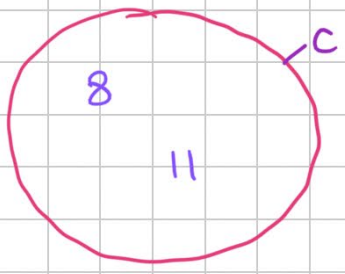
In real world examples, false positives and false negatives exist, so we need to come up with subject specific thresholds that are acceptable for each, in order to validate the total outcome.

Venn diagrams are a way of visualising commonality in sets

$A = \{1, 10, 5, 2\}$ and $B = \{5, -7, 10, 3\}$ and $C = \{8, 11\}$



$A \cap B = \{5, 10\}$



Set C is disjoint from both A and B, where

$A \cap C = \emptyset$ and $B \cap C = \emptyset$

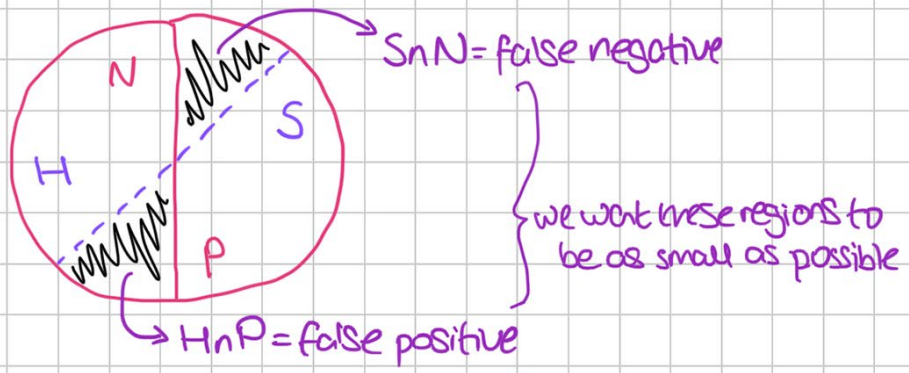
Sets Inclusion/Exclusion Formula

The inclusion/exclusion formula is given as:

$$|A \cup B| = |A| + |B| - |A \cap B|$$

so from the prior venn diagram $|A \cup B| = 6$, $|A| = 4$, $|B| = 4$, $|A \cap B| = 2$
$6 = 4 + 4 - 2$, which is true!

For our medical example we can visualise a venn diagram too



It is clear to see that $H \neq N$ and $S \neq P$ when visualised

The 'absolute value' of a number x is given as '$|x|$' (don't confuse with cardinality) which describes the distance from 0 to x, on the real number line.

General rule, for any real number $x \in \mathbb{R}$, where $\mathbb{R}$ denotes real number then the absolute value of x, or $|x|$, can be 1 of 2 things

$|x| = \{x\}$ if x is non-negative, ie $|8.7| = 8.7$
$|x| = \{-x\}$ if x is negative, ie $|-10.3| = 10.3$

Next we look at the types of intervals we can have on the real number line. These intervals become important when defining limits that may describe thresholds in certain calculations.

Intervals and Sigma Notation

A closed interval is denoted by square brackets

$[2, 3{\cdot}1] = \{x \in \mathbb{R} : 2 \leq x \leq 3{\cdot}1\}$ closed use inclusive $\leq$ and $\geq$

An open interval is given by round brackets

$(5, 8) = \{x \in \mathbb{R} : 5 < x < 8\}$ open use exclusive $<$ and $>$

While half-open intervals have both types of brackets

$(-7{\cdot}1, 15] = \{x \in \mathbb{R} : -7{\cdot}1 < x \leq 15\}$

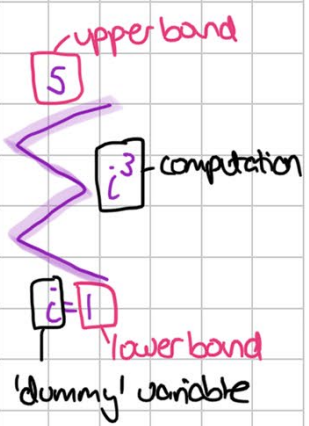
$[20{\cdot}3, 48) = \{x \in \mathbb{R} : 20{\cdot}3 \leq x < 48\}$

Sigma Notation: Σ means the 'sum' of an individual computation within a specified range

$$\sum_{i=1}^{4} i^2 = 1^2 + 2^2 + 3^2 + 4^2 = 30$$

$$\sum_{i=1}^{5} (2i + 3) = 5 + 7 + 9 + 11 + 13 = 45$$

$$\sum_{j=3}^{7} j/2 = \frac{3}{2} + \frac{4}{2} + \frac{5}{2} + \frac{6}{2} + \frac{7}{2} = \frac{25}{2}$$



Summation Problems

Summation problems can be split up into smaller unit summation problems when there is an addition between components, as numbers can be added in any order:

$$\sum_{i=1}^{5}(i^2+2i) = \sum_{i=1}^{5} i^2 + \sum_{i=1}^{5} 2i = 55 + 30 = 85$$

Note - if there are no index dependancies, just multiply the constant by the range specified:

$$\sum_{i=1}^{9} 5 = 5+5+5+5+5+5+5+5+5 = 45 \quad \text{(which is the same as } 5\times 9\text{)}$$

Sigma notation of mean, variance and standard deviation

$Z = \{1, 5, 12\}$, so $|Z| = 3$, then $\mu_Z = 6$, which is the mean of Z

the mean can be expressed in summation form:
where the reciprocal of the cardinality of the set is multiplied to the summation calculation.

$$\frac{1}{3}\sum_{i=1}^{n} xyz_i$$

In general, if $x = \{x_1, x_2, x_3, \ldots x_n\}$ then the mean is given as

$$\mu_x = \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)$$

However, the mean is not always a reliable or unique classifier of a set, ie $A = \{1, 5, 12\}$ $\mu_A = 6$, while $B = \{5, 6, 7\}$ $\mu_B = 6$, both sets have the same mean...
Perhaps another metric can provide further insight.

Variance and Standard Deviation

The variance gives us some more information about a set of data. It represents how spread out the data in the set is. The expression is squared as we don't care if the spread is negative or positive, just need to know the distance from the mean.

$$\sigma_x^2 = \frac{1}{n}\left[\sum_{i=1}^{n}\left(x_i - \mu_x\right)^2\right]$$

variance example: $\omega = \{5, 6, 7\}$ where $\omega_1 = 5, \omega_2 = 6, \omega_3 = 7$ and $\mu_\omega = 6$ and $|\omega| = 3$, plugging this into the equation above,

$$\sigma_\omega^2 = \frac{1}{3}\left[\sum_{i=1}^{3}\left(\omega_i - \mu_\omega\right)^2\right] \text{ which expands to}$$

$$= \frac{1}{3}\left[(5-6)^2 + (6-6)^2 + (7-6)^2\right]$$

$$= \frac{1}{3}\left[(-1)^2 + (0)^2 + (1)^2\right]$$

$$= \frac{1}{3}[2] \text{ or } \frac{2}{3}.$$

so the variance for set $\omega = \frac{2}{3}$, that is how spread out the data is

Now the standard deviation is just the square-root of the variance, so for set ω, the std. dev is given as $\sigma_\omega = \sqrt{\frac{2}{3}}$
A general form of the formula is usually expressed as:

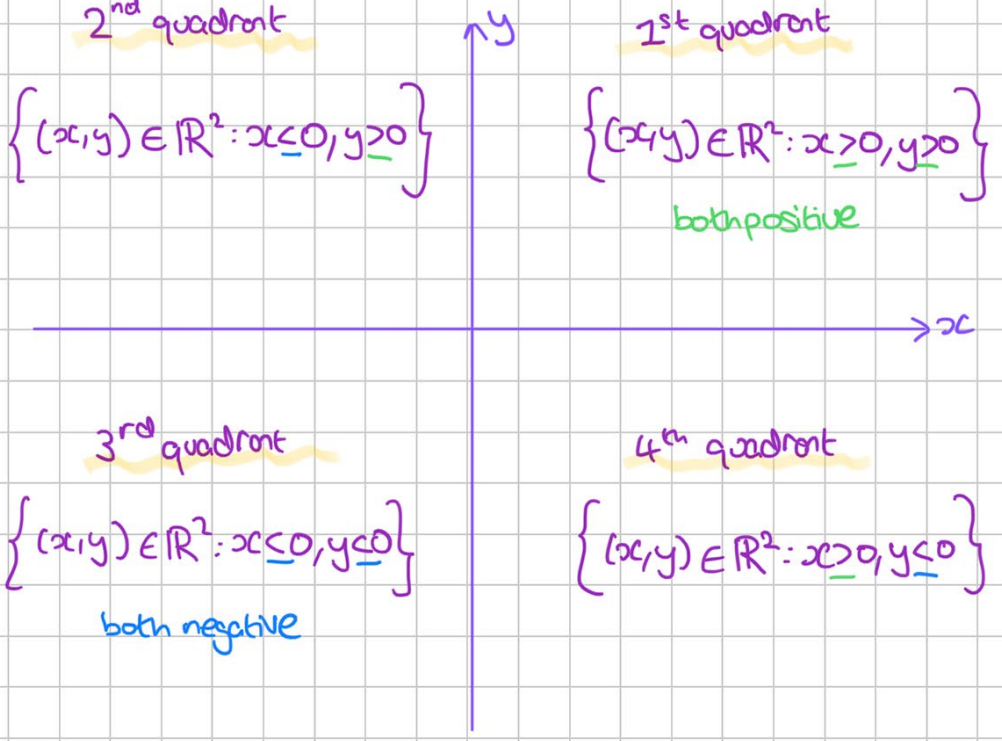
$$\sigma_x = \sqrt{\frac{\sum_{i=1}^{n}\left(x_i - \mu_x\right)^2}{n}}$$

Real Numbers

Real numbers are denoted by symbol $\mathbb{R}$ and the cartesian plane is represented as $\mathbb{R}^2$ note the cartesian plane quadrants are arranged anti-clockwise

$$x\text{-axis} = \{(x,y) \in \mathbb{R}^2 : y=0\}$$

$$y\text{-axis} = \{(x,y) \in \mathbb{R}^2 : x=0\}$$



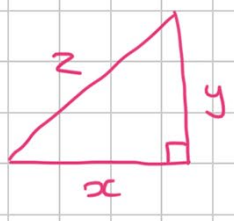
The real numbers can be anywhere is this 2D cartesian plane, each point on the plane will be described by either being on the x-axis or y-axis, or within one of the quadrants.

Distance in the Plane

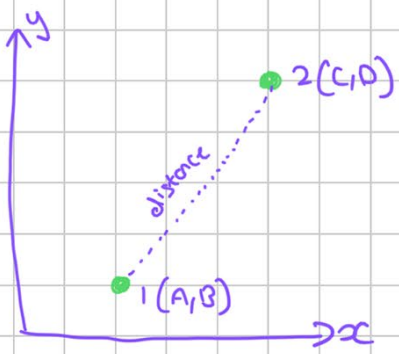
We will summarise the three main concepts below:
- the distance formula
- nearest neighbours (supervised learning)
- clustering (unsupervised learning)

The Pythagorean Theorem can give us the distance in the plane

$$z^2 = x^2 + y^2$$
$$\text{or } z = \sqrt{x^2 + y^2}$$



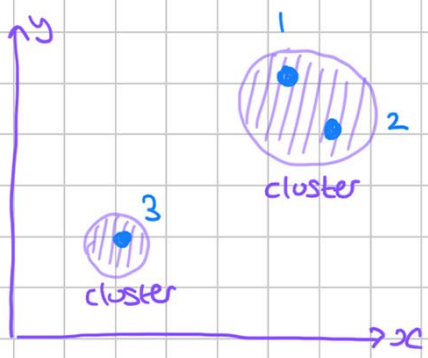
where z is known as the hypotenuse



The distance between two points on the plane, point 1 and point 2 are given as:

$$\text{Dist}(1,2) = \sqrt{\underbrace{(C-A)}_{\Delta x}^2 + \underbrace{(D-B)}_{\Delta y}^2}$$

Distance is a good way of defining membership to a cluster, in data science, where the distance between points within the same cluster is the smallest, relative to the distance between other points that belong to different clusters. This means we interpret points with a smaller distance as more similar.



where $\text{dist}(1,2) << \text{dist}(1,3)$ and $<< \text{dist}(2,3)$

Equation of a Line

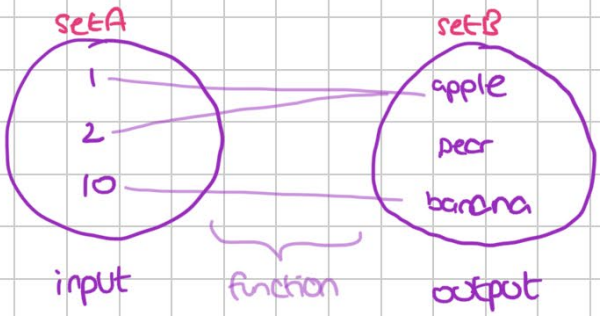
Point-slope formula:
If a line 'l' has a slope 'm' and if (x_0, y_0) is any point on that line, then l has the equation: $y - y_0 = m(x - x_0)$

Slope-intercept formula:
If a line 'l' has a slope 'm' and l crosses the y-axis at the point $(0, c)$, then the equation $y = mx + c$ is used to describe any true value point on the line

Mapping between sets, functions can be used to convert from one set to another, so a function that takes $f: A \to B$, can be a rule, function, machine which transforms each element $a \in A$ into $f(a) \in B$.
Where element a of set A is now transformed to be the result of function a which is now an element of set B.
So 'a' is the input and 'f(a)' is the output.

$A = \{1, 2, 10\}$

$B = \{\text{apple, pear, banana}\}$



Assume a function transforms set A into set B, so $f: A \to B$, which converts f(1): apple, f(2): apple, f(10): banana

supervised learning, is the process of figuring out the mysterious function from a few examples of inputs and outputs. Often you will be given some sample of an input set and output set, from which you obtain the method or rule that transforms the set from one to another. $f: A \xrightarrow{?} B$, so you may have some $a \in A$ and some $f(a) \in B$ from which you need to work out the relationship, trend or pattern.

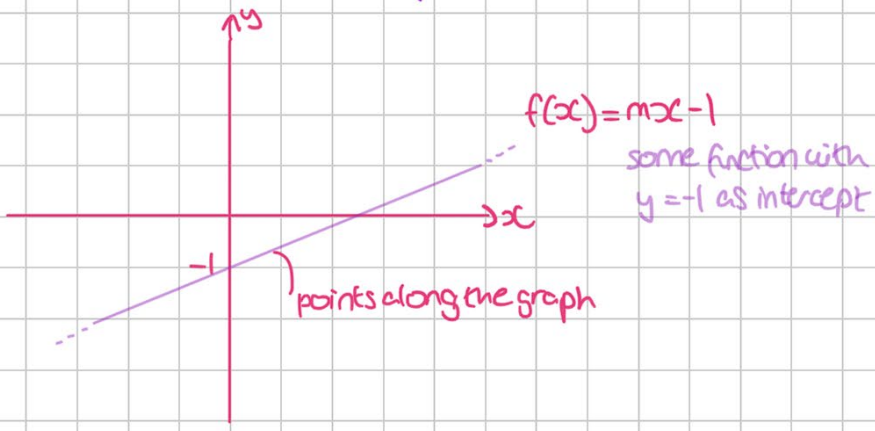
Function and Graphs

The actual function and the graph of that function a 2 separate things:

$$\text{graph}(f) = \{(x, y) \in \mathbb{R}^2 : y = f(x)\}$$

↑ visual

↑ satisfies function

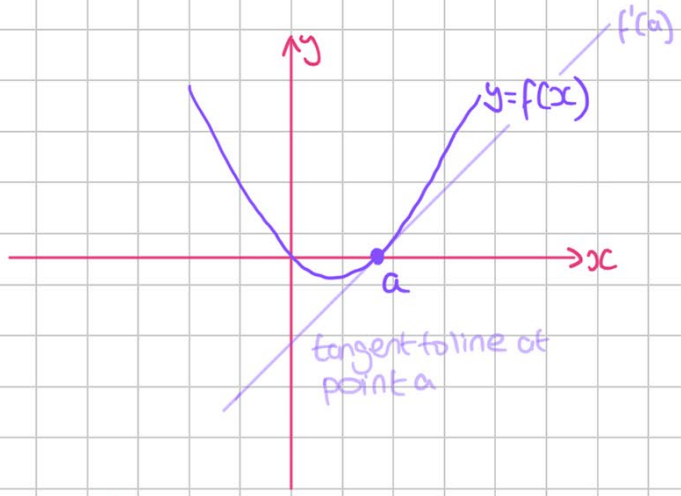


The vertical line test is a way to filter out impossible functions and the graphs of those functions. As a function cannot cut the y-axis twice, as it violates the basic operation of a function that strictly specifies the transformation of set A to set B, with no ambiguity. You cannot have 1 input converted into 2 outputs.

The horizontal line test is a way to figure out if a function is always increasing, decreasing, or neither. known as: strictly increasing, strictly decreasing or neither. If any horizontal line intersects the graph of a function once, it could be SI or SD, but if the line crosses the graph more than once, it is clearly not SI or SD.

Note, that if a graph of (f) fails the horizontal line test, then (f) cannot have an exactly opposite inverse function. So the only 'invertible' functions are ones that are strictly increasing, SI or strictly decreasing, SD.

Graphs and Limits



The slope of the tangent line gives only the instantaneous rate of change

This is also known as the derivate of the function at that specific point. The gradient is the derivative.

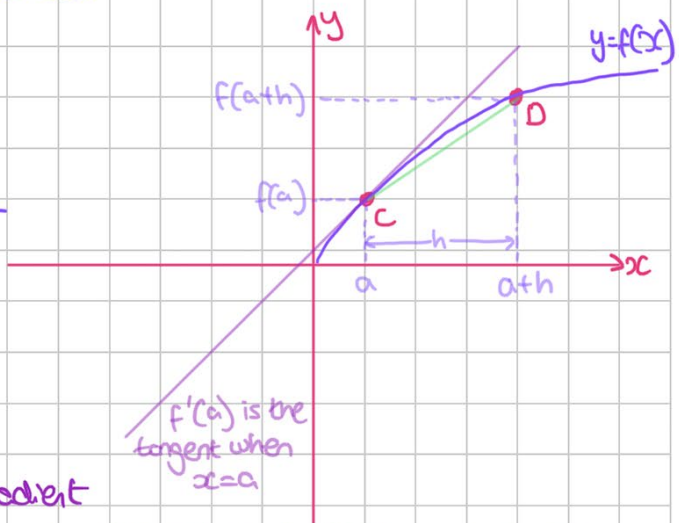
if $y=f(x)$ is the function

then $f'(a) = \lim_{h \to 0} \frac{f(a+h)-f(a)}{h}$ is the slope of the tangent line when $x=a$

The coordinates of points C and D are:

$C = (a, f(a))$

$D = ((a+h), f(a+h))$



The line segment $= \frac{f(a+h)-f(a)}{(a+h)-a}$

$= \frac{f(a+h)-f(a)}{h}$

We cannot simply calculate the gradient of the tangent $f'(a)$. But we can work out the slope of a line, so we choose a small distance, h, away from point a, then draw a line segment between them. We can work at the slope of that line segment. This is clearly not the final answer, but when $h \to 0$, the line segment between the two points decreases, the answer becomes more accurate.

Functions, Gradients and Limits

Suppose have the example function $f(x) = x^2$
to find the gradient of the tangent, we work through the algebra

$$f'(a) = \lim_{h \to 0} \frac{f(a+h) - f(a)}{h}$$ is the general formula to use and substitute in

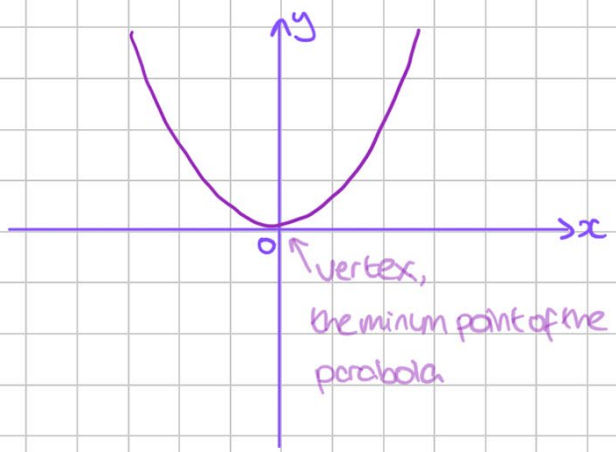
$$= \lim_{h \to 0} \frac{(a+h)^2 - a^2}{h}$$ in our example, we square the function

$$= \lim_{h \to 0} \frac{a^2 + 2ah + h^2 - a^2}{h}$$ cross out terms and factorise

$$= \lim_{h \to 0} \frac{h(2a+h)}{h}$$ cross out to simplify

So the limit for this example, when $h \to 0$ is just $2a$
This means that $f'(a) = 2a$, we now have a numerical value as the gradient of the tangent, at a specific point. $f(x) = x^2$, while $f'(x) = 2x$, this is known as the derivative function

Also, note that at the origin $(0,0)$ the derivate function $f'(x) = 2x$ becomes $f'(0) = 2 \times 0 = 0$, which means $y = 0$. This is correct as $y = 0$ is the horizontal x-axis and satisfies the geometry of the function $y = x^2$



Logarithms and Exponents I

These two concepts are related to eachother, they link the base, the exponent and the result.

$b^{x} = N$ — exponent form

"b raised to the power x is equal to N"

$x = \log_b(N)$ — logarithmic form

"x is equal to log to the base b of N"

The relationship between base, exponent and the result can be written in either form.

Note, any number raised to the power 0 is equal to 1, ie. $(2\pi)^0 = 1$
any log of any base of 1 is also 0, ie. $\log_{20}(1) = 0$

There are 3 logarithmic simplification rules: Product, Quotient, Power and Root

Product Rule:

$\log(xy) = \log(x) + \log(y)$ ⟶ factors can be added

$\log(35) = \log(7) + \log(5)$

Quotient Rule:

$\log(x/y) = \log(x) - \log(y)$ ⟶ denominator subtracted from numerator

$\log(64/2) = \log(64) - \log(2)$

Power and Root Rule:

$\log(x^n) = n\log(x)$ ⟶ power/roots are multiplied

$\log(\sqrt{x}) = \log(x^{\frac{1}{2}}) = \frac{1}{2}\log(x)$

Logarithms and Exponents II

Further compound examples:

$\log_2(16/4) = \log_2(16) - \log_2(4) = 2^4 - 2^2$

$\log_2(1{,}000)^{\frac{1}{3}} = \frac{1}{3}\log_2(1{,}000)$

$\log_{10}(7)^5 = 5\log_{10}(7)$

$\log_b(x^2 \cdot y^{-3}) = \log_b(x^2) + \log_b(y^{-3}) = 2\log_b(x) - 3\log_b(y)$

$\log_b\left(\frac{x^2}{y^{-\frac{1}{2}}}\right) = \log_b(x^2) - \log_b(y^{-\frac{1}{2}}) = 2\log_b(x) + \frac{1}{2}\log_b(y)$

Converting the base of logarithms:

Assume previous base is 10, where we have $\log_{10}(12)$, while new base is 2, so we want to get: $\log_2(12)$. first you need to divide the result of the old base, by the result of the new base, to get the actual final result of the new base.

If: $\log_{10}(12) = 1.079$
then: $\log_{10}(2) = 0.30103$
so: $\log_{10}(12) \div \log_{10}(2) = 3.585$
The final answer of $\log_2(12) = 3.585$

$$\log_a(b) = \frac{\log_x(b)}{\log_x(a)}$$

↑ General rule for converting the old base x, into new base a

Logarithms and Exponents III

Original base 2 → desired base 10

$\log_2(7)$ —convert bases→ $\log_{10}(7)$

$= 2.8073$ = desired result

Apply the conversion formula, by first calculating $\log_2(10)$ which is $= 3.3219$
so $2.8073 \div 3.3219 = 0.8540$

where: $\frac{\log_2(7)}{\log_2(10)} = \log_{10}(7)$ $\quad \frac{2.8073}{3.3219} = 0.8540$

You can have a discrete exponential rate of growth, or a continuous exponential rate of growth. Important to know the difference in data science.
continuous exponential growth rate involve euler's constant 'e' $= 2.71828$

Assume a baby elephant weighs 200kg, has continuously compounded growth rate, can use e to work out what weight it would be in 3 years:
$200 \times e^{(0.05)(3)}$ where 0.05 represents the continuous growth rate of 5% shown in decimal form, multiplied by the time period of 3 years.
The result of that calculation is then 232.4kg.

Note: log to the base of e or $\log_e = \ln(x)$ which is the 'natural log'
$\log_e = \ln(x)$. so a logarithm is related to the natural log through euler's constant. It is called 'natural' as most uses of it are associated with naturally occuring continuous rates of growth.

Logarithms and Exponents Problems

An investment of £1,600 is worth £7,400 after 8.5 years, what is the continuous compounded rate of return of the investment?

$\ln\left(\text{end}/\text{start}\right) = \ln\left(7400/1600\right) = \ln(4.625)$

$\ln(4.625) \div 8.5 = 0.18017$ ← divide by timeframe

$\therefore$ 18.02% is the growth rate of the investment

Want to know the time it takes for mass of rabbits equals the mass of Earth. Assume 1 pair of rabbits = 10kg. Then assume mass grows by a rate of 200%. The Earth weighs $\sim 5.972 \times 10^{24}$ kg. Remember rate of 200% is the same as multiplying by 2.

$5.972 \times 10^{24} = 10e^{2t}$ ← set up continuous equation

$5.972 \times 10^{23} = e^{2t}$ ← divide both sides by 10

$\ln(5.972 \times 10^{23}) = \ln(e^{2t})$ ← take natural logs of both sides

$\ln(5.972 \times 10^{23}) = 2t$ ← ln and e cancel eachother out

$\frac{\ln(5.972 \times 10^{23})}{2} = t$ ← rearrange to find unknown

Evaluating this situation gives the result of 27.37 years in order for the rabbit population to weigh as much as the Earth, assuming the parameters.

$5.972 \times 10^{24} = 10e^{2t}$ ← known growth rate, unknown time

↑ known end weight

↑ known start weight

There are many such problems where the rate or timeframe of growth, or change needs to be calculated. Key is to set up the calculation carefully.

Probability Distributions I

A probability distribution is a collection of statements, two or more, where those individual statements are both exclusive and exhaustive - when we have complete information.

↳ Exclusive = no more than 1 statement can be true

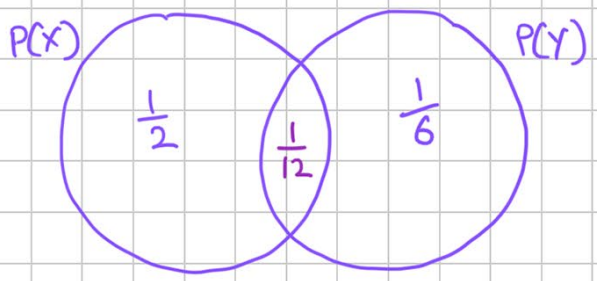
↳ Exhaustive = at least 1 statement from total must be true

$$P(\text{event}) = \frac{\text{outcomes defined by the event}}{\text{total outcomes in universe}}$$

$P(\text{queen card in shuffled deck}) = \frac{4}{52} = \frac{1}{13}$ given there are 4 queen cards in deck

Joint probabilities of $P(A)$ and $P(B)$ occuring can be written as $P(A,B)$, where ordering within brackets does not matter. It is independent when the joint distribution $P(x,y)$ = product distribution $P(x) \cdot P(y)$.

$P(x)$ = probability of getting heads in a coin toss
$P(y)$ = probability of getting 3 in a dice roll
} independent probabilities



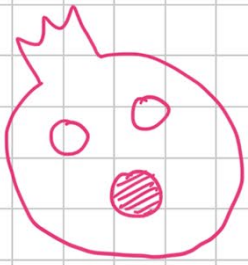
For both probabilities to be true: $P(X) \cap P(Y)$, multiply individual events together. If we need: $P(X) \cup P(Y)$, so one or the other needs to be true, then we use the inclusion/exclusion formula, subtracting probability of joint distribution so: $P(X \text{ or } Y) = P(X) + P(Y) - P(X,Y)$, which avoids double counting the intersection. In our case above: $\frac{1}{2} + \frac{1}{6} - \frac{1}{12} = \frac{6}{12} + \frac{2}{12} - \frac{1}{12} = \frac{7}{12}$.

Probability Distributions II

Permutations = order matters, where $\frac{n!}{(n-m)!}$ n = unique objects, m = attributes

Combinations = order does not matter, where in this case $\frac{n!}{(n-m)! \times m!}$

In many situations, there are 2 ways to draw from a process, either random or sequential. With replacement (**independent**), but without replacement it is known as (**dependent**).



with replacement: there is always $\frac{2}{3}$ chance for white and $\frac{1}{3}$ chance for pink.

without replacement: If 1st draw is white, then second white chance is reduced to $\frac{1}{2}$.

A **factorial** is a type of arithmetic that multiplies an integer by another integer one less than the first, until you reach 1.
ie. $5! = 5 \times 4 \times 3 \times 2 \times 1 = 120$. This is pronounced '5 factorial' is equal to 120
Note: $0! = 1$.

"m choose n"

$$\binom{m}{n} = \frac{m!}{(m-n)!\,n!}$$

the number of distinct groups of n items drawing from m items, without replacement

So if $m = 10$ and $n = 5$, it would be "10 chooses" $\binom{10}{5} = \frac{10!}{(5!)5!}$
In practice an example would be, how many teams of 5 people can be formed from a population of 10 people? it would be $\binom{10}{5}$.

Probability Distributions III

Marginal probability, or the table of known joint probabilities is equal to the sum of joint probabilities.

marginal events

marginal events	x_1	x_2	x_3	
y_1	$P(x_1,y_1)$	$P(x_2,y_1)$	$P(x_3,y_1)$	
y_2	$P(x_1,y_2)$	$P(x_2,y_2)$	$P(x_3,y_2)$	

ie. $P(x_1) = P(x_1,y_1) + P(x_1,y_2)$
= binary
ie. $P(y_2) = P(x_1,y_2) + P(x_2,y_2) + P(x_3,y_2)$
= non-binary

The sum rule for binary probability distribution : $P(A) = P(A,B) + P(A,\sim B)$
The sum rule for non-binary or n-series probability : $P(A) = P(A,B_1) \dots + P(A,B_n)$
ie. for $P(y_2)$ above, it is non-binary, where 1 of the element changes or iterates

Conditional probability : the probability that a statement is true, given that some other statement is true - with certainty. $P(A|B)$ means the probability of A given B is true with certainty, these are relationships with dependence.

$$\frac{\text{relevent outcomes that satisfy } P(A)}{\text{total outcomes when } P(B) \text{ is true}}$$

combining concepts of : joint probability, marginal probability and conditional probability, helps us to derive **The Product Rule** : $P(A|B) = \frac{P(A,B)}{P(B)}$

Expressed fully: The conditional probability of A, given the probability of B is true with certainty, is equal to the probability of both A and B being true, also known as the joint probability, divided by the marginal probability that B is true.
Note: the marginal probability of B being true, does not have to equal 1, we do not assume it to be true - Only the conditional probability statement, which is denoted by the vertical line assumes B is true with certainty.

Bayes' Theorem I

Previously we said the joint distribution was equal to the product distribution where $P(X,Y) = P(X) \cdot P(Y)$ now we have a new definition of independence so first divide both sides by $P(Y)$: $\frac{P(X,Y)}{P(Y)} = \frac{P(X) \cdot P(Y)}{P(Y)}$ which simplifies to $= P(X|Y) = P(X)$

This means that knowing $P(Y)$ is true with certainty does not always reveal information on $P(X)$ being true. so $P(Y)$ has no affect on $P(X)$ making them independent. However, when $P(X|Y) \neq P(X)$, then the events are dependent.

Bayes' Theorem is a key concept of probability theory:
first start with the product rule $P(A|B) = \frac{P(A,B)}{P(B)}$
then multiply both sides by $P(B)$
to get: $P(A|B) \cdot P(B) = P(A,B)$
substitute $P(B,A)$ for $P(A,B)$ on the right hand side of the equation
$P(A|B) \cdot P(B) = P(B,A)$ } inverse equivalents
$P(B|A) \cdot P(A) = P(A,B)$
Then rearrange to finally get the usual form of Bayes' Theorem: $P(A|B) = \frac{P(B|A)P(A)}{P(B)}$

Bayes Theorem useful for inverse probability problems.
where 'B' = observed data, and 'A_i' = possible process 'i', with probability parameter 'θ_i' pronounced theta sub i. Note that A_i and θ_i are regarded as equivalent, usually.

Assume the following: urn 1 = A_1, $P(\text{white marble}) = 20\%$
urn 2 = A_2, $P(\text{white marble}) = 10\%$
We observe 3 white marbles in a row, drawing with replacement.
(Q) what is the probability we are observing marbles taken from either urn 1 or urn 2?
↳ The full calculation will be shown in the next section.

Bayes' Theorem II

In the example problem previously shared, we have the known outcomes (the color of the marbles drawn), so we need to find out the probability that we observed either urn 1 or urn 2, via an unknown process. Note that in normal traditional probability questions we know the process but we dont know the outcomes.

$$P(A_i|B) = \frac{P(B|A_i) \cdot P(A_i)}{P(B)}$$ Bayes' Theorem

using data from our previous example:

$$P\left(\begin{matrix}\text{urn 1} & 0.2 \\ \text{urn 2} & 0.1\end{matrix} \;\middle|\; \text{3 white marbles}\right) = \frac{P(\text{observed}|\text{process parameter}) \cdot P(\text{process})}{P(\text{observed})}$$

process parameter — observed data

The denominator P(observed) or P(B) can be split-up using the sum rule to be the series of joint probabilities of $P(\text{data}|\text{process}_n) \cdot P(\text{process}_n)$, so P(B) in our example above can be written out as:

$P(B|A_1)P(A_1) + P(B|A_2)P(A_2) \ldots\ldots + P(B|A_n)P(A_n)$.

So, to answer this problem, first calculate the likelihoods, then use the principle of indifference, then use Bayes' theorem in its main form.

urn 1 likelihood = P(3 white marbles in a row | 20% white) = $0.2 \times 0.2 \times 0.2 = 0.008$

urn 2 likelihood = P(3 white marbles in a row | 10% white) = $0.1 \times 0.1 \times 0.1 = 0.001$

Apply principle of indifference: we have no basis for choosing between urns, before any data there is no given bias, we are neutral to the urn, so they have equal probability, where $P(A_1) = 0.5$ and $P(A_2) = 0.5$.

Bayes' Theorem III

Continuing from our previous problem to write out the equation fully:

$$P(A_1|B) = \frac{P(B|A_1)\cdot P(A_1)}{P(B|A_1)\cdot P(A_1) + P(B|A_2)\cdot P(A_2)} = \frac{0.008 \times 0.5}{[0.008 \times 0.5] + [0.001 \times 0.5]} = \frac{8}{9}$$

so $P(A_1|B) = \frac{8}{9}$ therefore $P(A_2|B) = \frac{1}{9}$ (where $1 - \frac{8}{9} = \frac{1}{9}$)

There was clearly a much greater probability that the marble draws we observed, came from urn 1, rather than urn 2.

↳ (A) from the assumed data the answer is $P(\text{urn1}) = \frac{8}{9}$ while $P(\text{urn2}) = \frac{1}{9}$.

Crucially, Bayes' theorem allows for updating probabilities based on new data. Imagine we add a 4th marble to the draw, which is also white. So, rather than using the principle of indifference, we take the generated probabilities as the 'new priors':

$P(\text{urn 1} \mid \text{3 whites in a row}) = 8/9$

$P(\text{urn 2} \mid \text{3 whites in a row}) = 1/9$

so now Bayes' theorem is stated: $P(\text{urn1} \mid \text{3 whites + 1 new white})$

$$= \frac{P(\text{white}|\text{urn1})\cdot P(\text{urn1})}{P(\text{white}|\text{urn1})\,P(\text{urn1}) + P(\text{white}|\text{urn2})\,P(\text{urn2})} = \frac{(0.2)(8/9)}{(0.2)(8/9) + (0.1)(1/9)}$$

$= P(A_1|B) = 94.12\%$, so $P(A_2|B) = 5.88\%$

↳ This new outcome event increases the probability of the situation we previously had coming from urn 1, where $8/9 = 88\%$ and $1/9 = 11\%$. But now given the addition of the 4th marble, the new probabilities are 94.12% and 5.88% respectively.

The result of Bayes' Theorem is known as the posterior probability, which denotes after new data are observed. The prior probability is what information we started with before any new data were observed, or indeed any data at all.

The standard forward probability part of the equation is known as the likelihood while the probability of the expression in the denominator is known as the marginal probability.

The Binomial Theorem

The binomial theorem is useful when there are 2 outcomes, ie. a success or non-success, hence only 2 outcomes are possible. The following equation is used, where n = number of independent trials (with replacement), while s = number of successes and p = probability of 1 success event occuring.

$$\binom{n}{s} \cdot p^s (1-p)^{n-s}$$

↑ probability of "s" successes in "n" trials, when the probability of 1 success is "p"

An example would be: 72 heads achieved in 100 coin tosses, where the coin is fair, so in this case $n=100$, $s=72$ and $p=0.5$.

$$\binom{100}{72} \cdot (0.5)^{72} (0.5)^{28} = 3.94 \times 10^{-6}$$

↑ which is an extremely small probability of achieving 72 heads within 100 tosses!

It is useful to have awareness of Bayes' Theorem and the Binomial Theorem to obtain information of the unknown process, with certainty, if you have relevant data on the outcomes. Both help to explore probability in an organised way.

Note, in its full form the Binomial Theorem is:

$$(a+b)^n = \sum_{s=0}^{n} \binom{n}{s} \cdot a^s b^{n-s}$$

where it is possible to expand a polynomial $(a+b)^n$ into its multiple terms.
ie. to expand $(x+y)^n$ we have the following:

$(x+y)^0 = 1$

$(x+y)^1 = x+y$

$(x+y)^2 = x^2 + 2xy + y^2$