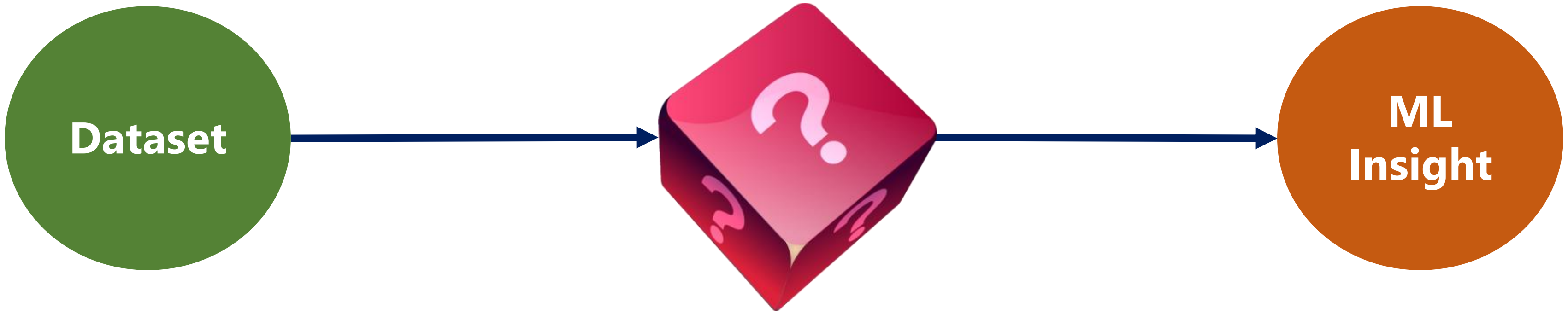
PROBABILITY FOR STATISTICS AND DATA SCIENCE

Introduction to Probability: Cheat Sheet

Probability Formula | Sample Space | Expected Values | Complements

Words of welcome



You are here because you want to comprehend the basics of probability before you can dive into the world of statistics and machine learning. Understanding the driving forces behind key statistical features is crucial to reaching your goal of mastering data science. This way you will be able to extract important insight when analysing data through supervised machine learning methods like regressions, but also fathom the outputs unsupervised or assisted ML give you.

Bayesian Inference is a key component heavily used in many fields of mathematics to succinctly express complicated statements. Through Bayesian Notation we can convey the relationships between elements, sets and events. Understanding these new concepts will aid you in interpreting the mathematical intuition behind sophisticated data analytics methods.

Distributions are the main way we lie to classify sets of data. If a dataset complies with certain characteristics, we can usually attribute the likelihood of its values to a specific distribution. Since many of these distributions have elegant relationships between certain outcomes and their probabilities of occurring, knowing key features of our data is extremely convenient and useful.

What is probability?

Probability is **the likelihood of an event occurring**. This event can be pretty much anything – getting heads, rolling a 4 or even bench pressing 225lbs. We measure probability with numeric values between 0 and 1, because we like to *compare* the relative likelihood of events. Observe the general probability formula.

$$P(X) = \frac{Preferred\ outcomes}{Sample\ Space}$$

Probability Formula:

- The Probability of event X occurring equals the *number* of preferred outcomes over the *number* of outcomes in the sample space.
- Preferred outcomes are the outcomes we want to occur or the outcomes we are interested in. We also call refer to such outcomes as "Favorable".
- Sample space refers to all possible outcomes that can occur. Its "size" indicates the amount of elements in it.

If two events are independent:

The probability of them occurring simultaneously equals the product of them occurring on their own.

P(A♥) = P(A) . P(♥)

Expected Values

Trial – Observing an event occur and recording the outcome.

Experiment – A collection of one or multiple trials.

Experimental Probability – The probability we assign an event, based on an experiment we conduct.

Expected value – the specific outcome we expect to occur when we run an experiment.

Example: Trial

Flipping a coin and recording the outcome.

Example: Experiment

Flipping a coin 20 times and recording the 20 individual outcomes.

In this instance, the **experimental probability** for getting heads would equal the number of heads we record over the course of the 20 outcomes, over 20 (the total number of trials).

The **expected value** can be numerical, Boolean, categorical or other, depending on the type of the event we are interested in. For instance, the expected value of the trial would be the more likely of the two outcomes, whereas the expected value of the experiment will be the number of time we expect to get either heads or tails after the 20 trials.

Expected value for **categorical** variables.

$$E(X) = n \times p$$

Expected value for **numeric** variables.

$$E(X) = \sum_{i=1}^{n} x_i \times p_i$$

Probability Frequency Distribution

What is a probability frequency distribution?:
A collection of the probabilities for each possible outcome of an event.

Why do we need frequency distributions?:
We need the probability frequency distribution to try and predict future events when the expected value is unattainable.

What is a frequency?:
Frequency is the number of times a given value or outcome appears in the sample space.

What is a frequency distribution table?:
The frequency distribution **table** is a table matching each distinct outcome in the sample space to its associated frequency.

How do we obtain the probability frequency distribution from the frequency distribution table?:
By dividing every frequency by the size of the sample space. (Think about the "favoured over all" formula.)

Sum	Frequency	Probability
2	1	1/36
3	2	1/18
4	3	1/12
5	4	1/9
6	5	5/36
7	6	1/6
8	5	5/36
9	4	1/9
10	3	1/12
11	2	1/18
12	1	1/36

Complements

The complement of an event is **everything** an event is **not.** We denote the complement of an event with an apostrophe.

$$A' = \text{Not } A$$

complement

opposite

original event

Characteristics of complements:

- Can never occur simultaneously.
- Add up to the sample space. (A + A′ = Sample space)
- Their probabilities add up to 1. (P(A) + P(A′) = 1)
- The complement of a complement is the original event. ((A′)′ = A)

Example:

- Assume event A represents drawing a spade, so P(A) = 0.25.
- Then, A′ represents **not** drawing a spade, so drawing a club, a diamond or a heart. P(A′) = 1 – P(A), so P(A′) = 0.75.

Permutations

Permutations represent the number of different possible ways we can **arrange** a number of elements.

$$P(n) = n \times (n-1) \times (n-2) \times \cdots \times 1$$

Permutations

Options for who we put first

Options for who we put second

Options for who we put last

Characteristics of Permutations:

- Arranging **all** elements within the sample space.
- No repetition.
- $P(n) = n \times (n-1) \times (n-2) \times \cdots \times 1 = n!$ (Called "n factorial")

Example:

- If we need to arrange 5 people, we would have P(5) = 120 ways of doing so.

Factorials

Factorials express the **product** of all integers from **1** to **n** and we denote them with the "!" symbol.

$$n! = n \times (n-1) \times (n-2) \times \cdots \times 1$$

Key Values:

- 0! = 1.
- If n<0, n! does not exist.

Rules for factorial multiplication. (For n>0 and n>k)

- $(n+k)! = n! \times (n+1) \times \cdots \times (n+k)$
- $(n-k)! = \frac{n!}{(n-k+1)\times\cdots\times(n-k+k)} = \frac{n!}{(n-k+1)\times\cdots\times n}$
- $\frac{n!}{k!} = \frac{k!\times(k+1)\times\cdots\times n}{k!} = (k+1) \times \cdots \times n$

Examples: n = 7, k = 4

- $(7+4)! = 11! = 7! \times 8 \times 9 \times 10 \times 11$
- $(7-4)! = 3! = \frac{7!}{4\times5\times6\times\times7}$
- $\frac{7!}{4!} = 5 \times 6 \times 7$

Variations

Variations represent the number of different possible ways we can **pick** and **arrange** a number of elements.

Variations **with** repetition

$$\bar{V}(n,p) = n^p$$

Number of different elements available

Number of elements we are arranging

Variations without repetition

$$V(n,p) = \frac{n!}{(n-p)!}$$

Number of different elements available

Number of elements we are arranging

Intuition behind the formula. (With Repetition)

- We have n-many options for the first element.
- We **still have n-many options** for the second element because repetition is allowed.
- We have n-many options for each of the p-many elements.
- $n \times n \times n \dots n = n^p$

Intuition behind the formula. (Without Repetition)

- We have n-many options for the first element.
- We **only have (n-1)-many options** for the second element because we cannot repeat the value for we chose to start with.
- We have less options left for each additional element.
- $n \times (n-1) \times (n-2) \dots (n-p+1) = \frac{n!}{(n-p)!}$

Combinations

Combinations represent the number of different possible ways we can **pick** a number of elements.

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

Combinations

Total number of elements in the sample space

Number of elements we need to select

Binomial coefficient (n-choose-k)

Characteristics of Combinations:

- Takes into account double-counting. (Selecting Johny, Kate and Marie is the same as selecting Marie, Kate and Johny)
- All the different permutations of a single combination are different variations.
- $C_p^n = \binom{n}{p} = \frac{V_p^n}{P} = \frac{n!/(n-p)!}{p!} = \frac{n!}{(n-p)!p!}$
- Combinations are symmetric, so $C_p^n = C_{n-p}^n$, since selecting p elements is the same as omitting n-p elements.

Combinations with separate sample spaces

Combinations represent the number of different possible ways we can **pick** a number of elements.

$$C = n_1 \times n_2 \times \cdots \times n_p$$

Combinations

Size of the first sample space.

Size of the second sample space.

Size of the last sample space.

Characteristics of Combinations with separate sample spaces:

- The option we choose for any element does not affect the number of options for the other elements.
- The order in which we pick the individual elements is arbitrary.
- We need to know the size of the sample space for each individual element. $(n_1, n_2 \ldots n_p)$

Winning the Lottery

To win the lottery, you need to satisfy two distinct independent events:

- Correctly guess the "Powerball" number. (From 1 to 26)
- Correctly guess the 5 regular numbers. (From 1 to 69)

Total number of Combinations

$$C = \underbrace{\frac{69!}{64!\,5!}}_{C_{5\ numbers}} \times \underbrace{26}_{C_{Powerball\ number}}$$

Intuition behind the formula:

- We consider the two distinct events as a combination of two elements with different sample sizes.
  - One event has a sample size of 26, the other has a sample size of C_5^{69}.
- Using the "favoured over all" formula, we find the probability of any single ticket winning equals $1/(\frac{69!}{64!5!} \times 26)$.

Combinations With Repetition

Combinations represent the number of different possible ways we can **pick** a number of elements. In special cases we can have repetition in combinations and for those we use a different formula.

$$\bar{C}(n, p) = \frac{(n + p - 1)!}{(n - 1)!\, p!}$$

Combinations with repetition

Total number of elements in the sample space

Number of elements we need to select

Now that you know what the formula looks like, we are going to walk you through the process of deriving this formula from the Combinations *without* repetition formula. This way you will be able to fully understand the intuition behind and not have to bother memorizing it.

Applications of Combinations with Repetition

To understand how combinations with repetition work you need to understand the instances where they occur.

We use combinations with repetition when the events we are dealing with, have sufficient quantity of each of the distinct values in the sample space.

One such example is the toppings on a pizza.

We can order extra of any given topping, so **repetition is allowed**. However, we **do not care about the order** in which the toppings are put on top of the pizza, so we cannot use variations.

Similar examples include picking the ice-cream flavours for a Sundae melt or the players for a Fantasy Football Team.

Pizza Example

To get a better grasp of the number of combinations we have, let us explore a specific example.

You are ordering a 3-topping pizza from your local pizza place but they only have 6 topping left because its late.

The toppings are as follows:
cheddar cheese, onions, green peppers, mushrooms, pepperoni and bacon.

Your pizza can have 3 different toppings, or you can repeat a topping up to 3 times.

You can either order a pizza with 3 different toppings, a pizza with 3 identical toppings or a pizza with 2 different toppings but having a double dose of one of them.

Methodology

The methodology we use for such combinations is rather abstract. We like to represent each type of pizza with a special sequence of 0s and 1s. To do so, we first need to select a specific order for the available ingredients.

We can reuse the order we wrote down earlier:
cheddar **c**heese, **o**nions, **g**reen peppers, **m**ushrooms, **p**epperoni and **b**acon.

For convenience we can refer to each ingredient by the associated letter we have highlighted (e.g "c" means cheese, and "o" means onions).

To construct the sequence for each unique type of pizza we follow 2 rules as we go through the ingredients in the order we wrote down earlier.

1. If we want no more from a certain topping, we write a **0 and move to the next topping**.
2. If we want to include a certain topping, we write a **1 and stay on the same topping**.
   - Not going to the next topping allows us to indicate if we want extra by adding another 1, before we move forward. Say, if we want our pizza to have extra cheese, the sequence would begin with "1, 1".
   - Also, we always apply rule 1 before moving on to another topping, so the sequence will actually start with "1, 1, 0".

Pizzas and Sequences

If we need to write a "0" after each topping, then every sequence would consist of 6 zeroes and 3 ones.

Let's look at some pizzas and the sequences they are expressed with.

A pizza with **cheese** and **extra peperoni** is represented by the sequence 1,0,0,0,0,1,1,0,0.

A vegan variety pizza with **onions, green peppers and mushrooms** would be represented by the sequence 0,1,0,1,0,1,0,0,0.

Now, what kind of pizza would the sequence **0,0,1,0,0,0,1,1,0** represent?

We can put the sequence into the table and see that it represents a pizza with green peppers and extra bacon.

C	O	G	M	P	B
1,0	0	0	0	1,1,0	0
0	1,0	1,0	1,0	0	0
0	0	1,0	0	0	1,1,0

Always Ending in 0

Notice how all the sequences we have examined end on a 0:

- 1,0,0,0,0,1,1,0,**0**
- 0,1,0,1,0,1,0,0,**0**
- 0,0,1,0,0,0,1,1,**0**

This is no coincidence, since according to rule 1 of our algorithm, we need to add a "0" at the end of the sequence, regardless of whether we wanted bacon or not.

That means that only the first 8 elements of the sequence can take different values.

Each pizza is characterized by the positions of the 3 "1"s in the sequence. Since only 8 of the positions in the sequence can take a value of "1", then the number of different pizzas would be the combination of any 3 of the 8 positions.

C	O	G	M	P	B
1,0	0	0	0	1,1,0	0
0	1,0	1,0	1,0	0	0
0	0	1,0	0	0	1,1,0

Positions

As stated before, we have 3 “1s” and 8 different positions. Therefore, the number of pizzas we can get would be the number of combinations of picking 3 elements out of a set of 8. This means we can transform combinations **with** repetition to combinations **without** repetition.

$$\bar{C}(6,3) = \mathrm{C}(8,3)$$

Let’s observe the values 3 and 8 for a moment and try to generalize the formula. “3” represents the amount of toppings we need to pick, so it is still equivalent to p.

$$\bar{C}(n,p) = \mathrm{C}(\mathrm{n}+\mathrm{p}-1,\mathrm{p})$$

“8” represents the number of positions we have available for the ones. We had 3 + 6, or 9 positions in total, but we knew the last one could not contain a “1”. Thus, we had “n + p - 1” many positions that could contain a 1.

The Final Step

Now that we know the relationship between the number of combinations **with** and **without** repetition, we can plug in "n+p-1" into the combinations without repetitio formula to get:

$$\bar{C}(n,p) = \mathrm{C}(\mathrm{n}+\mathrm{p}-1,\mathrm{p}) = \frac{(\mathrm{n}+\mathrm{p}-1)!}{((\mathrm{n}+\mathrm{p}-1)-\mathrm{p})!\,\mathrm{p}!} = \frac{(\mathrm{n}+\mathrm{p}-1)!}{(\mathrm{n}-1)!\,\mathrm{p}!}$$

This is the exact same formula we showed you at the beginning.
Before we you continue to the next lecture, let's make a quick recap of the algorithm and the formula.

1. We started by ordering the possible values and expressing every combinations as a sequence.
2. We examined that only certain elements of the sequence may differ.
3. We concluded that every unique sequence can be expressed as a combination of the positions of the "1" values.
4. We discovered a relationship between the formulas for combinations with and without repetition.
5. We used said relationship to create a general formula for combinations with repetition.

Symmetry of Combinations

Let's see the algebraic proof of the notion that selecting p-many elements out of a set of n is the same as omitting n-p many elements.

For starters, recall the combination formula:

$$\boldsymbol{C(n,p) = \frac{n!}{(n-p)!\,p!}}$$

If we plug in $n-p$ for p, we get the following:

$$\boldsymbol{C(n,n-p) = \frac{n!}{(n-(n-p))!\,(n-p)!} = \frac{n!}{(n-n+p))!\,(n-p)!} = \frac{n!}{p!\,(n-p)!} = \frac{n!}{(n-p)!\,p!} = C(n,p)}$$

Therefore, we can conclude that $\boldsymbol{C(n,p) = C(n,n-p)}$.

Bayesian Notation

A **set** is a collection of elements, which hold certain values. Additionally, every event has a set of outcomes that satisfy it.

The **null-set** (or **empty set**), denoted ∅, is an set which contain no values.

$$x \in A$$

Element (lower-case)

Set (upper-case)

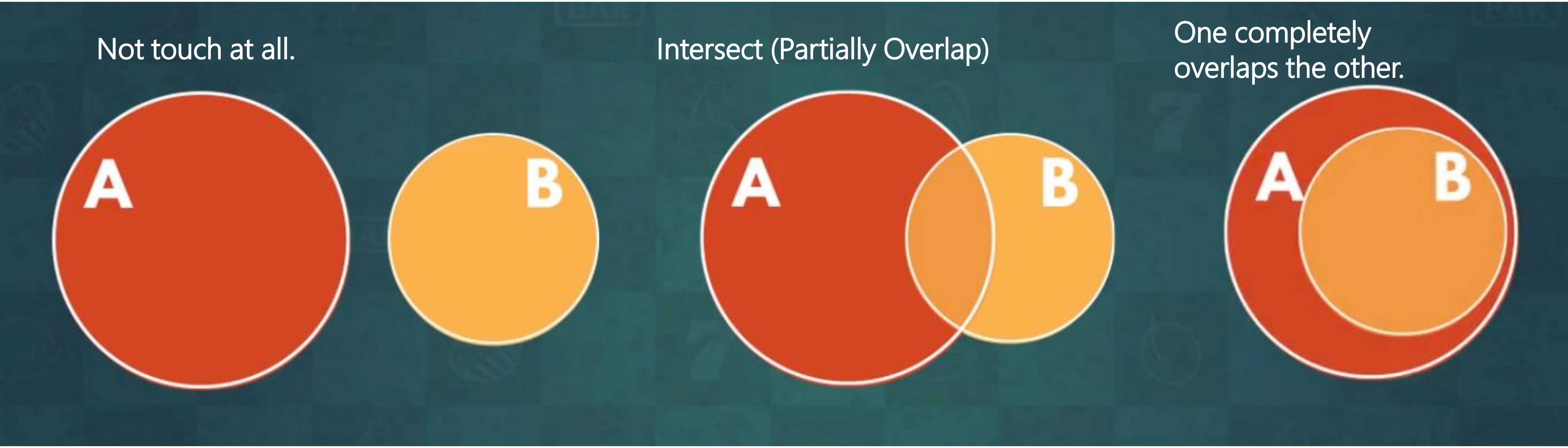
Notation:	Interpretation:	Example:
$x \in A$	"Element x is a part of set A."	$2 \in All\ even\ numbers$
$A \ni x$	"Set A contains element x."	$All\ even\ numbers \ni 2$
$x \notin A$	"Element x is NOT a part of set A."	$1 \notin All\ even\ numbers$
$\forall x$:	"For all/any x such that…"	$\forall x: x \in All\ even\ numbers$
$A \subseteq B$	"A is a subset of B"	$Even\ numbers \subseteq Integers$

Remember! Every set has at least 2 subsets.

- $A \subseteq A$
- $\emptyset \subseteq A$

Multiple Events

The sets of outcomes that satisfy two events A and B can interact in one of the following **3** ways.



Examples:

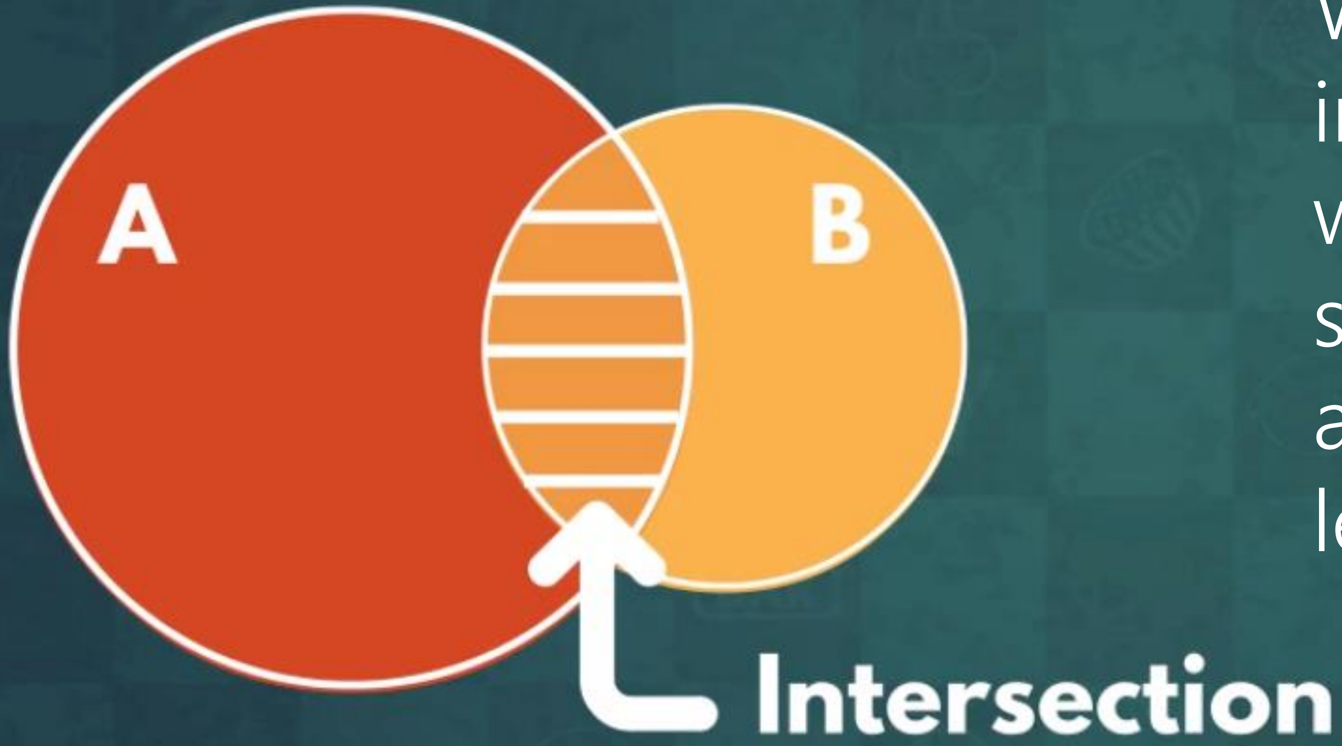
A -> Diamonds
B -> Hearts

Diamonds
Queens

Red Cards
Diamond

Intersection

The **intersection** of two or more events expresses the set of outcomes that satisfy all the events simultaneously. Graphically, this is the area where the sets intersect.

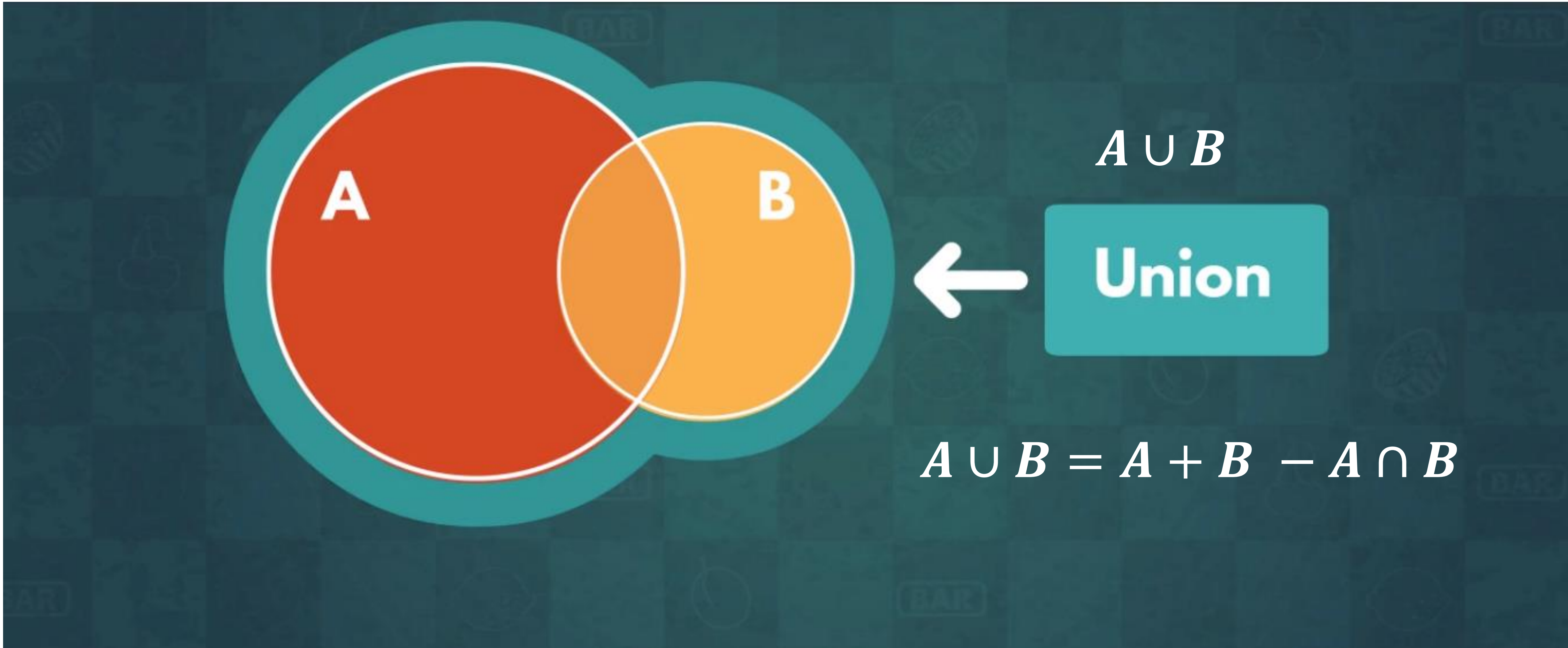


We denote the intersection of two sets with the “intersect” sign, which resembles an upside-down capital letter U:

$$\boldsymbol{A} \cap \boldsymbol{B}$$

Union

The **union** of two or more events expresses the set of outcomes that satisfy at least one of the events. Graphically, this is the area that includes both sets.

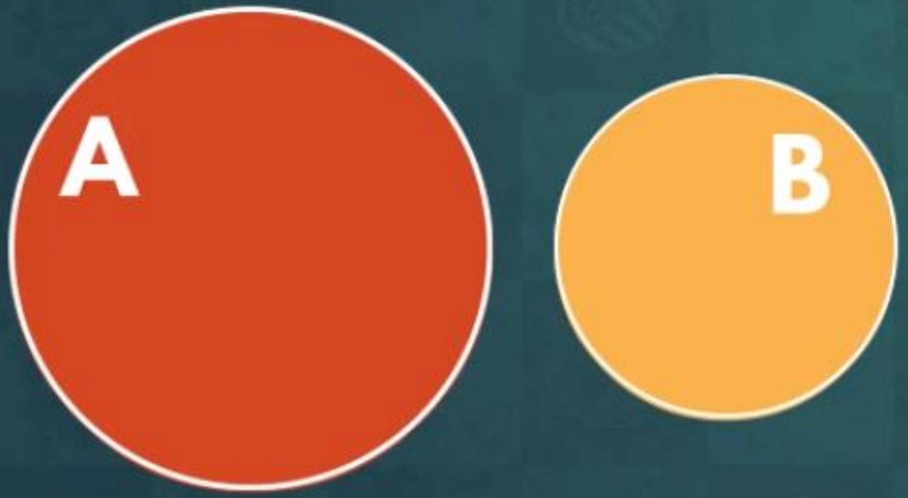


$$A \cup B$$

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

Mutually Exclusive Sets

Sets with no overlapping elements are called **mutually exclusive**. Graphically, their circles never touch.



If $A \cap B = \emptyset$, then the two sets are mutually exclusive.

Remember:
All complements are mutually exclusive, but not all mutually exclusive sets are complements.

Example:
Dogs and Cats are mutually exclusive sets, since no species is simultaneously a feline and a canine, but the two are not complements, since there exist other types of animals as well.

Independent and Dependent Events

If the likelihood of event A occurring (P(A)) is affected event B occurring, then we say that A and B are **dependent** events.

Alternatively, if it isn't – the two events are **independent**.

We express the probability of event A occurring, given event B has occurred the following way $\boldsymbol{P(A|B)}$.
We call this the conditional probability.

Independent:

- All the probabilities we have examined so far.
- The outcome of A does not depend on the outcome of B.
- $P(A|B) = P(A)$

Example

- A -> Hearts
- B -> Jacks

Dependent

- New concept.
- The outcome of A depends on the outcome of B.
- $P(A|B) \neq P(A)$

Example

- A -> Hearts
- B -> Red

Conditional Probability

For any two events A and B, such that the likelihood of B occurring is greater than 0 ($P(B) > 0$), the conditional probability formula states the following.

Probability of A, given B has occurred

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

Probability of the intersection.

Probability of event B

Intuition behind the formula:

- Only interested in the outcomes where B is satisfied.
- Only the elements in the intersection would satisfy A as well.
- Parallel to the "favoured over all" formula:
  - Intersection = "preferred outcomes"
  - B = "sample space"

Remember:

- Unlike the union or the intersection, changing the order of A and B in the conditional probability alters its meaning.
- $P(A|B)$ is not the same as $P(B|A)$, even if $P(A|B) = P(B|A)$ numerically.
- The two conditional probabilities possess **different meanings** even if they have equal values.

Law of total probability

The **law of total probability** dictates that for any set A, which is a union of many mutually exclusive sets $B_1, B_2, \dots, B_n$, its probability equals the following sum.

$$P(A) = \underbrace{P(A|B_1)}_{\substack{\text{Conditional Probability of A,} \\ \text{given } B_1 \text{ has occurred.}}} \times \underbrace{P(B_1)}_{\substack{\text{Probability of } B_1 \\ \text{occurring.}}} + \underbrace{P(A|B_2)}_{\substack{\text{Conditional Probability of A,} \\ \text{given } B_2 \text{ has occurred.}}} \times \underbrace{P(B_2)}_{\substack{\text{Probability of } B_2 \\ \text{occurring.}}} + \cdots + P(A|B_n) \times P(B_n)$$

Intuition behind the formula:

- Since P(A) is the union of mutually exclusive sets, so it equals the **sum of the individual sets**.
- The **intersection** of a union and one of its subsets is the entire subset.
- We can rewrite the conditional probability formula $\boldsymbol{P(A|B) = \frac{P(A \cap B)}{P(B)}}$ to get $\boldsymbol{P(A \cap B) = P(A|B) \times P(B)}$.
- Another way to express the law of total probability is:
  - $P(A) = P(A \cap B_1) + P(A \cap B_2) + \cdots + P(A \cap B_n)$

Additive Law

The additive law calculates the probability of the union based on the probability of the individual sets it accounts for.

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

Probability of the union

Probability of the intersection

Intuition behind the formula

- Recall the formula for finding the size of the Union using the size of the Intersection:
  - $A \cup B = A + B - A \cap B$
- The probability of each one is simply its size over the size of the sample space.
- This holds true for any events A and B.

The Multiplication Rule

The multiplication rule calculates the probability of the intersection based on the conditional probability.

$$P(A \cap B) = P(A|B) \times P(B)$$

Probability of the Intersection

Conditional Probability

Probability of event B

Intuition behind the formula

- We can multiply both sides of the conditional probability formula $P(A|B) = \frac{P(A \cap B)}{P(B)}$ by $P(B)$ to get $P(A \cap B) = P(A|B) \times P(B)$.
- If event B occurs in 40% of the time ($P(B) = 0.4$) and event A occurs in 50% of the time B occurs ($P(A|B) = 0.5$), then they would simultaneously occur in 20% of the time ($\mathrm{P}(\mathrm{A}|B) \times P(B) = 0.5 \times 0.4 = 0.2$).

Bayes' Law

Bayes' Law helps us understand the relationship between two events by computing the different conditional probabilities.

We also call it Bayes' Rule or Bayes' Theorem.

Conditional probability of B, given A.

Conditional probability of A, given B.

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

Intuition behind the formula

- According to the multiplication rule $P(A \cap B) = P(A|B) \times P(B)$, so $P(B \cap A) = P(B|A) \times P(A)$.
- Since $P(A \cap B) = P(B \cap A)$, we plug in $P(B|A) \times P(A)$ for $P(A \cap B)$ in the conditional probability formula $P(A|B) = \frac{P(A \cap B)}{P(B)}$
- Bayes' Law is often used in medical or business analysis to determine which of two symptoms affects the other one more.

An overview of Distributions

A distribution shows the possible values a random variable can take and how frequently they occur.

Important Notation for Distributions:

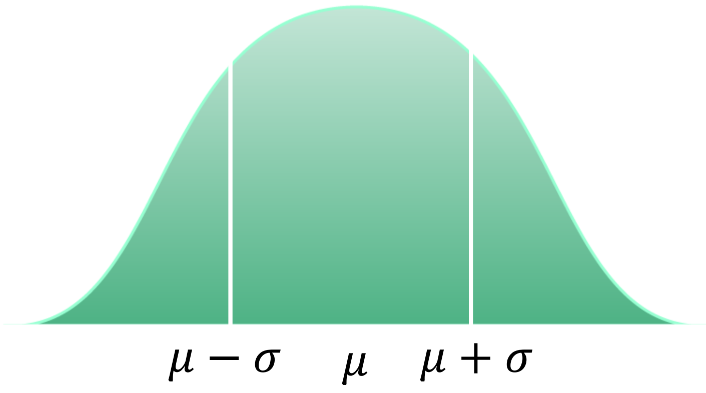
Y actual outcome

y one of the possible outcomes

$P(Y = y)$ is equivalent to $p(y)$.

We call a function that assigns a probability to each distinct outcome in the sample space, a **probability function.**

	Population	Sample
Mean	μ	$\bar{x}$
Variance	σ^2	s^2
Standard Deviation	σ	s



Types of Distributions

Certain distributions share characteristics, so we separate them into **types**. The well-defined types of distributions we often deal with have elegant statistics. We distinguish between two big types of distributions based on the type of the possible values for the variable – discrete and continuous.

Discrete

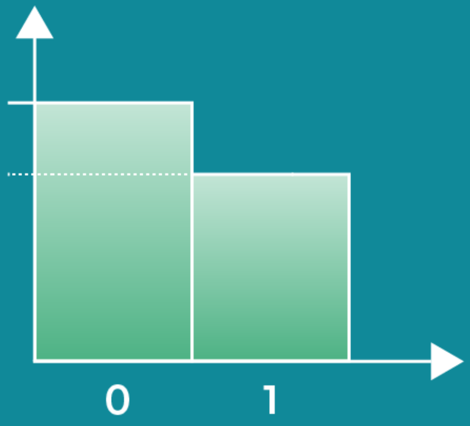
- Have a finite number of outcomes.
- Use formulas we already talked about.
- Can add up individual values to determine probability of an interval.
- Can be expressed with a table, graph or a piece-wise function.
- Expected Values might be unattainable.
- Graph consists of bars lined up one after the other.

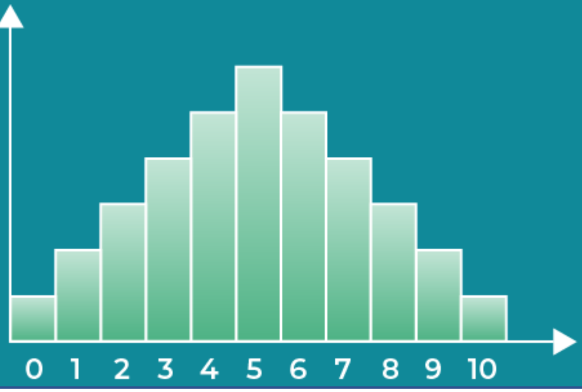
Continuous

- Have infinitely many consecutive possible values.
- Use new formulas for attaining the probability of specific values and intervals.
- Cannot add up the individual values that make up an interval because there are **infinitely many** of them.
- Can be expressed with a graph or a continuous function.
- Graph consists of a smooth curve.

Discrete Distributions

Discrete Distributions have finitely many different possible outcomes. They possess several key characteristics which separate them from continuous ones.





Key characteristics of discrete distribution

- Have a finite number of outcomes.
- Use formulas we already talked about.
- Can add up individual values to determine probability of an interval.
- Can be expressed with a table, graph or a piece-wise function.
- Expected Values might be unattainable.
- Graph consists of bars lined up one after the other.
- $\boldsymbol{P(Y \leq y) = P(Y < y + 1)}$

Examples of Discrete Distributions:

- Discrete Uniform Distribution
- Bernoulli Distribution
- Binomial Distribution
- Poisson Distribution

Uniform Distribution

A distribution where all the outcomes are equally likely is called a **Uniform Distribution**.

Notation:

- $Y \sim U(a, b)$
- * alternatively, if the values are categorical, we simply indicate the number of categories, like so: $Y \sim U(a)$

Key characteristics

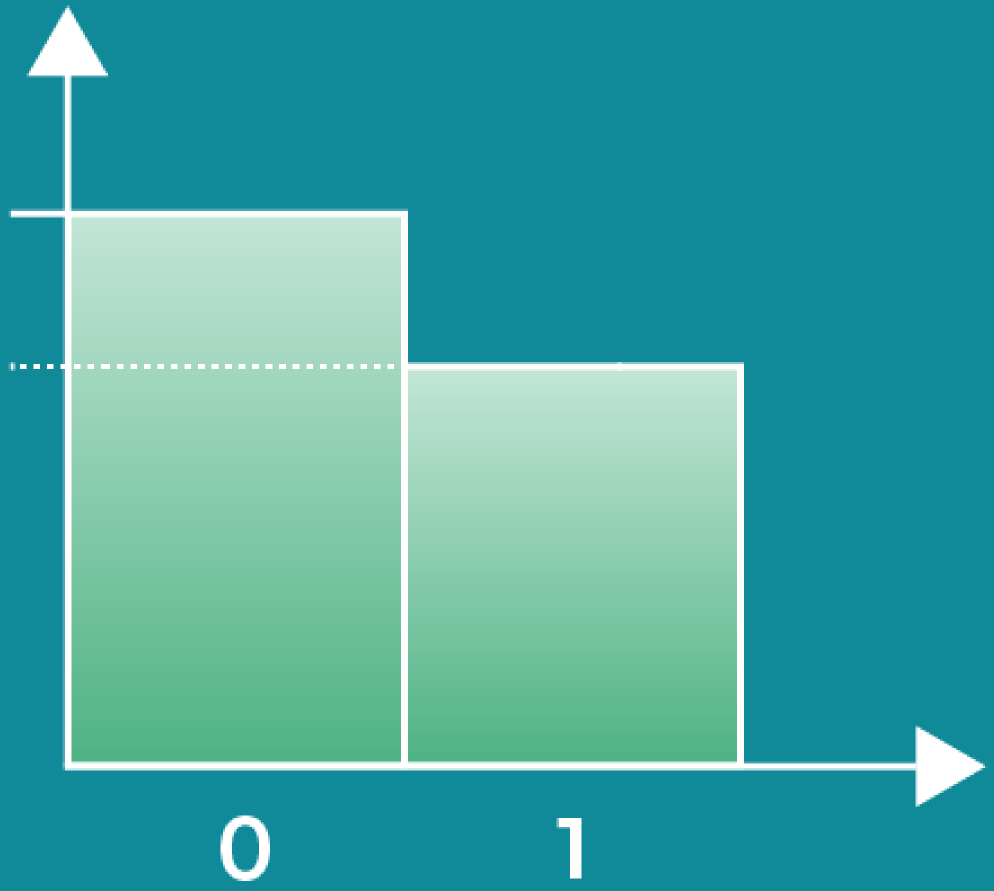
- All outcomes are equally likely.
- All the bars on the graph are equally tall.
- The expected value and variance have no predictive power.

Example and uses:

- Outcomes of rolling a single die.
- Often used in shuffling algorithms due to its fairness.

Bernoulli Distribution

A distribution consisting of a single trial and only two possible outcomes – success or failure is called a **Bernoulli Distribution.**



Notation:

- $Y \sim Bern(p)$

Key characteristics

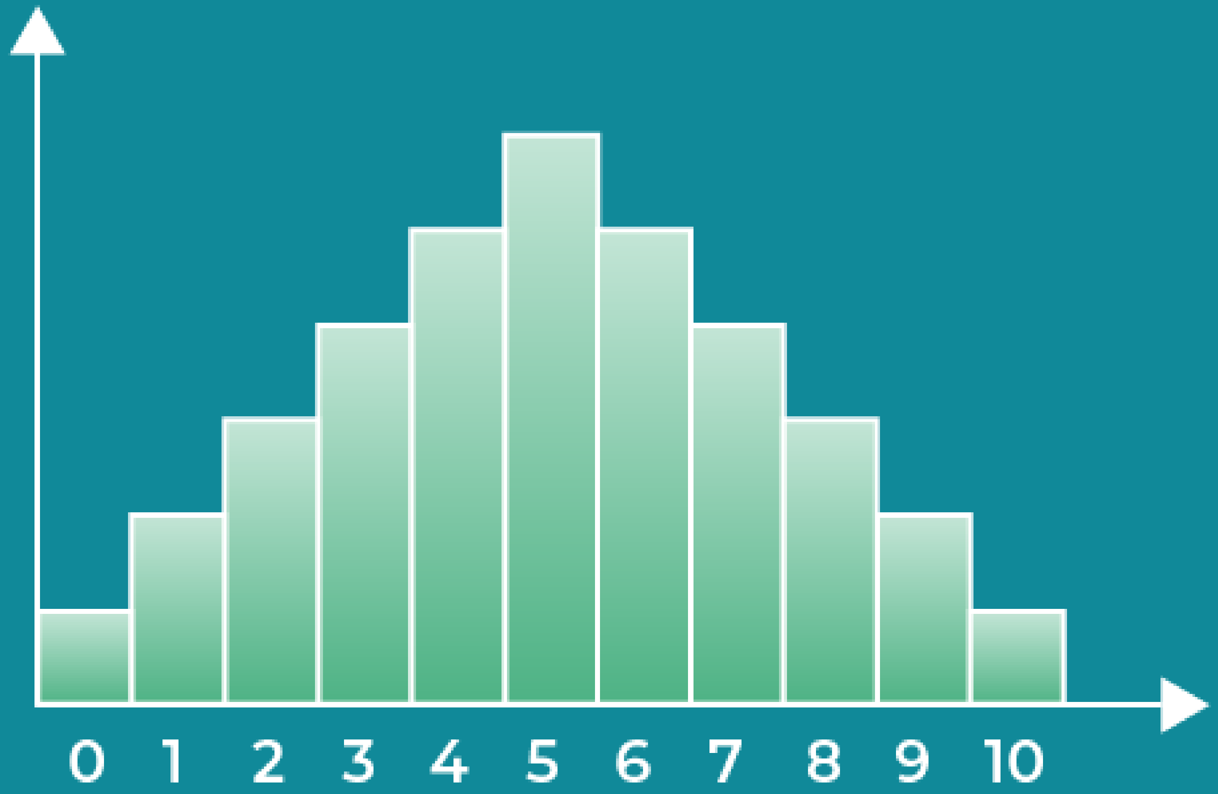
- One trial.
- Two possible outcomes.
- $E(Y) = p$
- $Var(Y) = p \times (1 - p)$

Example and uses:

- Guessing a single True/False question.
- Often used in when trying to determine what we expect to get out a single trial of an experiment.

Binomial Distribution

A sequence of identical Bernoulli events is called Binomial and follows a **Binomial Distribution**.



Notation:

- $Y \sim B(n, p)$

Key characteristics

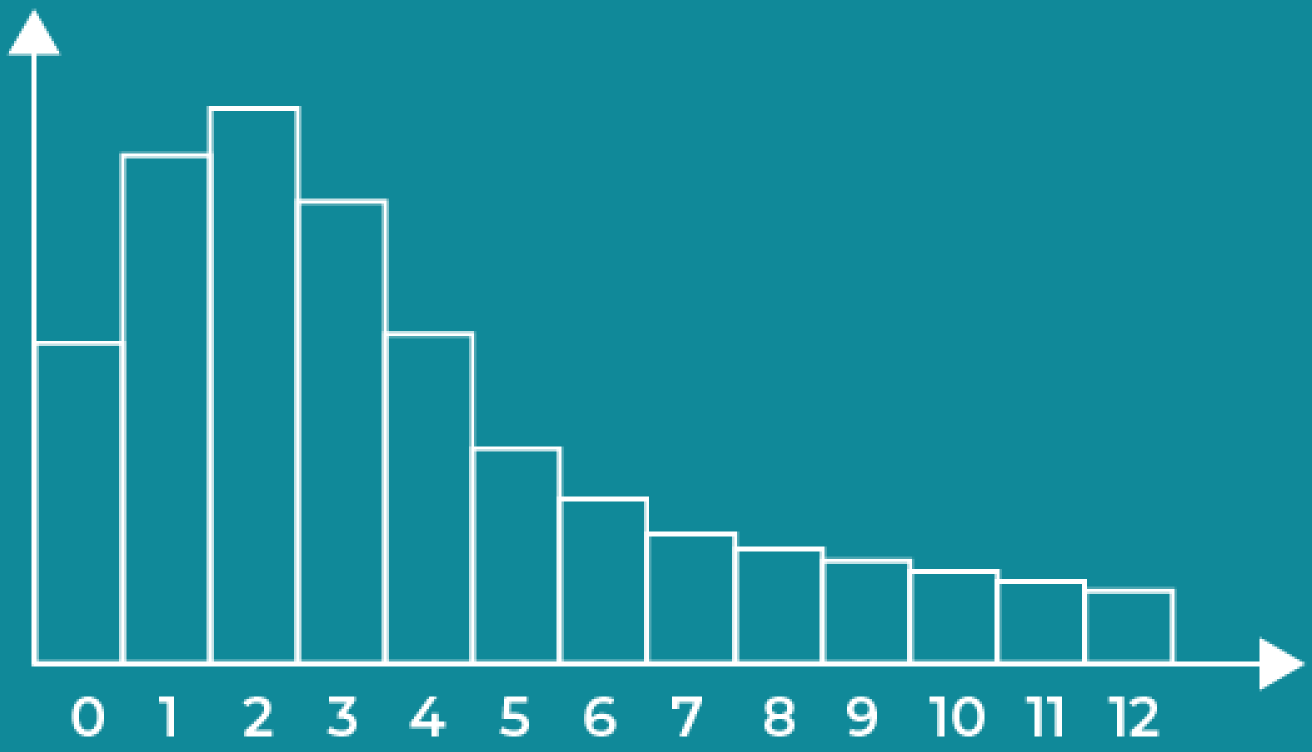
- Measures the frequency of occurrence of one of the possible outcomes over the n trials.
- $P(Y = y) = C(y, n) \times p^y \times (1 - p)^{n-y}$
- $E(Y) = n \times p$
- $Var(Y) = n \times p \times (1 - p)$

Example and uses:

- Determining how many times we expect to get a heads if we flip a coin 10 times.
- Often used when trying to predict how likely an event is to occur over a series of trials.

Poisson Distribution

When we want to know the likelihood of a certain event occurring over a given interval of time or distance, we use a **Poisson Distribution**.



Notation:

- $Y \sim Po(\lambda)$

Key characteristics

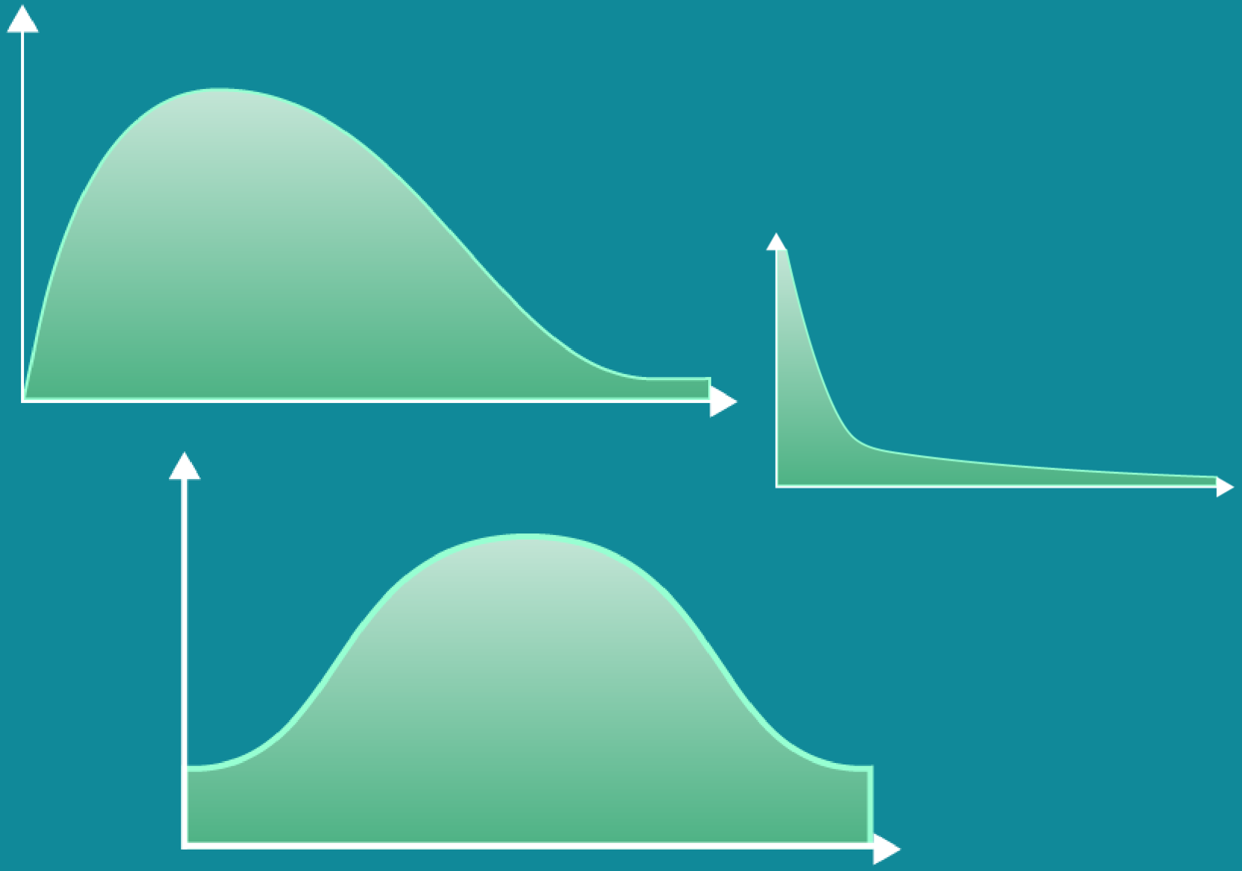
- Measures the frequency over an interval of time or distance. (Only non-negative values.)
- $P(Y = y) = \frac{\lambda^y e^{-\lambda}}{y!}$
- $E(Y) = \lambda$
- $Var(Y) = \lambda$

Example and uses:

- Used to determine how likely a specific outcome is, knowing how often the event **usually** occurs.
- Often incorporated in marketing analysis to determine whether above average visits are out of the ordinary or not.

Continuous Distributions

If the possible values a random variable can take are a sequence of infinitely many consecutive values, we are dealing with a **continuous distribution**.

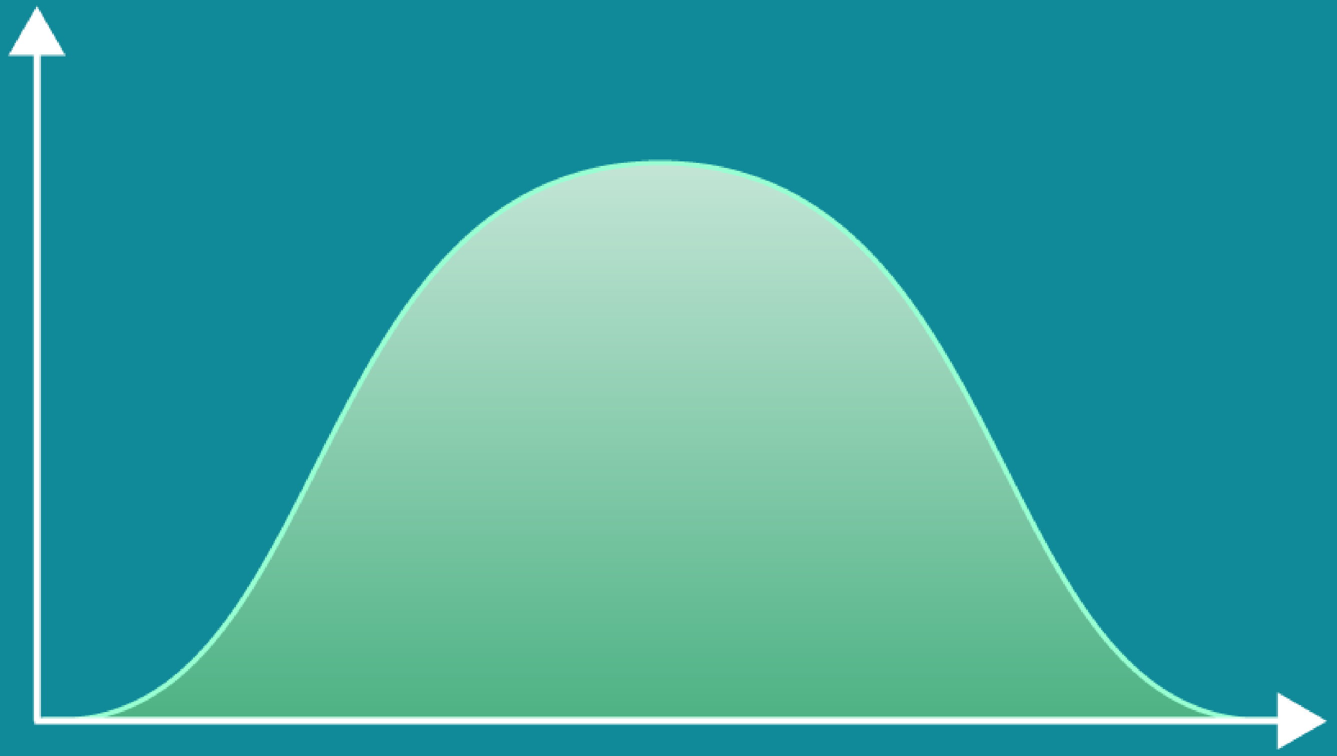


Key characteristics

- Have infinitely many consecutive possible values.
- Cannot add up the individual values that make up an interval because there are **infinitely many** of them.
- Can be expressed with a graph or a continuous function. Cannot use a table, be
- Graph consists of a smooth curve.
- To calculate the likelihood of an interval, we need integrals.
- They have important CDFs.
- $\boldsymbol{P(Y = y)} = 0$ for any individual value y.
- $\boldsymbol{P(Y < y) = P(Y \leq y)}$

Normal Distribution

A **Normal Distribution** represents a distribution that most natural events follow.



Notation:

- $\boldsymbol{Y} \sim \boldsymbol{N}(\mu, \sigma^2)$

Key characteristics

- Its graph is bell-shaped curve, symmetric and has thin tails.
- $\boldsymbol{E}(\boldsymbol{Y}) = \mu$
- $\boldsymbol{Var}(\boldsymbol{Y}) = \sigma^2$
- 68% of all its values should fall in the interval:
  - $(\mu - \boldsymbol{\sigma},\ \boldsymbol{\mu} + \boldsymbol{\sigma})$

Example and uses:

- Often observed in the size of animals in the wilderness.
- Could be standardized to use the Z-table.

Standardizing a Normal Distribution

To standardize any normal distribution we need to transform it so that the mean is 0 and the variance and standard deviation are 1.

Ensures mean is 0.

Using a transformation to create a new random variable z.

$$z = \frac{y - \mu}{\sigma}$$

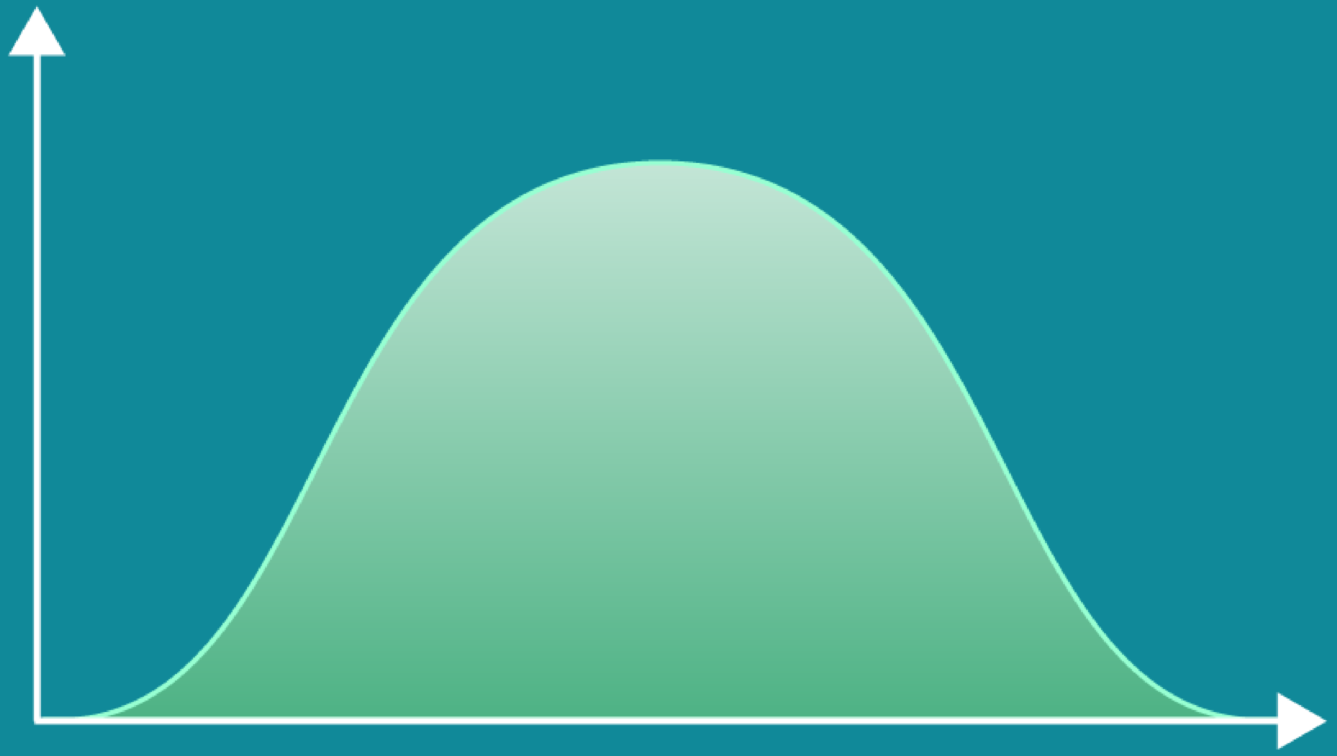
Ensures standard deviation is 1.

Importance of the Standard Normal Distribution.

- The new variable z, represents how many standard deviations away from the mean, each corresponding value is.
- We can transform any Normal Distribution into a Standard Normal Distribution using the transformation shown above.
- Convenient to use because of a table of known values for its CDF, called the Z-score table, or simply the Z-table.

Students’ T Distribution

A **Normal Distribution** represents a small sample size approximation of a Normal Distribution.



Notation:

- $Y \sim t\ (k)$

Key characteristics

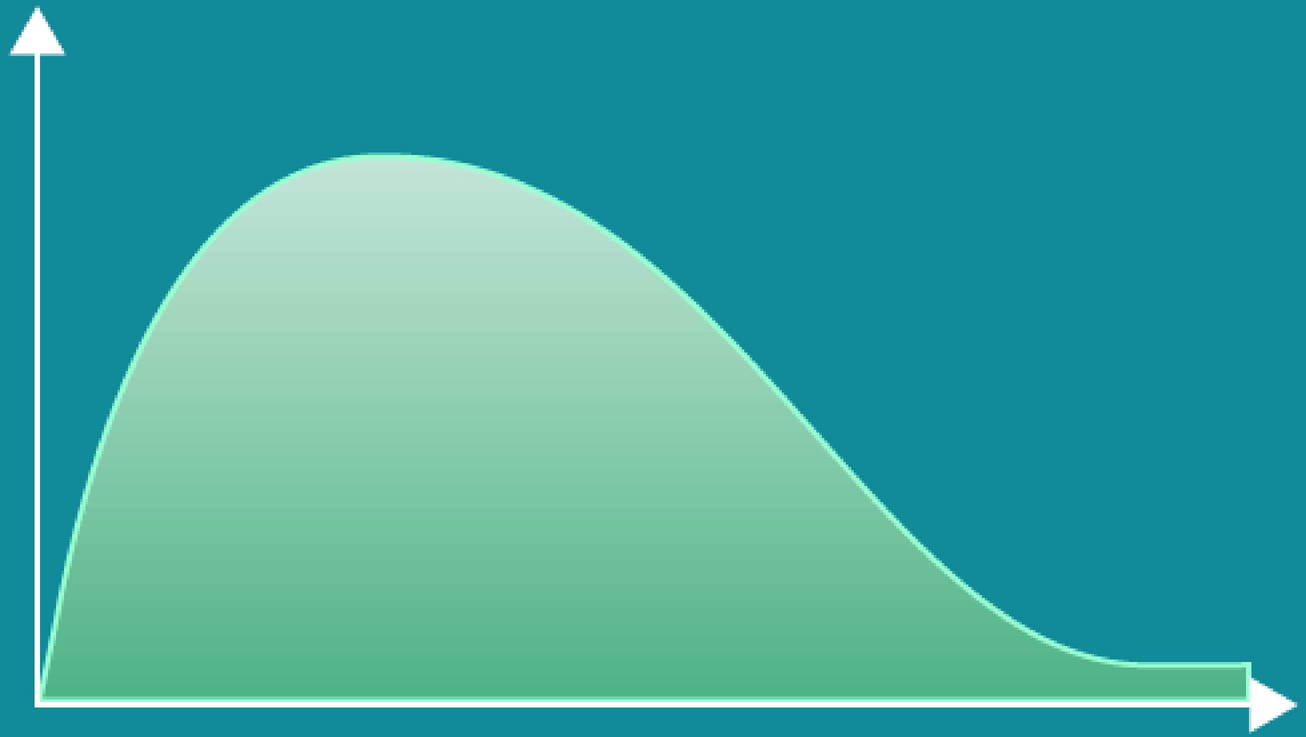
- A small sample size approximation of a Normal Distribution.
- Its graph is bell-shaped curve, symmetric, but has **fat** tails.
- Accounts for extreme values better than the Normal Distribution.
- If k>2: $E(Y) = \mu$ and $Var(Y) = s^2 \times \frac{k}{k-2}$

Example and uses:

- Often used in analysis when examining a small sample of data that usually follows a Normal Distribution.

Chi-Squared Distribution

A **Chi-Squared** distribution is often used.



Notation:

- $Y \sim \chi^2(k)$

Key characteristics

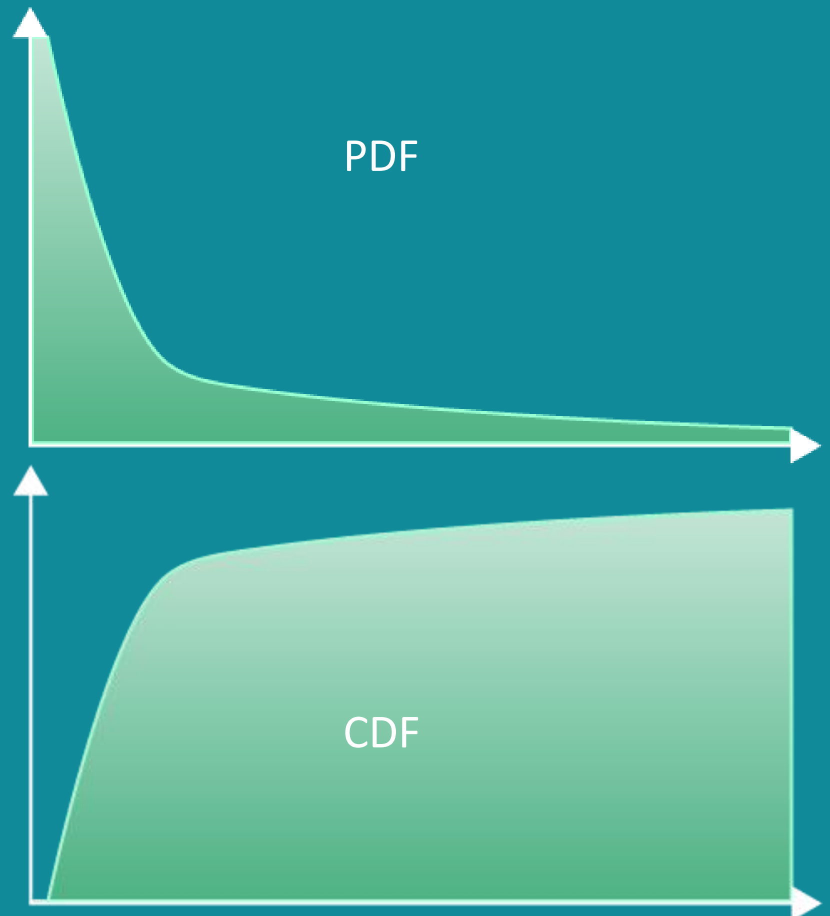
- Its graph is asymmetric and skewed to the right.
- $E(Y) = k$
- $Var(Y) = 2k$
- The Chi-Squared distribution is the square of the t-distribution.

Example and uses:

- Often used to test goodness of fit.
- Contains a table of known values for its CDF called the χ^2-table. The only difference is the table shows what part of the table

Exponential Distribution

The **Exponential Distribution** is usually observed in events which significantly change early on.



Notation:

- $Y \sim Exp\ (\lambda)$

Key characteristics

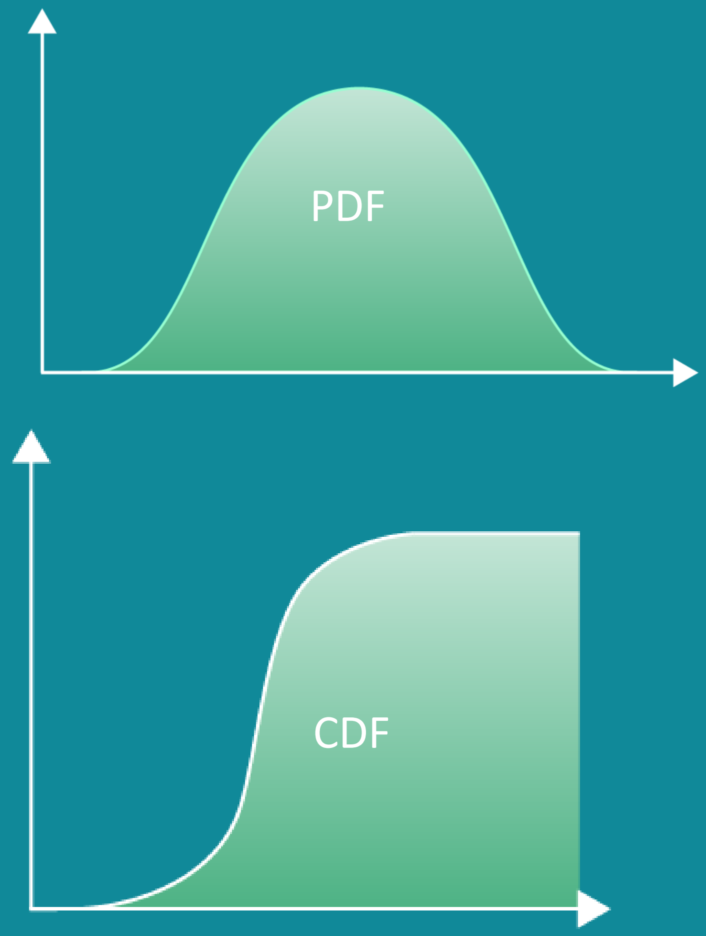
- Both the PDF and the CDF plateau after a certain point.
- $E(Y) = \frac{1}{\lambda}$
- $Var(Y) = \frac{1}{\lambda^2}$
- We often use the natural logarithm to transform the values of such distributions since we do not have a table of known values like the Normal or Chi-Squared.

Example and uses:

- Often used with dynamically changing variables, like online website traffic or radioactive decay.

Logistic Distribution

The **Continuous Logistic Distribution** is observed when trying to determine how continuous variable inputs can affect the probability of a binary outcome.



Notation:

- $\boldsymbol{Y \sim Logistic\ (\mu, s)}$

Key characteristics.

- $\boldsymbol{E(Y) = \mu}$
- $\boldsymbol{Var(Y) = \frac{s^2 \times \pi^2}{3}}$
- The CDF picks up when we reach values near the mean.
- The smaller the scale parameter, the quicker it reaches values close to 1.

Example and uses:

- Often used in sports to anticipate how a player's or team's performance can determine the outcome of the match.

Poisson Distribution: Expected Value and Variance:

Assume we have a random variable Y, such that $Y \sim Poisson(\lambda)$, then we can find its expected value and variance the following way. Recall that the expected value for any discrete random variable is a sum of all possible values multiplied by their likelihood of occurring P(y). Thus,

$$E(Y) = \sum_{y=0}^{\infty} y \frac{e^{-\lambda}\lambda^{y}}{y!}$$

Now, when y=0, the entire product is 0, so we can start the sum from y=1 instead. Additionally, we can divide the numerator and denominator by "y", since "y" will be non-zero in every case.

$$= \sum_{y=1}^{\infty} y \frac{e^{-\lambda}\lambda^{y}}{y!} = \sum_{y=1}^{\infty} \frac{e^{-\lambda}\lambda^{y}}{(y-1)!}$$

Since lambda is a constant number, we can take out $\lambda e^{-\lambda}$, in front of the sum.

$$= \lambda e^{-\lambda} \sum_{y=1}^{\infty} \frac{\lambda^{y-1}}{(y-1)!}$$

Since we have "y-1" in both the numerator and denominator and the sum starts from 1, this is equivalent to starting the sum from 0 and using y instead.

$$= \lambda e^{-\lambda} \sum_{y=0}^{\infty} \frac{\lambda^{y}}{y!}$$

Calculus dictates that for any constant "c", $\sum_{x=0}^{\infty} \frac{c^x}{x!} = e^c$. We use this to simply the expression to:

$$= \lambda e^{-\lambda} e^{\lambda}$$

Lastly, since any value to the negative power is the same as 1 divided by that same value, then $e^{-\lambda}e^{\lambda} = 1$.

$$= \lambda$$

Now, let's move on to the variance. We are first going to express it in terms of expected values and then we are going to apply a similar approach to the one we used for the expected value.

We start off with the well-known relationship between the expected value and the variance, the variance is equal to the expected value of the squared variable, minus the expected value of the variable, squared.

$$Var(Y) = E(Y^2) - E(Y)^2$$

This next step might seem rather unintuitive, but we are simply expressing the squared variable in a way which makes it easier to manipulate. Knowing how to do proper operations with expected values, allows us to simplify the expression and plug in values we already know:

$$= E\big((Y)(Y-1)+Y\big) - E(Y)^2$$

$$= E\big((Y)(Y-1)\big) + E(Y) - E(Y)^2$$

$$= E\big((Y)(Y-1)\big) + (\lambda - \lambda^2)$$

We turn to the definition of the expected value once again.

$$= \sum_{y=0}^{\infty} (y)(y-1)\frac{e^{-\lambda}\lambda^y}{y!} + (\lambda - \lambda^2)$$

From here on out, the steps are pretty much the same once we took for the expected value:

1) we change the starting value of the sum, since the first 2 are zeroes
2) we cross our repeating values in the numerator and denominator
3) we take out the constant factors in front of the sum
4) adjust the starting value of the sum once again
5) substitute the sum with e^{λ}

$$= \sum_{y=2}^{\infty} (y)(y-1)\frac{e^{-\lambda}\lambda^y}{y!} + (\lambda - \lambda^2)$$

$$= \sum_{y=2}^{\infty} \frac{e^{-\lambda}\lambda^y}{(y-2)!} + (\lambda - \lambda^2)$$

$$= \lambda^2 e^{-\lambda} \sum_{y=2}^{\infty} \frac{\lambda^{y-2}}{(y-2)!} + (\lambda - \lambda^2)$$

$$= \lambda^2 e^{-\lambda} \sum_{y=0}^{\infty} \frac{\lambda^y}{y!} + (\lambda - \lambda^2)$$

$$= \lambda^2 e^{-\lambda} e^{\lambda} + (\lambda - \lambda^2)$$

$$= \lambda^2 + \lambda - \lambda^2$$

$$= \lambda$$

Therefore, both the mean and variance for a Poisson Distribution are equal to lambda (λ).

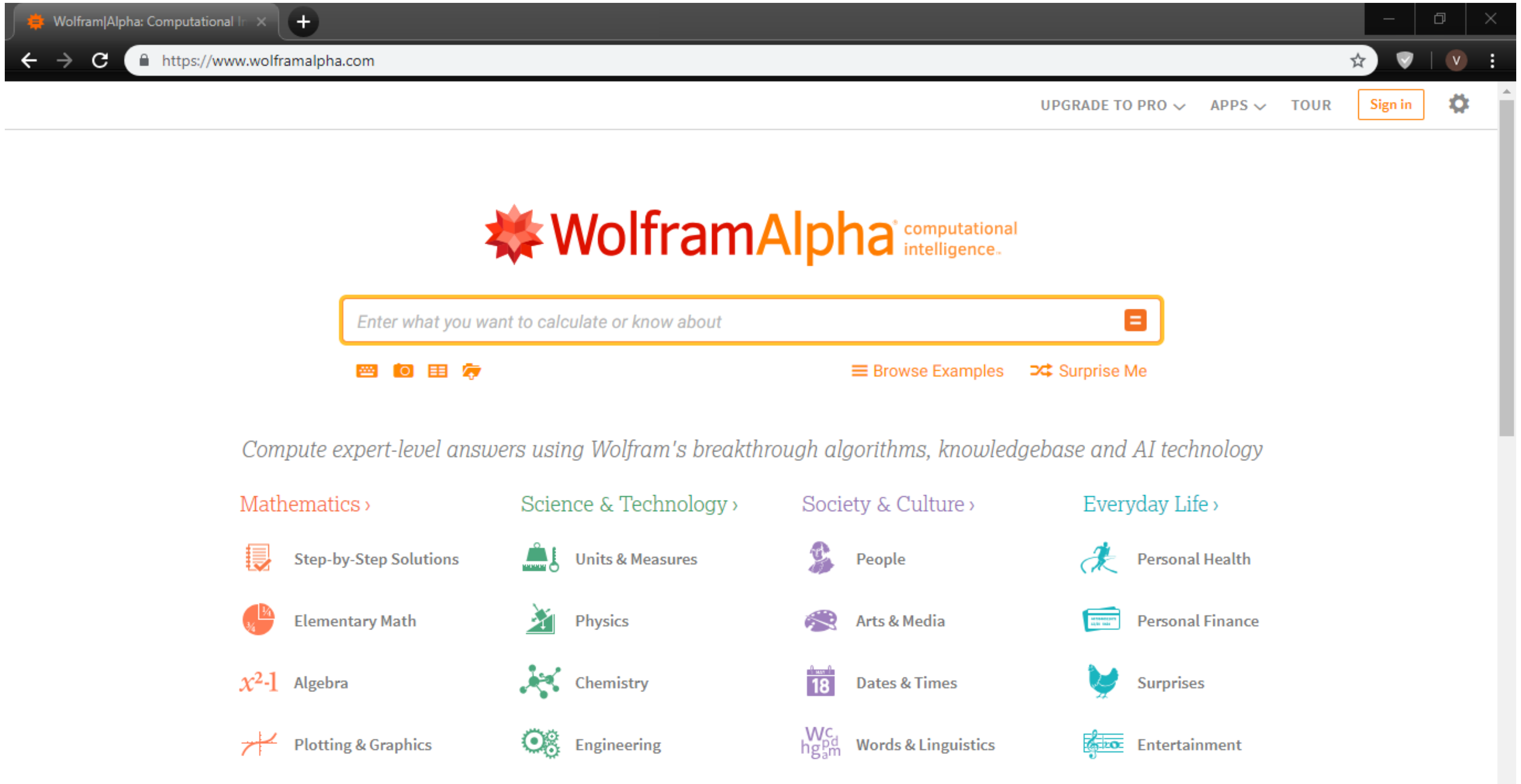
Setting up Wolfram Alpha

Step 1: Open a browser of your choosing. (We have opted for Google Chrome in this specific case, but any browser will work.)

Step2: Go to the https://www.wolframalpha.com/ website. (The page you are on, should look something like the image on the right.)

Step3: Click on the search bar which says "Enter what you want to calculate or know about." bar. It will let you type in any mathematical formula you are interested in.

Step 4: Enter the precise formula you are interested in before pressing the "Enter" key or clicking the compute button on the screen. (The one with the "=" sign.)



365√DataScience

Typing up the integral formula

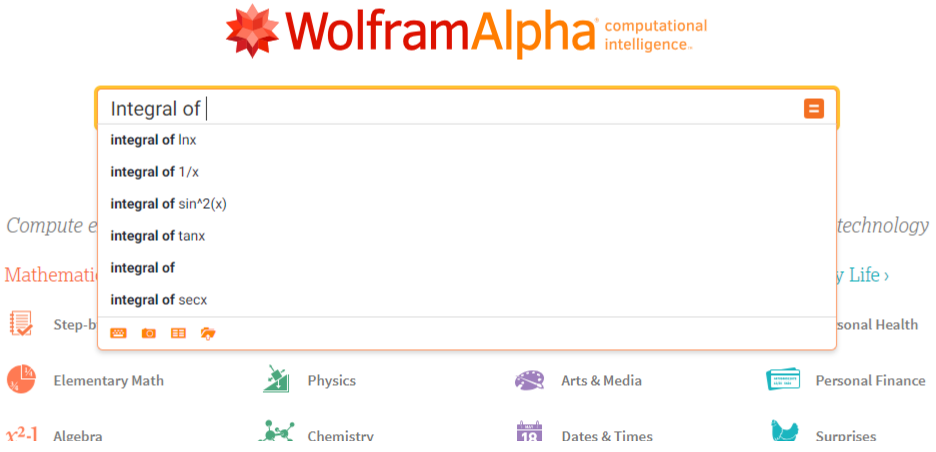
Now, let's focus on solving **integrals**.

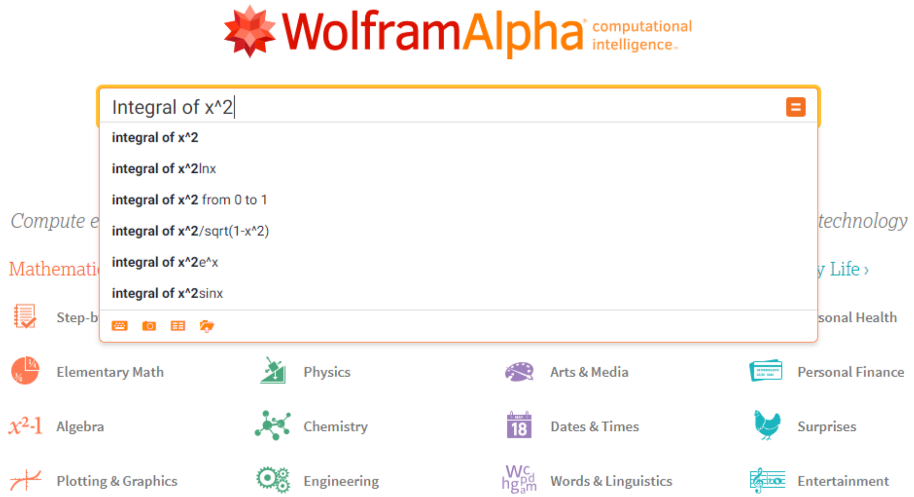
Step 1: We start by typing in the words "Integral of" into the search bar, like the image on the right.

Step 2: Wolfram Alpha's engine will likely recommend some of the most frequently searched integrals. If the one you are looking for is included, simply select it from the list. If not, proceed to the next step.

Step 3: Type in the equation, whose integral you want to solve. For instance, if the function you are interested in is "$y = x^2$", you type in "x^2".

Step 4: After we type in the appropriate function we press the "Enter" key.





Expressing Complex Formulas

Now, let's focus on the various notation we use for the **different complex functions**.

Power: If we want to express a number raised to a specific degree/power, we use the "^" sign. For instance, $\mathrm{y^\wedge n} = y^n$.

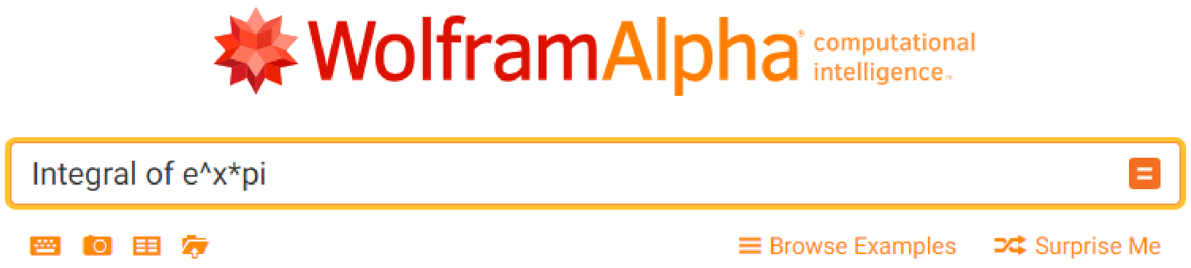
Euler's Constant: To express Euler's constant we simply type in the letter "e".

Pi: If the function we want to use features π, we simply type in the word "pi".

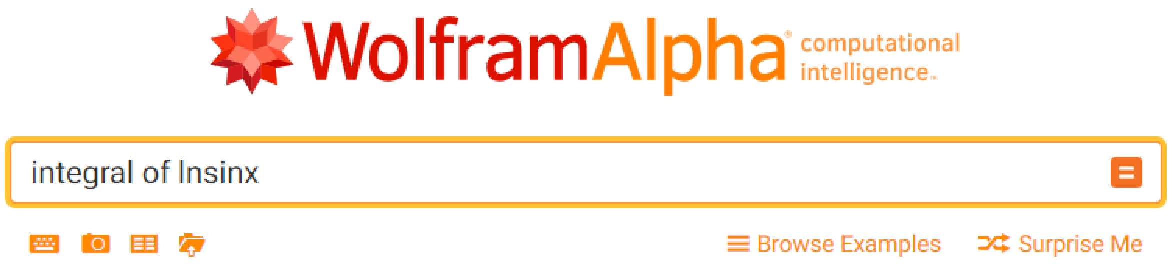
Natural Log: To express the natural log of a number, we just write the letters "ln" followed by the number

Trigonometric Functions: We simply type in the abbreviations for the given function **(e.g. sin, cos, ta cot)**

The expression below is equivalent to $\int e^x \pi dx$. Note, we omit typing "dx" because it is trivial when we only have one random variable.



The expression below is equivalent to $\int \ln \sin x \, dx$.



Understanding the Solution

Now, let's focus on a specific function like $\int e^x x dx$.

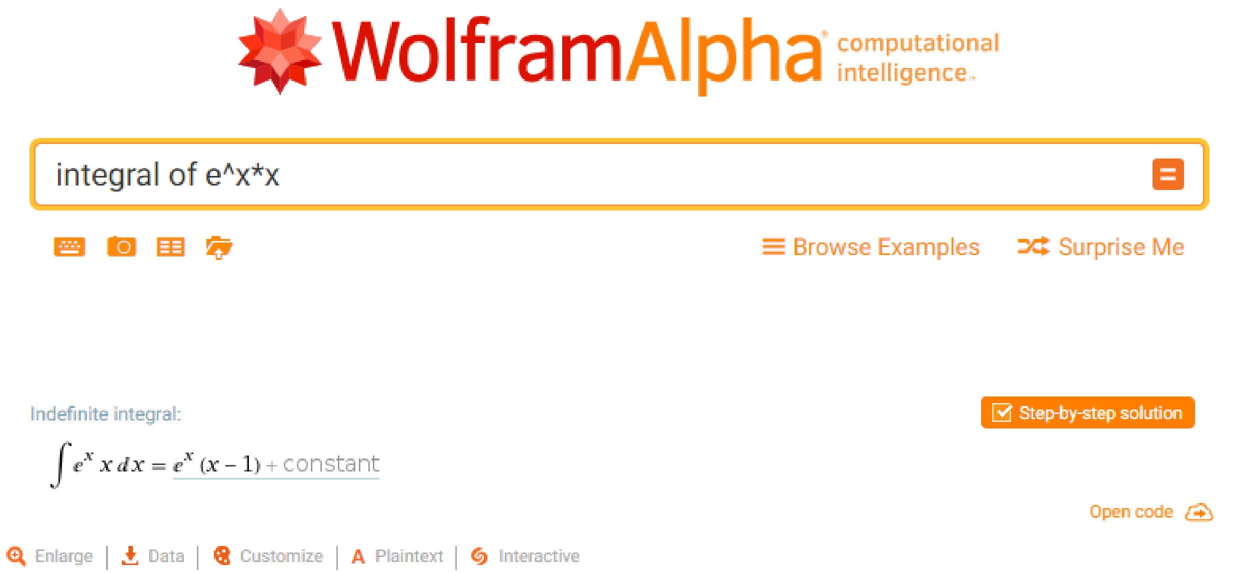
Step 1: Type in the integral like we explained earlier and press the "Enter" key:

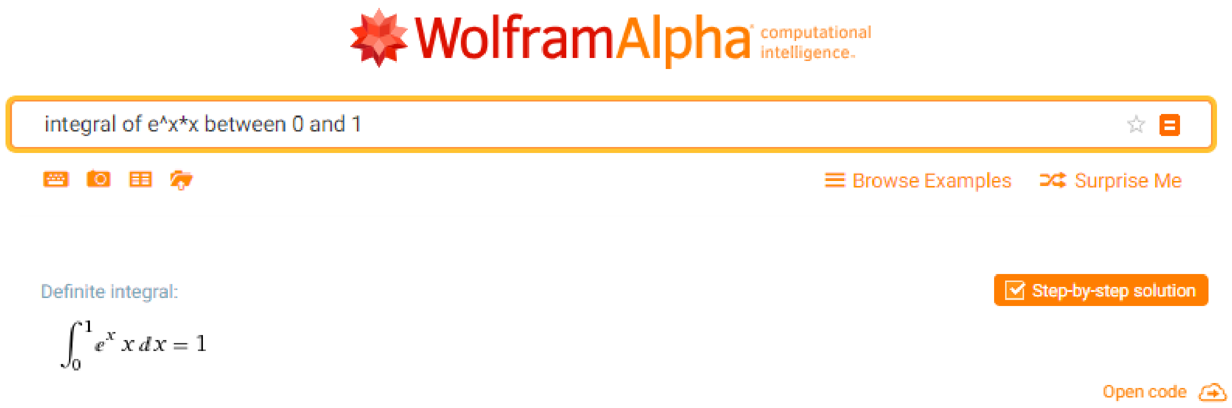
Step 2: The software provides us with several outputs after computing the integral. We focus on the first one which is called "Indefinite Integral".

Step 3: The solution we get is another function which provides an output for any value x we input into it.

Step 4: We might want to compute the values of this interval within a specific interval from a to b. In such instances, we add "between a and b" at the end of the integral before computing.

Step 5: If we have a well-defined range like this, we need to analyse the "Definite Integral" section for our solution.





Normal Distribution $E(Y), Var(Y)$

We find the expected value of a function by finding the sum of the products for each possible outcome and its chance of occurring.

Step 1: For continuous variables, this means using an integral going from negative infinity to infinity. The chance of each outcome occurring is given by the PDF, $f(y)$, so $\mathrm{E(Y)} = \int_{-\infty}^{\infty} y f(y)\, dy$.

Step 2: The PDF for a Normal Distribution is the following expression: $f(y) = \frac{1}{\sigma\sqrt{2\pi}}\, e^{\frac{-(y-\mu)^2}{2\sigma^2}}$

Step 3: Thus, the expected value equals: $\mathrm{E(Y)} = \int_{-\infty}^{\infty} y \frac{1}{\sigma\sqrt{2\pi}}\, e^{\frac{-(y-\mu)^2}{2\sigma^2}}\, dy$

Step 4: Since sigma and pi are constant numbers, we can take them out of the integral:

$$\mathrm{E(Y)} = \int_{-\infty}^{\infty} y \frac{1}{\sigma\sqrt{2\pi}}\, e^{\frac{-(y-\mu)^2}{2\sigma^2}}\, dy = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} y\, e^{\frac{-(y-\mu)^2}{2\sigma^2}}\, dy$$

Normal Distribution $E(Y), Var(Y)$

Step 5: We will substitute **t** in for $\frac{y-\mu}{\sqrt{2}\sigma}$ to make the integral more manageable. To do so, we need to transform **y** and **dy**. If t = $\frac{y-\mu}{\sqrt{2}\sigma}$, then clearly $y = \mu + \sqrt{2}\sigma t$. Knowing this, $\frac{dy}{dt} = \sqrt{2}\sigma$, so $dy = \sqrt{2}\sigma\, dt$. Therefore, we can substitute and take the constant out of the integral, before simplifying:

$$\frac{1}{\sigma\sqrt{2\pi}}\int_{-\infty}^{\infty} y\, e^{\frac{-(y-\mu)^2}{2\sigma^2}}\, dy = \frac{1}{\sigma\sqrt{2\pi}}\int_{-\infty}^{\infty}(\mu+\sqrt{2}\sigma t)\, e^{-t^2}\sqrt{2}\sigma\, dt = \frac{\sqrt{2}\sigma}{\sigma\sqrt{2\pi}}\int_{-\infty}^{\infty}(\mu+\sqrt{2}\sigma t)\, e^{-t^2}\, dt = \frac{1}{\sqrt{\pi}}\int_{-\infty}^{\infty}(\mu+\sqrt{2}\sigma t)\, e^{-t^2}\, dt$$

Step 6: We expand the expression within parenthesis and split the integral:

$$\frac{1}{\sqrt{\pi}}\int_{-\infty}^{\infty}(\mu+\sqrt{2}\sigma t)\, e^{-t^2}\, dt = \frac{1}{\sqrt{\pi}}\left[\mu\int_{-\infty}^{\infty} e^{-t^2}\, dt + \sqrt{2}\sigma\int_{-\infty}^{\infty} t\, e^{-t^2}\, dt\right]$$

Step 7: We solve the two simpler integrals:

$$\frac{1}{\sqrt{\pi}}\left[\mu\int_{-\infty}^{\infty} e^{-t^2}\, dt + \sqrt{2}\sigma\int_{-\infty}^{\infty} t\, e^{-t^2}\, dt\right] = \frac{1}{\sqrt{\pi}}\left[\mu\sqrt{\pi} + \sqrt{2}\sigma\left(-\frac{1}{2}e^{-t^2}\right)_{-\infty}^{\infty}\right]$$

Step 8: Since the exponential tends to 0, we get the following:

$$\frac{1}{\sqrt{\pi}}\left[\mu\sqrt{\pi} + \sqrt{2}\sigma\left(-\frac{1}{2}e^{-t^2}\right)_{-\infty}^{\infty}\right] = \frac{1}{\sqrt{\pi}}[\mu\sqrt{\pi} + \mathbf{0}] = \frac{\mu\sqrt{\pi}}{\sqrt{\pi}} = \mu$$

Step 9: Using Calculus we just showed that for a variable y which follows a Normal Distribution and has a PDF of $f(y) = \frac{\mathbf{1}}{\sigma\sqrt{2\pi}}\boldsymbol{e}^{\frac{-(y-\mu)^2}{2\sigma^2}}$, the expected value equals μ.

To find the Variance of the distribution, we need to use the relationship between Expected Value and Variance we already know, namely:

$$Var(Y) = E(Y^2) - [E(Y)]^2$$

Step 1: We already know the expected value, so we can plug in μ^2 for $[E(Y)]^2$, hence:

$$Var(Y) = E(Y^2) - \mu^2$$

Step 2: To compute the expected value for Y^2, we need to go over the same process we did when calculating the expected value for Y, so let's quickly go over the obvious simplifications.

$$\mathrm{E}(Y^2) - \mu^2 = \int_{-\infty}^{\infty} y^2 \frac{\mathbf{1}}{\sigma\sqrt{2\pi}} \boldsymbol{e}^{\frac{-(\boldsymbol{y}-\boldsymbol{\mu})^2}{2\sigma^2}} dy - \mu^2 =$$

$$= \frac{\mathbf{1}}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} y^2\, \boldsymbol{e}^{\frac{-(\boldsymbol{y}-\boldsymbol{\mu})^2}{2\sigma^2}} dy \, - \mu^2 =$$

$$= \frac{\mathbf{1}}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} \left(\sqrt{2}\sigma t + \mu\right)^{\mathbf{2}} \boldsymbol{e}^{-\boldsymbol{t}^{\mathbf{2}}} \sqrt{2}\sigma\, dt - \mu^2 =$$

$$= \frac{\sqrt{2}\sigma}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} \left(\sqrt{2}\sigma t + \mu\right)^{\mathbf{2}} e^{-t^2} dt - \mu^2 =$$

$$= \frac{1}{\sqrt{\pi}} \int_{-\infty}^{\infty} \left(\sqrt{2}\sigma t + \mu\right)^{\mathbf{2}} e^{-t^2} dt - \mu^2 =$$

$$= \frac{1}{\sqrt{\pi}} \left[\int_{-\infty}^{\infty} (\boldsymbol{2\sigma^2 t^2} + \boldsymbol{2\sqrt{2}\sigma\mu t} \; + \; \boldsymbol{\mu^2})\, e^{-t^2} dt \right] - \mu^2 =$$

$$= \frac{1}{\sqrt{\pi}} \left[2\sigma^2 \int_{-\infty}^{\infty} t^2\, e^{-t^2} dt + 2\sqrt{2}\sigma\mu \int_{-\infty}^{\infty} t\, e^{-t^2} dt + \mu^2 \int_{-\infty}^{\infty} e^{-t^2} dt \right] - \mu^2$$

Step 3: We already evaluated two of the integrals when finding the expected value, so let's just use the results and simplify.

$$\frac{1}{\sqrt{\pi}}\left[2\sigma^2\int_{-\infty}^{\infty} t^2\, e^{-t^2} dt + 2\sqrt{2}\sigma\mu \int_{-\infty}^{\infty} t\, e^{-t^2} dt + \mu^2 \int_{-\infty}^{\infty} e^{-t^2} dt\right] - \mu^2 =$$

$$= \frac{1}{\sqrt{\pi}}\left[2\sigma^2\int_{-\infty}^{\infty} t^2\, e^{-t^2} dt + 2\sqrt{2}\sigma\mu \times \mathbf{0} + \mu^2\sqrt{\boldsymbol{\pi}}\right] - \mu^2 =$$

$$= \frac{1}{\sqrt{\pi}}\left[2\sigma^2\int_{-\infty}^{\infty} t^2\, e^{-t^2} dt\right] + \frac{1}{\sqrt{\pi}}\mu^2\sqrt{\pi} - \mu^2 =$$

$$= \frac{1}{\sqrt{\pi}}\left[2\sigma^2\int_{-\infty}^{\infty} t^2\, e^{-t^2} dt\right] + \mu^2 - \mu^2 =$$

$$= \frac{2\sigma^2}{\sqrt{\pi}}\int_{-\infty}^{\infty} t^2\, e^{-t^2} dt$$

Step 4: We need to integrate by parts next:

$$\frac{2\sigma^2}{\sqrt{\pi}}\int_{-\infty}^{\infty} t^2\, e^{-t^2} dt = \frac{2\sigma^2}{\sqrt{\pi}}\left(\left[-\frac{t}{2}e^{-t^2}\right]_{-\infty}^{\infty} + \frac{1}{2}\int_{-\infty}^{\infty} e^{-t^2} dt\right)$$

Step 5: The exponential tends to 0 once again, so we get the following:

$$\frac{2\sigma^2}{\sqrt{\pi}}\left(\left[-\frac{t}{2}e^{-t^2}\right]_{-\infty}^{\infty}+\frac{1}{2}\int_{-\infty}^{\infty}e^{-t^2}dt\right)=$$
$$=\frac{2\sigma^2}{\sqrt{\pi}}\left(\mathbf{0}+\frac{1}{2}\int_{-\infty}^{\infty}e^{-t^2}dt\right)=$$
$$=\frac{2\sigma^2}{\sqrt{\pi}}\frac{1}{2}\int_{-\infty}^{\infty}e^{-t^2}dt=$$
$$=\frac{\sigma^2}{\sqrt{\pi}}\int_{-\infty}^{\infty}e^{-t^2}dt$$

Step 6: As we computed earlier, $\int_{-\infty}^{\infty}e^{-t^2}dt=\sqrt{\pi}$, which means:

$$\frac{\sigma^2}{\sqrt{\pi}}\int_{-\infty}^{\infty}e^{-t^2}dt=\frac{\sigma^2}{\sqrt{\pi}}\sqrt{\boldsymbol{\pi}}=\boldsymbol{\sigma^2}$$

Thus, the variance for a Variable, whose PDF looks like: $f(y)=\frac{\mathbf{1}}{\boldsymbol{\sigma}\sqrt{2\pi}}\boldsymbol{e}^{\frac{-(y-\mu)^2}{2\sigma^2}}$, equals $\boldsymbol{\sigma^2}$.