MATH CONCEPTS, NOTATION & FORMULAE

Bayes' Theorem

Formula

$$P(A \mid B) = \frac{P(B \mid A) \cdot P(A)}{P(B)}$$

In probability theory and statistics, Bayes's theorem, named after Reverend Thomas Bayes (also known as "Bayes's law" or "Bayes's rule"), describes the probability of an event, based on prior knowledge of conditions that might be related to that event.

Bayes's theorem is stated mathematically in a equation, where A and B are events and $P(B) \neq 0$:

- $P(A \mid B)$ is a conditional probability: the likelihood of event A occurring given that B is true.
- $P(B \mid A)$ is also a conditional probability: the likelihood of event B occurring given that A is true.
- $P(A)$ and $P(B)$ are the probabilities of observing A and B respectively. They are known as the marginal probability.

One of the many applications of Bayes's theorem is Bayesian inference, a particular approach to statistical inference. When applied, the probabilities involved in Bayes's theorem may have different probability interpretations. With Bayesian probability interpretation, the theorem expresses how a degree of belief, expressed as a probability, should rationally change to account for the availability of related evidence. Bayesian inference is fundamental to Bayesian statistics.

Examples

If the risk of developing health problems is known to increase with age, Bayes's theorem allows the risk to an individual of a known age to be assessed more accurately (by conditioning it on his age) than simply assuming that the individual is typical of the population as a whole.

Bayesian Inference

Formula

$$P(H \mid E) = \frac{P(E \mid H) \cdot P(H)}{P(E)}$$

Bayesian inference is a method of statistical inference in which Bayes' theorem is used to update the probability for a hypothesis as more evidence or information becomes available. Bayesian inference

is an important technique in statistics, and especially in mathematical statistics. Bayesian updating is particularly important in the dynamic analysis of a sequence of data.

Bayesian inference has found application in a wide range of activities, including science, engineering, philosophy, medicine, sport, and law. In the philosophy of decision theory, Bayesian inference is closely related to subjective probability, often called "Bayesian probability".

Examples

TBD

Binomial Distribution, Binomial Coefficient and Combination

Formula

Binomial coefficient (equivalent notations):

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

Binomial distribution formula:

$$P(k) = \binom{n}{k} p^k \left(1-p\right)^{n-k}$$

In mathematics, a combination is a selection of items from a collection, such that the order of selection does not matter (unlike permutations). For example, given three fruits, say an apple, an orange and a pear, there are three combinations of two that can be drawn from this set: an apple and a pear; an apple and an orange; or a pear and an orange.

The set of all k-combinations of a set S is often denoted by $\binom{S}{k}$. More formally, a k-combination of a set S is a subset of k distinct elements of S. If the set has n elements, the number of k-combinations is equal to the binomial coefficient which can be written using factorials, whenever $k \leq n$, and which is zero when $k > n$.

The binomial coefficients occur in many areas of mathematics, and especially in combinatorics. The symbol $\binom{n}{k}$ is usually read as "n choose k" because there are $\binom{n}{k}$ ways to choose an (unordered) subset of k elements from a fixed set of n elements.

For example, there are $\binom{4}{2} = 6$ ways to choose 2 elements from $\{1,2,3,4\}$, namely $\{1,2\}$, $\{1,3\}$, $\{1,4\}$, $\{2,3\}$, $\{2,4\}$ and $\{3,4\}$.

The binomial coefficients can be generalized to $\binom{z}{k}$ for any complex number z and integer $k \geqslant 0$, and many of their properties continue to hold in this more general form.

Examples

Calculate how many unique teams of 5 people can be formed from 10 people:

$$\binom{n}{k} = \frac{n!}{k!\,(n-k)!} = \frac{10!}{5!\,(10!-5!)} = \frac{10!}{5!\,5!} = \frac{10\cdot 9\cdot 8\cdot 7\cdot 6\cdot 5\cdot 4\cdot 3\cdot 2\cdot 1}{5\cdot 4\cdot 3\cdot 2\cdot 1} = 252$$

Calculate the probability of getting exactly 6 heads in 10 throws if a coin is bent so that it has a 40% probability of coming up heads:

$$P(6) = \binom{10}{6} 0.4^6\, 0.6^{10-6} = 0.11147673600000005$$

In [5]:

```
"""
Probability of exactly 8 events 0.010616832000 (1.06%)
Probability of X ≤ 8 events     0.998322278400 (99.83%)
Probability of X > 8 events     0.001677721600 (0.17%)
Probability of X < 8 events     0.987705446400 (98.77%)
Probability of X ≥ 8 events     0.012294553600 (1.23%)
"""
import math
import scipy.special

# Calculate the probability of a bent coin with a p probability
# of coming up heads, to get k heads in n throws:
k = 8    # Number of target successful outcomes
n = 10   # Number of trials to be performed
p = 0.5  # Percentage of expected successful outcomes (in decimal notation)
q = 1-p  # Percentage of expected unsucessful outcomes (in decimal notation)

# Check if comb() and binom() return the same result
# as doing the binomial calculation without them
print("Sanity check: ", scipy.special.comb(n, k) ==
    scipy.special.binom(n, k) == (math.factorial(n) /
    (math.factorial(k) * math.factorial(n-k))))

# Use a binomial Probability Distribution Function (PDF) to calculate for
# "exactly k events":
# P(k) = (n!/k!(n-k)!) (p^k) (q^(n-k))
pdf = scipy.special.comb(n, k) * (p**k) * (q**(n-k))
print(f"PDF (Probability of exactly {k} events in {n} trials): " +
    "{round(pdf * 100, 2)}% ({pdf})")

cdf = 0
for i in range(0, k):
    cdf = cdf + scipy.special.comb(n, i) * (p**i) * (q**(n-i))

# Use a binomial Cumulative Distribution Function (CDF) to calculate for
# "equal or greater than k events":
# P(k) = (1 - sum( (n!/k!(n-k)!) (p^k) (q^(n-k)) {0 >= k} ))
print(f"CDF (Probability of greater or equal to {k} events in {n} trials): " +
    "{round((1 - cdf) * 100, 2)}% ({1 - cdf})")

# Use a binomial Cumulative Distribution Function (CDF) to calculate for
# "less than k events":
# P(k) = (sum( (n!/k!(n-k)!) (p^k) (q^(n-k)) {0 < k} ))
print(f"CDF (Probability of less than {k} events in {n} trials): " +
    "{round((cdf) * 100, 2)}% ({cdf})")
```

```
Sanity check:  True
PDF (Probability of exactly 8 events in 10 trials): {round(pdf * 100, 2)}% ({pdf})
CDF (Probability of greater or equal to 8 events in 10 trials): {round((1 - cdf) * 100,
```

```
2)}% ({1 - cdf})
CDF (Probability of less than 8 events in 10 trials): {round((cdf) * 100, 2)}% ({cdf})
```

Coefficient of Determination

Formula

$$R^2 = 1 - \frac{TSS(Total\,Sum\,Of\,Squares)}{RSS(Residual\,Sum\,Of\,Squares)}$$

The coefficient of determination, also designated R^2 or R squared, is the proportion of the variance in the response variable that can be explained by the predictor variable. The coefficient of determination can range from 0 to 1. A value of 0 indicates that the response variable cannot be explained by the predictor variable at all. A value of 1 indicates that the response variable can be perfectly explained without error by the predictor variable.

The Residual Sum of Squares (RSS), also known as the Sum of Squared Residuals (SSR) or the Sum of Squared Estimate of Errors (SSE), is the sum of the squares of residuals (deviations predicted from actual empirical values of data). It is a measure of the discrepancy between the data and an estimation model. A small RSS indicates a tight fit of the model to the data. It is used as an optimality criterion in parameter selection and model selection.

Examples

TBD

Compound Interest

Formula

$$A = P\left(1 + \frac{r}{n}\right)^{nt}$$

The concept of compound interest is that interest is added back to the principal sum so that interest is gained on that already accumulated interest, during the next compounding period, or in other words, interest on interest. It is the result of reinvesting interest, rather than paying it out, so that interest in the next period is then earned on the principal sum plus previously accumulated interest. Compound interest is standard in finance and economics.

To use the compound interest formula, we need figures for principal amount, annual interest rate, time factor and the number of compound periods, so that:

- P is the principal investment amount (the initial deposit or loan amount).
- r is the annual interest rate (in decimal form).
- n is the number of times that interest is compounded per unit t.
- t is the time the money is invested or borrowed for.

- A is the future value of the investment/loan, including interest.

Examples

Calculate the value of an investment of an amount of $5,000 deposited into a savings account after 10 years, at an annual interest rate of 5%, compounded monthly:

$$A = 5000 \cdot \left(1 + \frac{0.05}{12}\right)^{(12 \cdot 10)} = 8235.05$$

In [ ]:
```python
import math

P = 5000  # Principal (initial investment)
r = 0.05  # Annual interest rate
n = 12    # Number of times to calculate interest within each time unit
t = 10    # Total time of the investment

total_investment_value = P * (1 + r/n) ** (n*t)
round(total_investment_value, 2) == 8235.05
```

Conditional Probability

In probability theory and statistics, given two jointly distributed random variables X and Y, the conditional probability distribution of Y given X is the probability distribution of Y when X is known to be a particular value; in some cases the conditional probabilities may be expressed as functions containing the unspecified value x of X as a parameter. When both X and Y are categorical variables, a conditional probability table is typically used to represent the conditional probability. The conditional distribution contrasts with the marginal distribution of a random variable, which is its distribution without reference to the value of the other variable.

If the conditional distribution of Y given X is a continuous distribution, then its probability density function is known as the "conditional density function". The properties of a conditional distribution, such as the moments, are often referred to by corresponding names such as the conditional mean and conditional variance.

More generally, one can refer to the conditional distribution of a subset of a set of more than two variables; this conditional distribution is contingent on the values of all the remaining variables, and if more than one variable is included in the subset then this conditional distribution is the conditional joint distribution of the included variables.

Examples

TBD

Continuous Compound Return

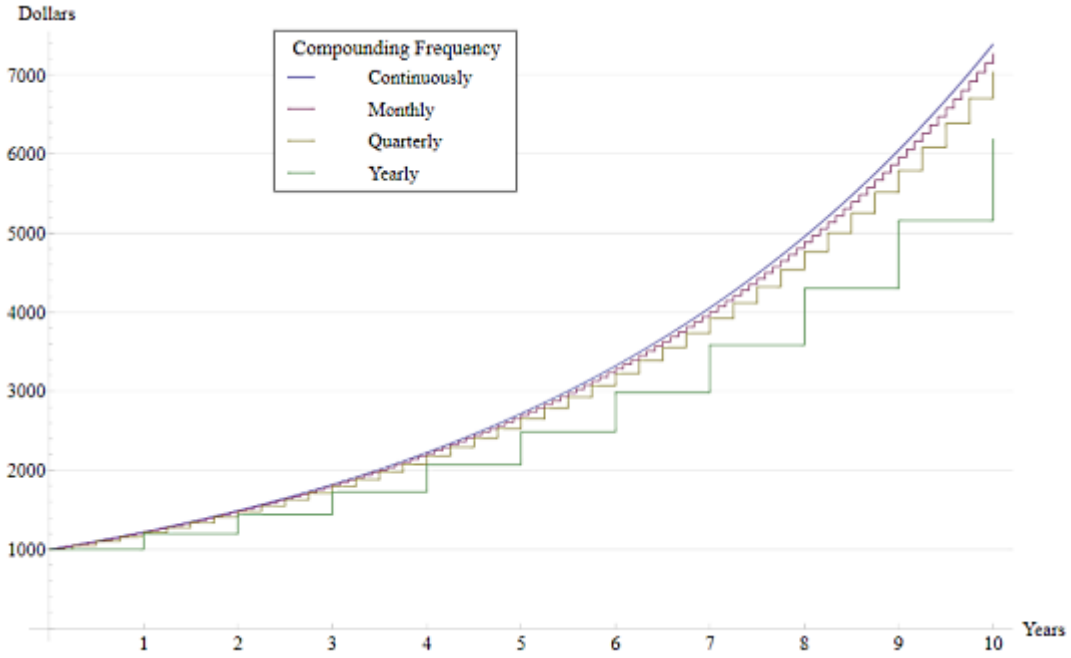
Formula

$$r_{continuous} = \frac{\ln \frac{Value_{End}}{Value_{Start}}}{Periods}$$

A compound interest is interest calculated on the initial principal and also on the accumulated interest of previous periods of a deposit or loan. The effect of compound interest depends on frequency.

Continuous compounding is the mathematical limit that compound interest can reach. It can be thought of as making the compounding period infinitesimally small, achieved by taking the limit as n goes to infinity. It is an extreme case of compounding, since most interest is compounded on a monthly, quarterly or semiannual basis.

As the number of compounding periods n reaches infinity in continuous compounding, the continuous compound interest rate is referred to as the "force of interest" δ.



Examples

Calculate the continuously compounded rate of return of a 1,600 investment which is worth 7,400 after 8.5 years:

$$\frac{\ln \frac{7400}{1600}}{8.5} = 0.18017 = 18\%$$

Calculate the total weight of a baby elephant of 200Kg weight, after 3 years, assuming a continuous growth of 5% per year:

$$200 \cdot e^{(0.05)(3)} = 232.4\text{Kg}$$

Convert an interest rate from a compounding basis to another compounding basis:

$$r_2 = \left[\left(1 + \frac{r_1}{n_1}^{\frac{n1}{n2}}\right) - 1\right] n_2$$

Convert from a continuously compounding interest to another compounding interest rate, where δ is the interest rate on a continuous compounding basis, and r is the stated interest rate, with a compounding frequency n:

$$\delta = n \ln\left(1 + \frac{r}{n}\right)$$

In [ ]:
```python
import math

# Calculate the continuously compounded rate of return of a
# 1,600 investment which is worth 7,400 after 8.5 years
math.log(7400/1600) / 8.5 == 0.18017369070169278
```

Derivative

Notation

Common notations to express derivatives:

Lagrange

In Lagrange's notation, the derivative of f is expressed as f' (pronounced "f *prime*").

This notation is probably the most common when dealing with functions with a single variable. If, instead of a function, we have an equation such as $y = f(x)$, we can also write y' to represent the derivative. This, however, is less common to do.

Leibniz

In Leibniz's notation, the derivative of f is expressed as $\frac{d}{dx}f(x)$. When we have an equation such as $y = f(x)$, we can express the derivative as $\frac{dy}{dx}$. Here, $\frac{d}{dx}$ serves as an operator that indicates a differentiation with respect to x. This notation also allows us to directly express the derivative of an expression without using a function or a dependent variable. For example, the derivative of x^2 can be expressed as $\frac{d}{dx}(x^2)$.

This notation, while less comfortable than Lagrange's notation, becomes very useful when dealing with integral calculus, differential equations, and multivariable calculus.

Newton

In Newton's notation, the derivative of f is expressed as $\dot{f}$ and the derivative of $y = f(x)$ is expressed as $\dot{y}$.

This notation is mostly common in Physics and other sciences where calculus is applied in a real-world context.

Derivatives are the result of performing a differentiation process upon a function or an expression. The derivative of a function of a real variable measures the sensitivity to change of the function value (output value) with respect to a change in its argument (input value).

Derivatives are a fundamental tool of Calculus. For example, the derivative of the position of a moving object with respect to time is the object's velocity: it measures how quickly the position of the object changes when time advances.

The derivative of a function of a single variable at a chosen input value, when it exists, is the slope of the tangent line to the graph of the function at that point. The tangent line is the best linear approximation of the function near that input value.

For this reason, the derivative is often described as the "instantaneous rate of change", the ratio of the instantaneous change in the dependent variable (the y-axis on a graph) to that of the independent variable (the x-axis on a graph).

The process of finding a derivative is called "differentiation". The reverse process is called "antidifferentiation". The fundamental theorem of calculus relates antidifferentiation with integration. Differentiation and integration constitute the two fundamental operations in single-variable calculus.

In mathematics and computer algebra, automatic differentiation, also called algorithmic differentiation, computational differentiation, auto-differentiation, or simply *autodiff*, is a set of techniques to numerically evaluate the derivative of a function specified by a computer program.

Automatic differentiation exploits the fact that every computer program, no matter how complicated, executes a sequence of elementary arithmetic operations (addition, subtraction, multiplication, division, etc.) and elementary functions (exp, log, sin, cos, etc.). By applying the chain rule (a formula to compute the derivative of a composite function) repeatedly to these operations, derivatives of arbitrary order can be computed automatically, accurately to working precision, and using at most a small constant factor more arithmetic operations than the original program.

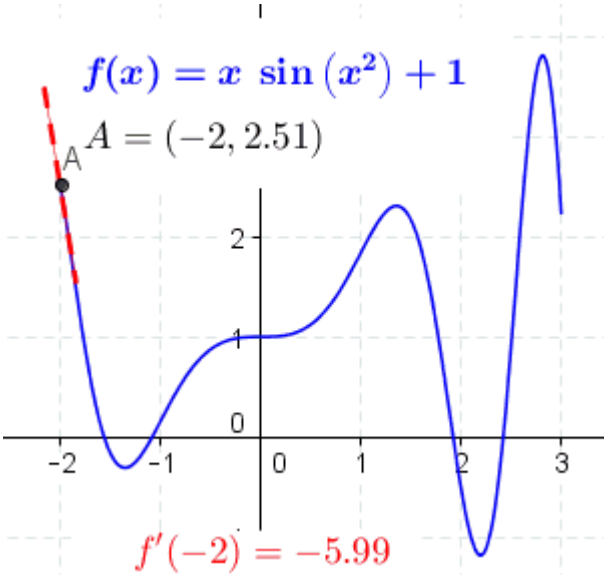
Examples

$f(x) = x \sin\left(x^2\right) + 1 \left\{-2 < x < 3\right\}$

$g(x) = \frac{d}{dx} f\left(x\right)$

$A = \left(a,\ f\left(a\right)\right)$

$y = g\left(a\right)\left(x - a\right) + f\left(a\right)$



Distance

Formula

$$d(A, B) = \sqrt{(\Delta x)^2 + (\Delta y)^2}$$

The distance formula is a variant of the Pythagorean Theorem used in geometry. The distance from a point A to a point B is sometimes denoted as $|AB|$. In most cases, "distance from A to B" is interchangeable with "distance from B to A".

In mathematics, the Euclidean distance or Euclidean metric is the "ordinary" straight-line distance between two points in Euclidean space. With this distance, Euclidean space becomes a metric space.

Examples

Calculate the distance between point $A = (2, 2)$ and point $B = (-1, -2)$:

$$d(A, B) = \sqrt{(-1 - 2)^2 + (-2 - 2)^2} = 5$$

In [ ]:
```python
import math

A = [2, 2]
B = [-1, -2]

# The math.hypot() function calculates the hypotenuse
# of a right triangle, which is given by sqrt(a*a + b*b)
d_AB = math.hypot(B[1] - A[1], B[0] - A[0])

# Output is 5.0
print(d_AB)
```

Element Of

Symbol

$$\in$$

The relation *is an element of*, also called "set membership", is denoted by the symbol $\in$. Writing $x \in A$ means that "x is an element of A". Some equivalent expressions are "x is a member of A", "x belongs to A", "x is in A" and "x lies in A".

The negation of set membership is denoted by the symbol $\notin$.

For the relation $\in$, the *converse relation* or *transpose* $\in^T$ may be written as $A \ni x$, meaning A contains or includes x.

Examples

If set $A = \{1, 2, -3, 7\}$ then $2 \in A$, $A \ni 2$, $5 \notin A$, and $A \not\ni 5$.

Empty Set

Symbol

$$\varnothing \text{ or } \emptyset \text{ or } \{\}$$

In mathematics, the empty set is the unique set having no elements. Its size or cardinality (count of elements in a set) is zero. Some axiomatic set theories ensure that the empty set exists by including an axiom of empty set, while in other theories, its existence can be deduced. Many possible properties of sets are vacuously true for the empty set.

In some textbooks and popularizations, the empty set is referred to as the "null set". However, null set is a distinct notion within the context of measure theory, in which it describes a set of measure zero (which is not necessarily empty). The empty set may also be called the "void set".

It is important not to confuse $\emptyset$ with $\{\emptyset\}$. The former has no elements, while the latter has one element. If we visualize the empty set as an empty paper bag, then we can visualize $\{\emptyset\}$ as a paper bag inside a paper bag.

Examples

TBD

Euler's Number

Symbol

e

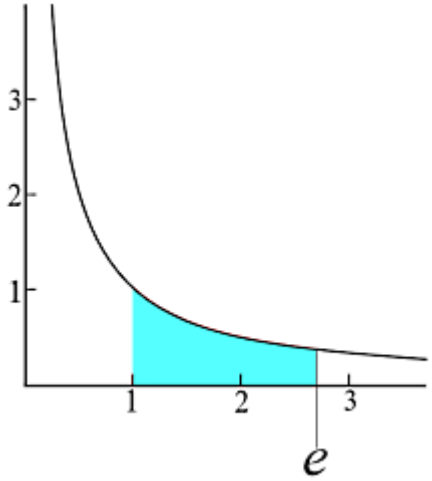
The number e, also known as "Euler's number", is a mathematical constant approximately equal to 2.718281828459045, and can be characterized in many ways. It is the base of the natural logarithm. It is the limit of $(1+1/n)^n$ as n approaches infinity, an expression that arises in the study of compound interest. It can also be calculated as the sum of the infinite series:

$$e = \sum_{n=0}^{\infty} \frac{1}{n!} = \frac{1}{1} + \frac{1}{1} + \frac{1}{1 \cdot 2} + \frac{1}{1 \cdot 2 \cdot 3} + \cdots$$

It is also the unique positive number a such that the graph of the function $y = ax$ has unit slope at $x = 0$.

The (natural) exponential function $f(x) = e^x$ is the unique function which is equal to its own derivative, with the initial value $f(0) = 1$ (and hence one may define e as $f(1)$).

The natural logarithm, or logarithm to base e, is the inverse function to the natural exponential function. The natural logarithm of a number $k > 1$ can be defined directly as the area under the curve $y = 1/x$ between $x = 1$ and $x = k$, in which case e is the value of k for which this area equals one:



The discovery of the constant itself is credited to Jacob Bernoulli in 1683, who attempted to find the value of the following expression (which is equal to e):

$$\lim_{n \to \infty} \left(1 + \frac{1}{n}\right)^n$$

Examples

TBD

Event

In probability theory, an event is a set of possible outcomes of an experiment (a subset of the sample space) to which a probability is assigned. Any subset of S is an event. A single outcome may

be an element of many different events, and different events in an experiment are usually not equally likely, since they may include very different groups of outcomes.

An event defines a complementary event, namely the complementary set (the event not occurring), and together these define a Bernoulli trial: did the event occur or not?

Typically, when the sample space is finite, any subset of the sample space is an event (i.e. all elements of the power set of the sample space are defined as events). However, this approach does not work well in cases where the sample space is uncountably infinite. So, when defining a probability space, it is possible, and often necessary, to exclude certain subsets of the sample space from being events.

Examples

Considering E_1 as the experiment of tossing a die and getting an even number event, then $A_1 = \{2, 4, 6\}$, and $A'_1 = \{1, 3, 5\}$.

Factorial

Formula

$$n! = \prod_{i=1}^{n} i \text{ or } n! = n \cdot (n-1) \cdot (n-2) \cdot (n-3) \cdot \ldots 3 \cdot 2 \cdot 1$$

In mathematics, the factorial of a positive integer n, denoted by $n!$, is the product of all positive integers less than or equal to n. The value of $0!$ is 1, according to the convention for an empty product.

The factorial operation is encountered in many areas of mathematics, notably in combinatorics, algebra, and mathematical analysis. Its most basic use counts the possible distinct sequences – the permutations – of n distinct objects: there are $n!$.

The factorial function can also be extended to non-integer arguments while retaining its most important properties by defining $x! = \Gamma(x + 1)$, where Γ is the gamma function; this is *undefined* when x is a negative integer.

Examples

$5! = 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1 = 120$

$\frac{7!}{5!} = \frac{7 \cdot 6 \cdot 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1}{5 \cdot 4 \cdot 3 \cdot 2 \cdot 1} = 42$

$0! = 1$

In [ ]:
```
import math

math.factorial(5) == 120
```

Laplace's Rule

Formula

$$P(A) = \frac{m}{n}$$

Laplace's rule is extremely important because it allows us to compute the probability of an event, always that the elementary events are equiprobable, that is, that all possible outcomes have the same probability. Under these conditions, the probability of an event is obtained by dividing m, the number of results that form the event (favorable cases), by n, the number of possible outcomes (possible cases). However, it is important to note that this rule only works when all cases are equiprobable.

Examples

1. If a family has two children, and we assume that the probability of being a man is the same as that of being a woman, what is the probability of having both children of the same sex?

   As we are going to apply Laplace's rule, we will consider the results in order. So, in this case, the favorable cases are MM and WW, and our sample space is $\Omega = \{MM, MW, WM, WW\}$.

   Therefore, the probability is:

   $$P(A) = \frac{2}{4} = \frac{1}{2} = 0.5 = 50\%$$

2. What is the probability of throwing two coins and getting heads both times?

   Every time you throw a coin it is equally probable to come out with heads or tails, so that all possible outcomes are equiprobable.

   Our sample space has four elements $\Omega = \{HH, HT, TH, TT\}$, and there is only one case in favor of event "get heads twice".

   Therefore, the probability is:

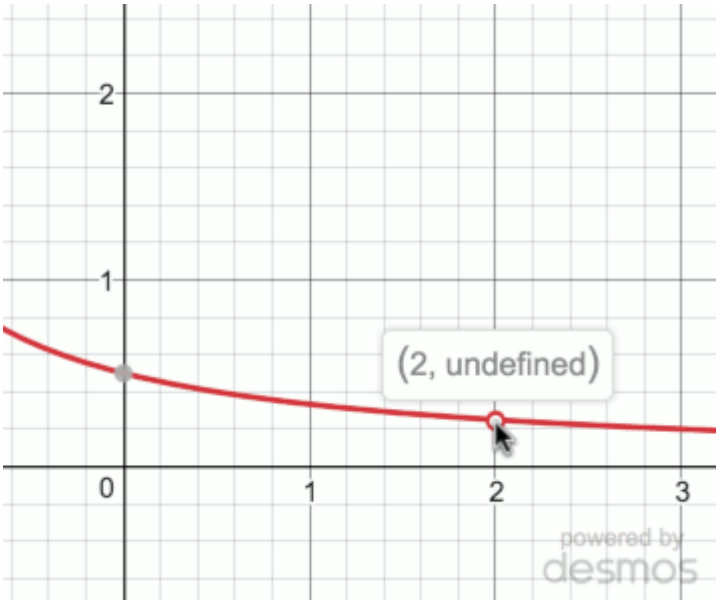
   $$P(A) = \frac{1}{4} = 0.25 = 25\%$$

Limit

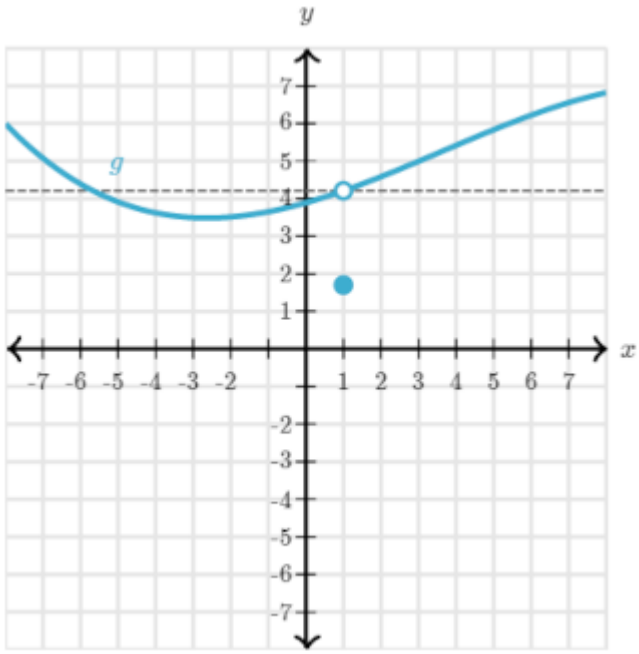
Symbol

$$\lim_{x \to \infty} f(x)$$

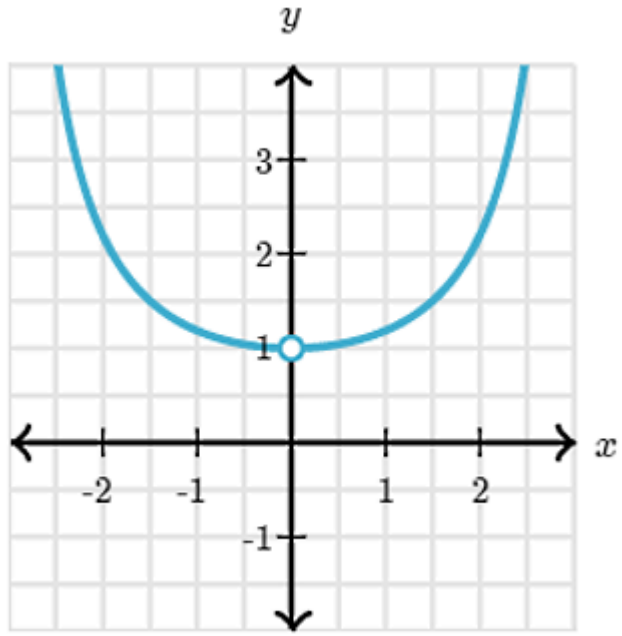
There's an important difference between the value a function is approaching (what we call the limit) and the value of the function itself. In the graph below, as we get closer and closer to $x = 2$ from both the left and the right, we seem to approach $y = 0.25$. We see that the function value is *undefined*, but the limit value is approximately 0.25.



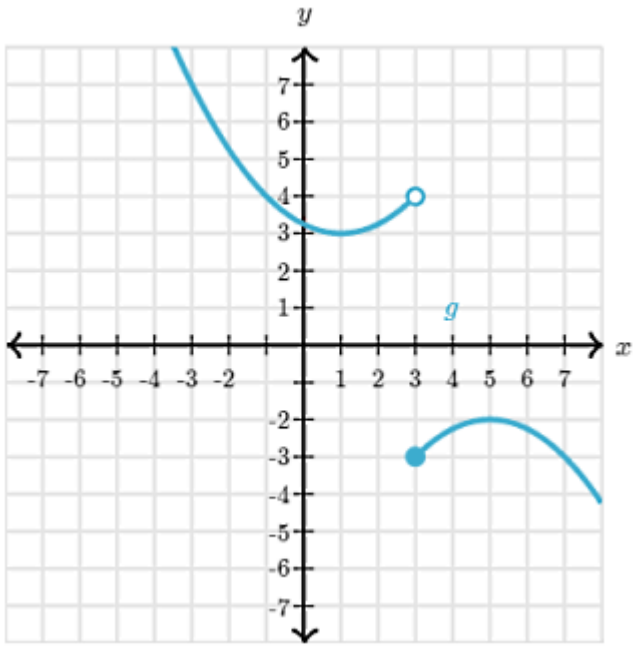
It is also possible for the function value to be different from the function limit value. In the next graph, the limit seems to be somewhere between $y = 4$ and $y = 5$, but slightly closer to $y = 4$.



Just because a function is *undefined* for some x-value, doesn't mean there is no limit. Holes in graphs happen with rational functions, which become *undefined* when their denominators are zero. The graph of $y = x/sin(x)$ is a classic example. Notice there is a hole at $x = 0$ because the function is *undefined* there. So, the function value is irrelevant to find the limit. All that matters is figuring out what the y-values are approaching as we get closer and closer:



When the function is defined for some x-value, that doesn't mean that the limit necessarily exists. The following graph shows something that can happen when we're working with piecewise functions. Notice how we're not approaching the same y-value from both sides of $x = 3$:



A graph can help us to approximate a limit by allowing to estimate the finite y-value we are approaching as we get closer and closer to some x-value (from both sides). But even using a graphing calculator, we can only zoom in so far. We can't get infinitely close to the x-value we're interested in. The best we can do is reason about what the y-values seem to be approaching as our x-values get closer and closer to some number.

Examples

TBD

Marginal Distribution (Probability)

In probability theory and statistics, the marginal distribution of a subset of a collection of random variables is the probability distribution of the variables contained in the subset. It gives the probabilities of various values of the variables in the subset without reference to the values of the other variables. This contrasts with a conditional distribution, which gives the probabilities contingent upon the values of the other variables.

Marginal variables are those variables in the subset of variables being retained. These concepts are "marginal" because they can be found by summing values in a table along rows or columns, and writing the sum in the margins of the table. The distribution of the marginal variables (the marginal distribution) is obtained by marginalizing – that is, focusing on the sums in the margin – over the distribution of the variables being discarded, and the discarded variables are said to have been marginalized out.

The context here is that the theoretical studies being undertaken, or the data analysis being done, involves a wider set of random variables but that attention is being limited to a reduced number of those variables. In many applications, an analysis may start with a given collection of random variables, then first extend the set by defining new ones (such as the sum of the original random variables) and finally reduce the number by placing interest in the marginal distribution of a subset (such as the sum). Several different analyses may be done, each treating a different subset of variables as the marginal variables.

Examples

TBD

Mean

Symbol

μ or $\bar{x}$

In statistics, the term *average* refers to any of the measures of central tendency. The mean is the most commonly used and readily understood measure of central tendency in a data set.

To express a population mean, use the μ or μ_x symbols. To express the arithmetic mean of a sample of the population, use a $\bar{x}$ symbol.

The arithmetic mean of a set of observed data is defined as being equal to the sum of the numerical values of each and every observation, divided by the total number of observations.

Symbolically, if we have a data set consisting of the values a_1, a_2, ..., a_n, then the arithmetic mean $\bar{x}$ is defined by the formula:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{n}(x_1 + \cdots + x_n)$$

Examples

The arithmetic mean of Z set is 6:

If $Z = \{1, 5, 12\}$

Then $\mu_z = \frac{1+5+12}{3} = \frac{18}{3} = 6$

Or (using summation notation) $\mu_z = \frac{1}{n}(\sum_{i=1}^{n} z_i) = 6$

Mean Squared Error

Formula

$$MSE = \frac{1}{n}\sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$$

In statistics, the Mean Squared Error (MSE) or Mean Squared Deviation (MSD) of an estimator (a procedure for estimating an unobserved quantity) measures the average of the squares of the errors (the average squared difference between the estimated values and the actual value).

MSE is a risk function, corresponding to the expected value of the squared error loss. The fact that MSE is almost always strictly positive (and not zero) is because of randomness or because the estimator does not account for information that could produce a more accurate estimate.

The MSE is a measure of the quality of an estimator, it is always non-negative, and values closer to zero are better.

The MSE is the second moment (about the origin) of the error, and thus incorporates both the variance of the estimator (how widely spread the estimates are from one data sample to another) and its bias (how far off the average estimated value is from the true value). For an unbiased estimator, the MSE is the variance of the estimator. Like the variance, MSE has the same units of measurement as the square of the quantity being estimated.

In an analogy to standard deviation, taking the square root of MSE yields the Root Mean Squared Error or Root Mean Squared Deviation (RMSE or RMSD), which has the same units as the quantity being estimated. For an unbiased estimator, the RMSE is the square root of the variance, known as the standard error.

Examples

Considering a set of actual and predicted values, compute the respective squared error by squaring the difference between each predicted value ($\hat{y}$) and its actual (observed) value (y).

For example, for row 1 in the table below, calculate $(28 - 19)^2 = 9^2 = 81$.

Actual Values	Predicted Values	Squared Error
19	28	81
37	33	16
25	20	25
9	16	49
22	15	49

After calculating the squared error for each row, sum all of those and then divide the sum's total by the number of rows, to get the MSE of the whole set:

$$\frac{81 + 16 + 25 + 49 + 49}{5} = 44$$

The Mean Absolute Error (MAE) is similar to the MSE, however, it does not square the error of each result. Instead, it sums up their absolute values:

$$\frac{1}{N}\sum_{i=1}^{N} |y_i - \hat{y}_i|^2$$

Another common error function is the Root Mean Squared Error (RMSE), which only difference from the MAE is to take the squared root of the sum's total (instead of dividing it by the number of input values):

$$\sqrt{81 + 16 + 25 + 49 + 49} = 6.63$$

Median

Symbol

$\tilde{X}$ or M or Med

Formally, a median of a population is any value such that at most half of the population is less than the proposed median and at most half is greater than the proposed median.

Medians may not be unique. If each set contains less than half the population, then some of the population is exactly equal to the unique median.

The median is well-defined for any ordered (one-dimensional) data, and is independent of any distance metric. The median can thus be applied to classes which are ranked but not numerical (e.g. working out a median grade when students are graded from A to F), although the result might be halfway between classes if there is an even number of cases.

The median can be used as a measure of location when one attaches reduced importance to extreme values, typically because a distribution is skewed, extreme values are not known, or outliers

are untrustworthy, i.e., may be measurement/transcription errors.

Examples

$\tilde{X} = \frac{1}{2}\left(x_{\lfloor (n+1)/2 \rfloor} + x_{\lceil (n+1)/2 \rceil}\right)$

Natural Logarithm

Symbol

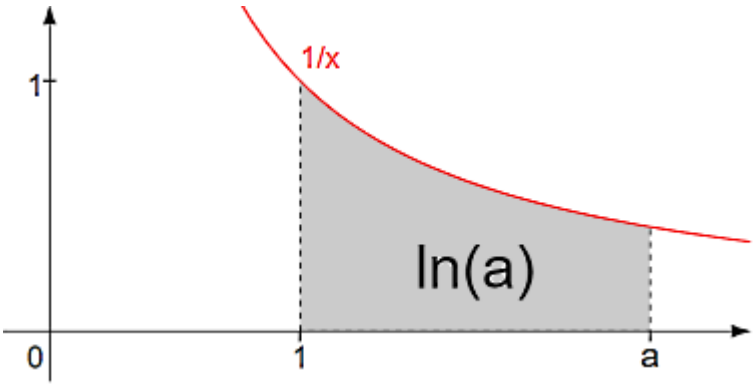
$\ln x$ or $\ln(x)$ or $\log x$ or $\log(x)$ or $\log_e x$ or $\log_e(x)$

Logarithms are useful for solving equations in which the unknown appears as the exponent of some other quantity. For example, logarithms are used to solve for the half-life, decay constant, or unknown time in exponential decay problems. They are important in many branches of mathematics and scientific disciplines, and are used in finance to solve problems involving compound interest.

The natural logarithm of a number is its logarithm to the base of the mathematical constant e. The natural logarithm of x is generally written as $\ln x$, $\log_e x$, or sometimes, if the base e is implicit, simply $\log x$. To prevent ambiguity, parentheses may also be added, giving $\ln(x)$, $\log_e(x)$, or $\log(x)$, particularly when the argument to the logarithm is not a single symbol.

The natural logarithm of x is the power to which e would have to be raised to equal x. The natural logarithm of e itself, $\ln_e$, is 1, because $e^1 = e$, while the natural logarithm of 1 is 0, since $e^0 = 1$.

The natural logarithm can be defined for any positive real number a as the area under the curve $y = 1/x$ from 1 to a (with the area being negative when $0 < a < 1$):



The simplicity of this definition, which is matched in many other formulas involving the natural logarithm, leads to the term "natural".

Examples

The natural logarithm of x is the power to which e would have to be raised to equal x:

$\ln(7.5) = 2.0149030205422647$ because $e^{2.0149030205422647} = 7.5$

In [ ]:
```
import math

round(math.pow(math.e, math.log(7.5, math.e)), 14) == 7.5
```

Natural Numbers

Symbol

$\mathbb{N}$

In mathematics, the natural numbers are those used for counting (as in "there are six coins on the table") and ordering (as in "this is the third largest city in the country"). In common mathematical terminology, words colloquially used for counting are "cardinal numbers", and words used for ordering are "ordinal numbers". The natural numbers can, at times, appear as a convenient set of codes (labels or "names"); that is, as what linguists call nominal numbers, forgoing many or all of the properties of being a number in a mathematical sense. The set of natural numbers is often denoted by the symbol $\mathbb{N}$.

Some definitions, including the standard ISO 80000-2, begin the natural numbers with 0, corresponding to the non-negative integers 0, 1, 2, 3, ... (collectively denoted by the symbol $\mathbb{N}_0$), whereas others start with 1, corresponding to the positive integers 1, 2, 3, ... (collectively denoted by the symbol $\mathbb{N}_1$).

Texts that exclude zero from the natural numbers sometimes refer to the natural numbers together with zero as the whole numbers, while in other writings, that term is used instead for the integers (including negative integers).

Properties of the natural numbers, such as divisibility and the distribution of prime numbers, are studied in number theory. Problems concerning counting and ordering, such as partitioning and enumerations, are studied in combinatorics.

In common language, particularly in primary school education, natural numbers may be called counting numbers to intuitively exclude the negative integers and zero, and also to contrast the discreteness of counting to the continuity of measurement — a hallmark characteristic of real numbers.

Examples

- $\mathbb{N}_0 = \mathbb{N}^0 = \mathbb{N} \cup \{0\} = \{0, 1, 2, 3, \ldots\}$
- $\mathbb{N}^* = \mathbb{N}^+ = \mathbb{N}_1 = \mathbb{N}_{>0} = \{1, 2, 3, \ldots\}$

Norm

In mathematics, a norm is a function from a vector space over the real or complex numbers to the non-negative real numbers, that satisfies certain properties pertaining to scalability and additivity

and takes the value zero only if the input vector is zero. A pseudonorm or seminorm satisfies the same properties, except that it may have a zero value for some nonzero vectors.

The Euclidean norm, or 2-norm, is a specific norm on a Euclidean vector space that is strongly related to the Euclidean distance. It is also equal to the square root of the inner product of a vector with itself.

A vector space on which a norm is defined is called a normed vector space. In a similar manner, a vector space with a seminorm is called a seminormed vector space.

Some of the most useful operators in linear algebra are norms. Informally, the norm of a vector tells us how big a vector is. The notion of size under consideration here concerns not dimensionality but rather the magnitude of the components. A vector norm is a function f that maps a vector to a scalar, satisfying a handful of properties.

In deep learning, we are often trying to solve optimization problems: *maximize* the probability assigned to observed data; *minimize* the distance between predictions and the ground-truth observations. Assign vector representations to items (like words, products, or news articles) such that the distance between similar items is minimized, and the distance between dissimilar items is maximized. Oftentimes, the objectives, perhaps the most important components of deep learning algorithms (besides the data), are expressed as norms.

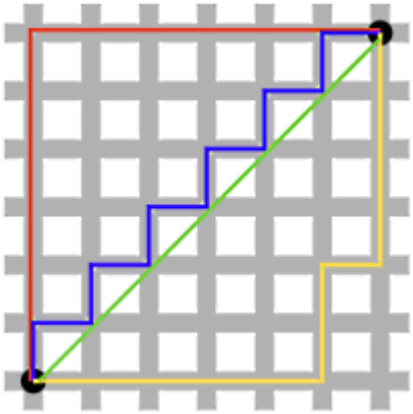
Examples

The Euclidean distance is a norm: specifically, it is the $L2$ norm. Suppose that the elements in the n-dimensional vector x are $x_1, \ldots, x_n$. The L_2 norm of x is the square root of the sum of the squares of the vector elements:

$$\|\mathbf{x}\|_2 = \sqrt{\sum_{i=1}^{n} x_i^2}$$

In deep learning, the squared L_2 norm is more often used, however, it is also frequent to find the L_1 norm, which is expressed as the sum of the absolute values of the vector elements:

$$\|\mathbf{x}\|_1 = \sum_{i=1}^{n} |x_i|$$



In *taxicab geometry* (a L_1 norm example), the red, yellow, and blue paths all have the same shortest path length of 12. In Euclidean geometry, the green line has length $6\sqrt{2} \approx 8.49$ and is the unique

shortest path.

As compared with the L_2 norm, the L_1 norm is less influenced by outliers.

Both the L_2 norm and the L_1 norm are special cases of the more general L_p norm:

$$\|\mathbf{x}\|_p = \left(\sum_{i=1}^{n} |x_i|^p\right)^{1/p}$$

Analogous to L_2 norms of vectors, the Frobenius norm of a matrix $X \in \mathbb{R}^{m \times n}$ is the square root of the sum of the squares of the matrix elements:

$$\|\mathbf{X}\|_F = \sqrt{\sum_{i=1}^{m}\sum_{j=1}^{n} x_{ij}^2}$$

In [15]:
```
import numpy as np

u = np.array([3, -4])

# Calculate the L1 norm of a vector
# Output is 7
print(np.abs(u).sum())

# Calculate the L2 norm of a vector
# Output is 5.0
print(np.linalg.norm(u))

# Calculate the Frobenius norm of a matrix
# Output is 6.0
print(np.linalg.norm(np.ones((4, 9))))
```

```
7
5.0
6.0
```

Normal Distribution

Symbol

$\mathcal{N}(\mu, \sigma^2)$

In probability theory, a normal distribution (or Gaussian or Gauss or Laplace–Gauss or "bell curve" - though there are other bell shaped curves which are not a normal distribution) is a type of continuous probability distribution for a real-valued random variable. The general form of its PDF (Probability Density Function) is:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

The parameter μ is the mean or expectation of the distribution (and also its median and mode), while the parameter σ is its standard deviation. The variance of the distribution is σ^2. A random variable with a Gaussian distribution is said to be normally distributed, and is called a normal deviate.

Normal distributions are important in statistics and are often used in the natural and social sciences to represent real-valued random variables whose distributions are not known. Their importance is partly due to the central limit theorem. It states that, under some conditions, the average of many samples (observations) of a random variable with finite mean and variance is itself a random variable—whose distribution converges to a normal distribution as the number of samples increases. Therefore, physical quantities that are expected to be the sum of many independent processes, such as measurement errors, often have distributions that are nearly normal.

Examples

Standard normal distribution

The simplest case of a normal distribution is known as the standard normal distribution. This is a special case when $\mu = 0$ and $\sigma = 1$, and it is described by this probability density function (PDF):

$$\varphi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2}$$

General normal distribution

Every normal distribution is a version of the standard normal distribution, whose domain has been stretched by a factor σ (the standard deviation) and then translated by μ (the mean value):

$$f(x \mid \mu, \sigma^2) = \frac{1}{\sigma}\varphi\left(\frac{x-\mu}{\sigma}\right)$$

Outcome

In probability theory, an outcome is a possible result of an experiment or trial. Each possible outcome of a particular experiment is unique, and different outcomes are mutually exclusive (only one outcome will occur on each trial of the experiment).

The set of all the possible outcomes of an experiment form the elements of a *sample space*, which is usually noted as S or Ω.

For the experiment where we flip a coin twice, the four possible outcomes that make up our sample space are (H, T), (T, H), (T, T) and (H, H), where "H" represents a "heads", and "T" represents a "tails". Outcomes should not be confused with events, which are sets (or informally, "groups") of outcomes. For comparison, we could define an event to occur when "at least one *heads*" is flipped in the experiment - that is, when the outcome contains at least one *heads*. This event would contain all outcomes in the sample space except the element (T, T).

Since individual outcomes may be of little practical interest, or because there may be prohibitively (even infinitely) many of them, outcomes are grouped into sets of outcomes that satisfy some condition, which are called "events." The collection of all such events is a sigma-algebra.

An event containing exactly one outcome is called an *elementary event*. The event that contains all possible outcomes of an experiment is its sample space. A single outcome can be a part of many different events.

Examples

1. An example of a simple random experiment (e.g. E_1) is to toss a die and observe the outcome. In this case, the sample space of E_1 is: $S_1 = \{1, 2, 3, 4, 5, 6\}$, and $S_1' = \{$"odd", "even"$\}$. This also shows that the sample space for a given experiment may not be unique.
2. The sample space for tossing a coin is (assuming "H = Heads" and "T = Tails"): $S_2 = \{HHH, HHT, HTH, HTT, THH, THT, TTH, TTT\}$
3. See how long a bulb lasts: $S_3 = \{\, t \,|t \geq 0\}$.

Permutation

In mathematics, a permutation of a set is, loosely speaking, an arrangement of its members into a sequence or linear order, or if the set is already ordered, a rearrangement of its elements. The word "permutation" also refers to the act or process of changing the linear order of an ordered set.

Permutations differ from combinations, which are selections of some members of a set regardless of order. For example, written as tuples, there are six permutations of the set $\{1, 2, 3\}$, namely: $(1, 2, 3)$, $(1, 3, 2)$, $(2, 1, 3)$, $(2, 3, 1)$, $(3, 1, 2)$, and $(3, 2, 1)$. These are all the possible orderings of this three-element set. Anagrams of words whose letters are different are also permutations: the letters are already ordered in the original word, and the anagram is a reordering of the letters. The study of permutations of finite sets is an important topic in the fields of combinatorics and group theory.

Permutations are used in almost every branch of mathematics, and in many other fields of science. In computer science, they are used for analyzing sorting algorithms; in quantum physics, for describing states of particles; and in biology, for describing RNA sequences.

The number of permutations of n distinct objects is n factorial, usually written as $n!$, which means the product of all positive integers less than or equal to n.

Examples

TBD

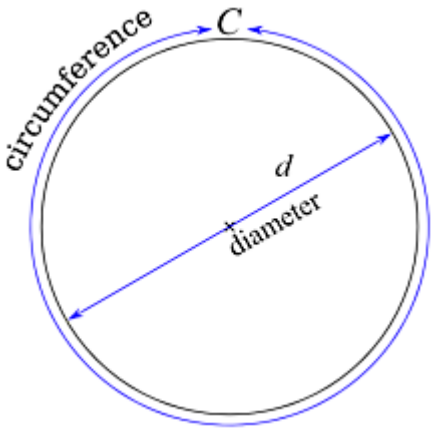
Pi

Symbol

π

The number π is a mathematical constant. π is commonly defined as the ratio of a circle's circumference C to its diameter d:

$$\pi = \frac{C}{d}$$



π appears in many formulas in all areas of mathematics and physics. Its value is approximately equal to 3.141592653589793. It has been represented by the Greek letter π since the mid-18th century, and it is also referred to as "Archimedes' constant".

Being an **irrational number**, π cannot be expressed as a common fraction. Although fractions such as $22/7$ and $355/113$ are commonly used as an approximation, no common fraction (ratio of whole numbers) can be its exact value.

Equivalently, π decimal representation never ends and never settles into a permanently repeating pattern. Its decimal (or other base) digits appear to be randomly distributed, and are conjectured to satisfy a specific kind of statistical randomness.

It is known that π is also a *transcendental number*, meaning, it is not the root of any polynomial with rational coefficients. The transcendence of π implies that it is impossible to solve the ancient challenge of squaring the circle with a compass and straightedge.

Examples

TBD

Point Slope Form

Formula

$$y - y_1 = m(x - x_1)$$

The point slope form is an equation of a straight line. In this formula, y is the unknown y coordinate, and y_1 is the given y coordinate, which is to be placed in the formula.

Similarly, x is the unknown x coordinate, and x_1 is the given x coordinate, which is to be used in the formula.

The variable m represents the slope of the straight line.

We can use the point slope formula when we have the coordinates of one point on the line and the slope of the line.

Examples

Find the equation of the straight line that has slope $m = 4$ and passes through the point $(-1, -6)$:

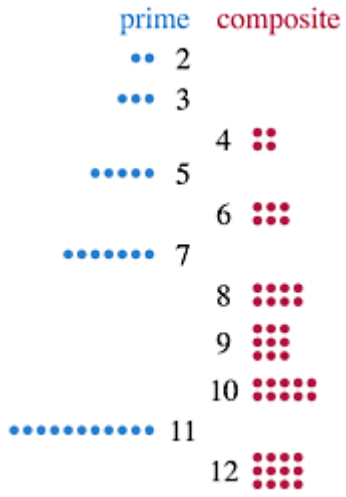
$y - y_1 = m(x - x_1)$
$y - (-6) = (4)(x - (-1))$
$y + 6 = 4(x + 1)$
$y + 6 = 4x + 4$
$y = 4x + 4 - 6$
$y = 4x - 2$

Prime Number

A prime number (or a prime) is a natural number greater than 1 that is not a product of two smaller natural numbers. A natural number greater than 1 that is not prime is called a composite number. For example, 5 is prime because the only ways of writing it as a product, $1 \cdot 5$ or $5 \cdot 1$, involve 5 itself. However, 4 is composite because it is a product $(2 \cdot 2)$ in which both numbers are smaller than 4.

Primes are central in number theory because of the fundamental theorem of arithmetic: every natural number greater than 1 is either a prime itself or can be factorized as a product of primes that is unique up to their order.

Composite numbers can be arranged into rectangles, but prime numbers cannot:



There are infinitely many primes, as demonstrated by Euclid around 300 BC. No known simple formula separates prime numbers from composite numbers. However, the distribution of primes within the natural numbers in the large can be statistically modelled. The first result in that direction is the prime number theorem, proven at the end of the 19th century, which says that the probability of a randomly chosen number being prime is inversely proportional to its number of digits, that is, to its logarithm.

Examples

The first 25 prime numbers (which are all the prime numbers less than 100):

$2, 3, 5, 7, 11, 13, 17, 19, 23, 29, 31, 37, 41, 43, 47, 53, 59, 61, 67, 71, 73, 79, 83, 89$ and 97.

Probability

Symbol

$P(A)$ or $p(A)$ or $Pr(A)$

Probability is the branch of mathematics concerning numerical descriptions of how likely an event is to occur, or how likely it is that a proposition is true. Probability describes random variation in systems. The probability of an event is a number between 0 and 1, where, roughly speaking, 0 indicates impossibility of the event and 1 indicates certainty.

The higher the probability of an event, the more likely it is that the event will occur. A simple example is the tossing of a fair (unbiased) coin. Since the coin is fair, the two outcomes ("heads" and "tails") are both equally probable; the probability of "heads" equals the probability of "tails"; and since no other outcomes are possible, the probability of either "heads" or "tails" is $1/2$ (which could also be written as 0.5 or 50%).

A joint probability is the probability that two separate events with separate probability distributions are both true. $P(A \text{ and } B)$ is written $P(A, B)$, and read “the joint probability of A and B” or “the probability that A is true and B is true.”

The *opposite* or *complement* of an event A is the event $\sim A$ (that is, the event of A not occurring), often denoted as A', A^c, $\overline{A}$, $A^\complement$, $\neg A$, or $\sim A$. Its probability is given by $P(\sim A) = 1 - P(A)$.

Probability theory is applied in everyday life in risk assessment and modeling. The insurance industry and markets use actuarial science to determine pricing and make trading decisions. Governments apply probabilistic methods in environmental regulation, entitlement analysis (reliability theory of aging and longevity), and financial regulation.

Axioms

1. $0 \leq P(A) \leq 1$: a probability value is always between 0 and 1.
2. $P(S) = 1$: the probability of having an outcome is 1.
3. If $A \cap B = \emptyset$ then $P(A \cup B) = P(A) + P(B)$.

Theorems

1. $P(\emptyset) = 0$.
2. $P(A' = 1 - P(A)$.
3. $P(A \cup B) = P(A) + P(B) - P(A \cap B)$.
4. For any events A, B, C:
   $P(A \cup B \cup C) = P(A) + P(B) + P(C) - P(A \cap B) - P(A \cap C) - P(B \cap C) + P(A \cap B \cap$

.

5. $A \leq B = P(A) \leq P(B)$.

Examples

1. Find the probability that the event of heads comes up after tossing a coin. This sample space is $S = \{H, T\}$, therefore, $P(H) = \frac{1}{2} = 50\%$ probability.

2. Find the probability of raining tomorrow, assuming:

   - 40% chance of cold (C).
   - 10% chance of cold and rain (C and R).
   - 80% chance of cold or rain (C or R, or both)

   So, $P(R) = P(C \cup R) - P(C) + P(C \cap R).

   Then, 0.8 - 0.4 + 0.1 = 0.5 = 50$% probability.

3. Find the probability that any person likes at least one of the three following activities, assuming:

   - 75% people like jogging (J).
   - 20% like ice cream (I).
   - 40% enjoy music (M).
   - 15% $J \cap I$.
   - 30% $J \cap M$.
   - 10% $I \cap M$.
   - 5% $J \cap I \cap M$.

   So,
   $P(J \cup I \cup M) = P(J) + P(I) + P(M) - P(J \cap I) - P(J \cap M) - P(I \cap M) + P(J \cap I \cap M$
   .

   Then, $0.75 + 0.20 + 0.40 - 0.15 - 0.30 - 0.10 + 0.05 = 0.85 = 85\%$ probability.

4. Based on previous example assumptions, find the probability of one person to like only one of the three activities (e.g. to like jogging, but not ice cream or music).

   So, $P(J \cap \overline{I} \cap \overline{M}) + P(\overline{J} \cap I \cap \overline{M}) + P(\overline{J} \cap \overline{I} \cap M) = 0.35 + 0.0 + 0.05 = 0.40$.

   This result is better illustrated using a Venn diagram:

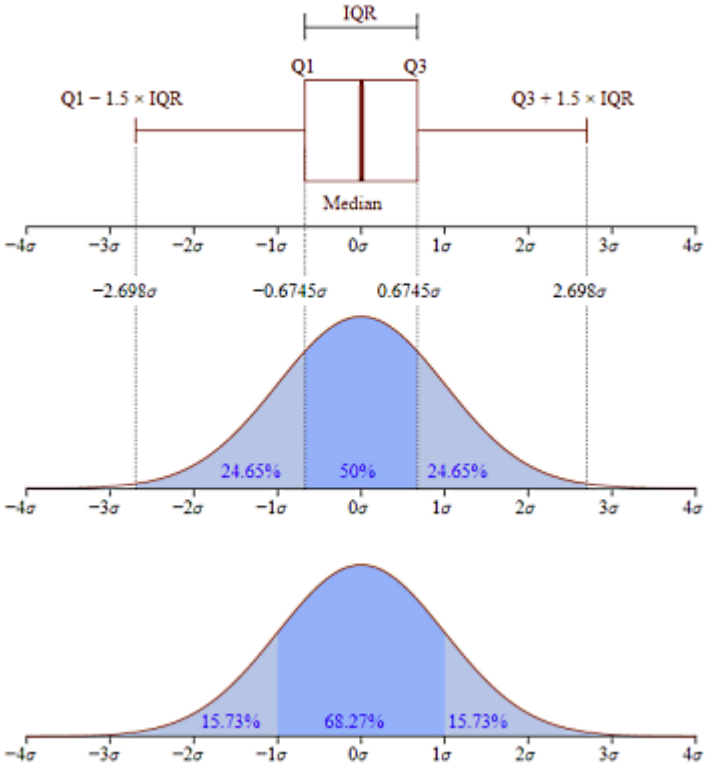
Note: The "principle of inclusion-exclusion" generalizes this case to n events, e.g. A, B, C, D, E.

Probability Density Function

In probability theory, a probability density function (PDF), or density of a continuous random variable, is a function whose value at any given sample (or point) in the sample space (the set of possible values taken by the random variable) can be interpreted as providing a relative likelihood that the value of the random variable would equal that sample.

In other words, while the absolute likelihood for a continuous random variable to take on any particular value is 0 (since there are an infinite set of possible values to begin with), the value of the PDF at two different samples can be used to infer, in any particular draw of the random variable, how much more likely it is that the random variable would equal one sample compared to the other sample.

In a more precise sense, the PDF is used to specify the probability of the random variable falling within a particular range of values, as opposed to taking on any one value. This probability is given by the integral of this variable's PDF over that range—that is, it is given by the area under the density function but above the horizontal axis and between the lowest and greatest values of the range. The probability density function is non-negative everywhere, and its integral over the entire space is equal to 1.



Examples

TBD

Probability Distribution

In probability theory and statistics, a probability distribution is the mathematical function that gives the probabilities of occurrence of different possible outcomes for an experiment. It is a mathematical description of a random phenomenon in terms of its sample space and the probabilities of events (subsets of the sample space).

What defines a probability distribution is a collection of statements, two or more, where those statements are exclusive and exahustive. Exclusive means that no more than one statement can be true. Exhaustive means that, assuming we have complete information, at least one statement must be true.

For instance, if X is used to denote the outcome of a coin toss ("the experiment"), then the probability distribution of X would take the value 0.5 for X=heads, and 0.5 for X=tails (assuming that the coin is fair).

Examples of random phenomena include the weather condition in a future date, the height of a person, the fraction of male students in a school, the results of a survey, etc.

A probability distribution can be described in various forms, such as by a probability mass function or a cumulative distribution function. One of the most general descriptions, which applies for continuous and discrete variables, is by means of a probability function $P : \mathcal{A} \to \mathbb{R}$ whose input space $\mathcal{A}$ is related to the sample space, and gives a probability as its output.

Examples

Complement Rule

$P(A) = 1 - P(A^c)$

Addition Rule for Mutually Exclusive Events

$P(A \text{ or } B) = P(A) + P(B)$

Multiplication Rule for Independent Events

We can calculate the probability of two events from independent distributions occuring at the same time. The probability independence definition states that when the joint distribution equals the product distribution, the two distributions are independent.

So, for example, considering the probability of getting heads from throwing a fair coin to be $1/2$, and the probability of getting a 3 when throwing a fair dice to be $1/3$, and as those two events belong to independent distributions, we can calculate their joint probability using the following formula:

$$P(x_1, y_1) = P(x_1) \cdot P(y_1) = \left(\frac{1}{2}\right) \cdot \left(\frac{1}{6}\right) = \frac{1}{12}$$

At Least One Rule

$P(\text{at least one}) = 1 - P(\text{none})$

Probability Space

In probability theory, a probability space or a probability triple $(\Omega, \mathcal{F}, P)$ is a mathematical construct that provides a formal model of a random process or "experiment". For example, one can define a probability space which models the throwing of a die.

A probability space consists of three elements:

- A sample space, Ω, which is the set of all possible outcomes.
- An event space, which is a set of events, $\mathcal{F}$, an event being a set of outcomes in the sample space.
- A probability function, which assigns each event in the event space a probability, which is a number between 0 and 1.

Examples

TBD

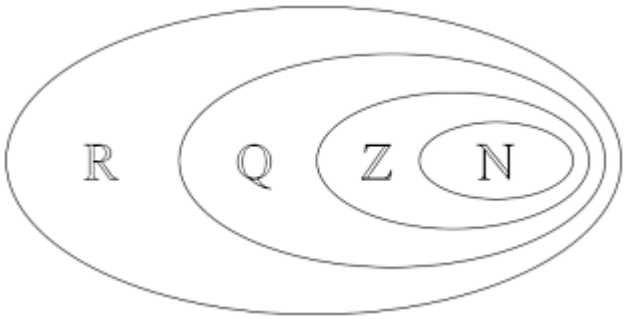
Real Numbers

Symbol

$\mathbb{R}$

A real number is a value of a continuous quantity that can represent a distance along a line (or alternatively, a quantity that can be represented as an infinite decimal expansion).

The real numbers ($\mathbb{R}$) include the rational numbers ($\mathbb{Q}$), which themselves include the integers ($\mathbb{Z}$), which in turn include the natural numbers ($\mathbb{N}$).



Therefore, the real numbers include all the rational numbers, such as the integer -5 and the fraction $4/3$, and all the irrational numbers, such as $\sqrt{2}$ (the square root of 2). Included within the irrationals are the transcendental numbers, such as π.

The adjective real in this context was introduced in the 17$^{\text{th}}$ century by René Descartes, who distinguished between real and imaginary roots of polynomials.

In addition to measuring distance, real numbers can be used to measure quantities such as time, mass, energy, velocity, and many more.

$\mathbb{R}^n$ is a coordinate space over the real numbers. This means that it is the set of the n-tuples of real numbers (sequences of n real numbers). With component-wise addition and scalar multiplication, it is a real vector space.

Typically, the Cartesian coordinates of the elements of a Euclidean space form a real coordinate space. This explains the name of coordinate space and the fact that geometric terms are often used when working with coordinate spaces.

For example, $\mathbb{R}^2$ is a plane (a two-dimensional flat surface that extends infinitely far):

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

Coordinate spaces are widely used in geometry and physics, as their elements allow locating points in Euclidean spaces, and computing with them.

Examples

The number 2 is an element of the Real Numbers set:

$2 \in \mathbb{R}$

Sensitivity and Specificity

Sensitivity and specificity are statistical measures of the performance of a binary classification test that are widely used in medicine. The terms were introduced by American biostatistician Jacob Yerushalmy in 1947:

- Sensitivity measures the proportion of positives that are correctly identified (e.g., the percentage of sick people who are correctly identified as having some illness).
- Specificity measures the proportion of negatives that are correctly identified (e.g., the percentage of healthy people who are correctly identified as not having some illness).

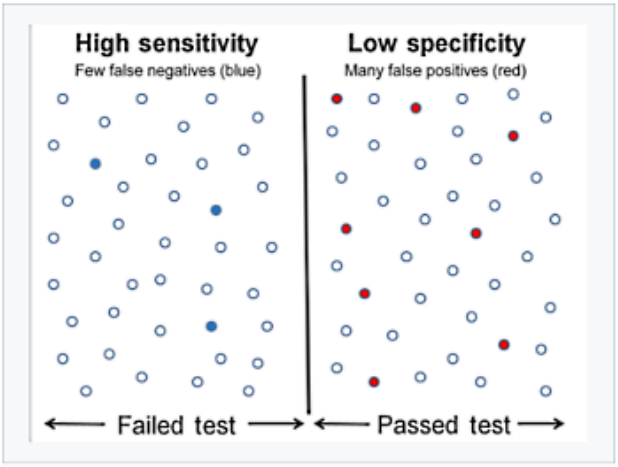
The terms *positive* and *negative* do not refer to benefit, but to the presence or absence of a condition. For example, if the condition is a disease, "positive" means "diseased" and "negative" means "healthy".

In many tests, including diagnostic medical tests, sensitivity is the extent to which true positives are not overlooked, thus false negatives are few, and specificity is the extent to which true negatives are classified as such, thus false positives are few. A sensitive test rarely overlooks a true positive (for example, showing nothing wrong despite a problem existing); a specific test rarely registers a positive classification for anything that is not the target of testing (for example, finding one bacterial species and mistaking it for another closely related one that is the true target).

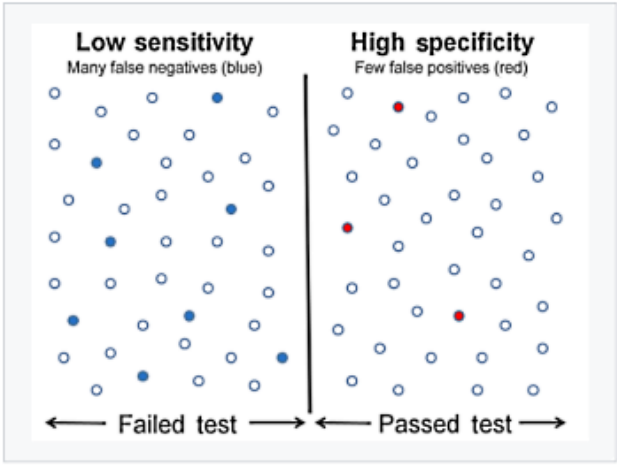
A perfect predictor would be 100% sensitive, meaning all sick individuals are correctly identified as sick, and 100% specific, meaning no healthy individuals are incorrectly identified as sick.

There is usually a trade-off between measures. For instance, in airport security, since testing of passengers is for potential threats to safety, scanners may be set to trigger alarms on low-risk items like belt buckles and keys (low specificity) in order to increase the probability of identifying dangerous objects and minimize the risk of missing objects that do pose a threat (high sensitivity).

Examples



High sensitivity and low specificity



Low sensitivity and high specificity

Set Theory

A set is a collection of objects. Each object in the set is an element of the set. The order of the elements in a set does not matter.

The cardinality of A, sometimes mentioned as the absolute value of A, is noted as $|A|$, and refers to the number of elements in A.

The complement of set A, is expressed as $\overline{A}$, or A^c, and it refers to all the elements that are not elements of A.

Examples

Set A contains all integer numbers from 1 to 10:

$A = \{1, 2, 3, 4, 5, 6, 7, 8, 9, 10\}$

Number 2 is an element of set A:

$2 \in A$

Number 11.4 is not an element of set A:

$11.4 \notin A$

Set B contains two sport modalities:

$B = \{\text{"baseball", "football"}\}$

Set C refers to the interval in the real numbers line that contains all the numbers between 0 and 1 ("$|$" or ":" means "such that"):

$C = \{\, x \,|\, 0 \leq x \leq 1 \,\}$

Scientific Notation

Formula

$$m \cdot 10^n$$

In scientific notation, all numbers are written in the form "m times ten raised to the power of n", where the exponent n is an integer, and the coefficient m is any real number.

The scientific notation is a way of expressing real numbers that are too large or too small to be conveniently written in decimal form. It may also be referred to as "scientific form" or "standard index form", or "standard form" in the UK.

It is a base ten notation, commonly used by scientists, mathematicians, and engineers, in part because it can simplify certain arithmetic operations.

Examples

The rule for the conversion from decimal to scientific notation is to keep all the significant digits, and then to always leave one digit to the left of the decimal:

Decimal notation	Scientific notation
2	2×10^{0}
300	3×10^{2}
4 321.768	$4.321\,768 \times 10^{3}$
−53 000	-5.3×10^{4}
6 720 000 000	6.72×10^{9}
0.2	2×10^{-1}
987	9.87×10^{2}
0.000 000 007 51	7.51×10^{-9}

The Earth's mass value (in Kg):

$$5,972,000,000,000,000,000,000,000 = 5.972 \cdot 10^{24}$$

An electron's mass value (in Kg):

$$0.00000000000000000000000000009109 = 9.109 \cdot 10^{-31}$$

In [ ]:
```
earth_mass = 5972000000000000000000000
electron_mass = 0.00000000000000000000000000009109

# Using format() function:
print(format(earth_mass, ".3E"))
print(format(electron_mass, ".3E"))

# Alternative format() syntax:
print("{:.3E}".format(earth_mass))
print("{:.3E}".format(electron_mass))

# Using f-strings:
print(f"{earth_mass:.3E}")
print(f"{electron_mass:.3E}")
```

Slope

Formula

$$m = \frac{\Delta y}{\Delta x} = \frac{y_2 - y_1}{x_2 - x_1}$$

In mathematics, the slope or gradient of a line is a number that describes both the direction and the steepness of the line. Slope is often denoted by the letter m.

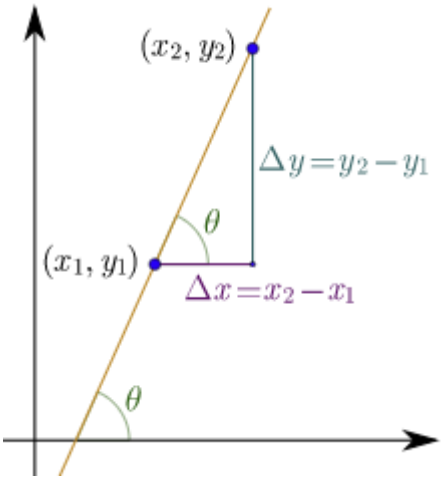
The slope is calculated by finding the ratio of the "vertical change" to the "horizontal change" between (any) two distinct points on a line. Sometimes the ratio is expressed as a quotient ("rise over run"), giving the same number for every two distinct points on the same line. A line that is

decreasing has a negative "rise". The line may be practical - as set by a road surveyor, or in a diagram that models a road or a roof either as a description or as a plan.

The steepness, incline, or grade of a line is measured by the absolute value of the slope. A slope with a greater absolute value indicates a steeper line.

The direction of a line is either increasing, decreasing, horizontal or vertical:

- A line is increasing if it goes up from left to right. In this case, the slope is positive: $m > 0$.
- A line is decreasing if it goes down from left to right. In this case, the slope is negative: $m < 0$.
- If a line is horizontal, the slope is zero. This is a constant function.
- If a line is vertical, the slope is undefined.



Examples

Calculate the slope between point $A = (2, -1)$ and point $B = (4, 3)$:

$$m = \frac{B_y - A_y}{B_x - A_x} = \frac{3 - (-1)}{4 - 2} = \frac{4}{2} = 2$$

Slope Intercept Form

Formula

$$y = mx + b$$

The slope intercept form is used to find the equation of a straight line. The formula identifies the line slope, represented as m, and the y-intercept, represented as b, which has this name because it is where the line intercepts the y-axis.

The slope measures how steep the line is. It is the inclination, or gradient, of a line. If it is positive, the values of y increase with increasing x, and if it is negative, y decreases with an increasing x.

The y-intercept is the place where the line crosses the y axis, so we can also think of the y-intercept as the value of y when x equals zero.

In summary, we have:

- y: dependent variable
- m: slope of the line
- x: independent variable
- b: y-intercept

Examples

Least-Squares Regression

The most common method for fitting a regression line is the method of least-squares. This method calculates the best-fitting line for the observed data by minimizing the sum of the squares of the vertical deviations from each data point to the line (if a point lies on the fitted line exactly, then its vertical deviation is 0).

Because the deviations are first squared, then summed, there are no cancellations between positive and negative values.

To calculate the LSR, assuming we have x and y number sets, each one with n data points:

- **Step 1**: Calculate the slope (m):

$$m = \frac{n\sum xy - (\sum x)(\sum y)}{n\sum x^2 - (\sum x)^2}$$

- **Step 2**: Calculate the y-intercept (b):

$$b = \frac{(\sum y - m(\sum x))}{n}$$

- **Step 3**: Calculate the predicted value (y):

$$y = mx + b$$

Some of the things to consider when implementing the least-squares regression method:

- The data must be free of outliers because they might lead to a biased and wrongful line of best fit.
- The line of best fit can be drawn iteratively until we get a line with the minimum possible squares of errors.
- The method works well even with non-linear data.
- Technically, the difference between the actual value of y and the predicted value of y is called a "residual" (because it denotes the error).

Root Mean Squared Error

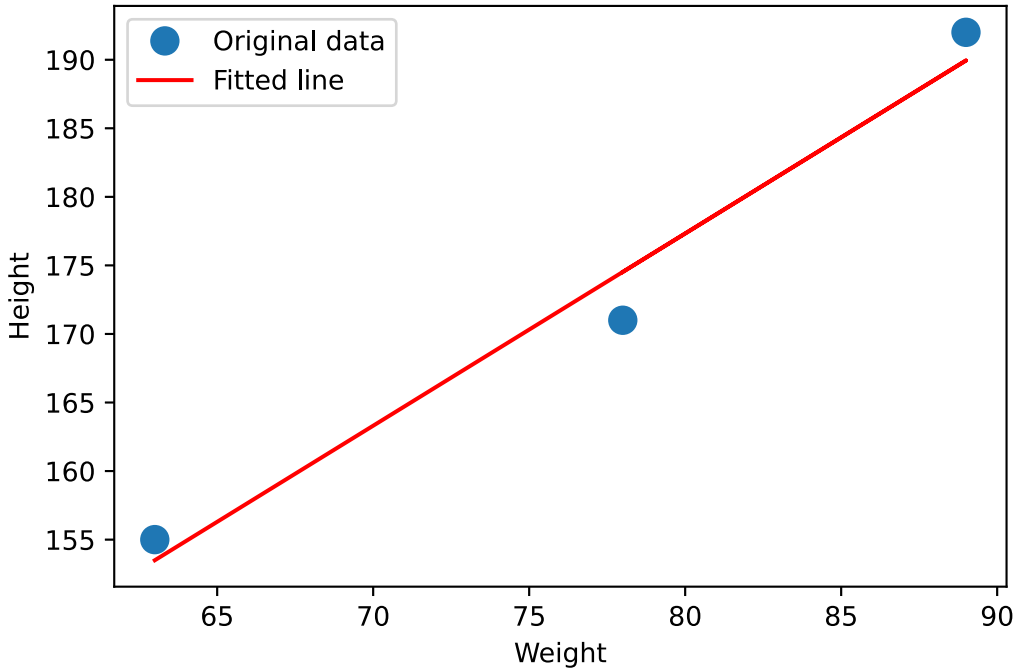
The Root Mean Squared Error (RMSE) allows evaluating the quality of a prediction model. It is calculated by getting the square root of the sum of all errors divided by the total number of values. $\hat{y}_i$ is the $_i$th predicted value:

$$RMSE = \sqrt{\sum_{i=1}^{n} \frac{1}{n}(\hat{y}_i - y_i)^2}$$

In [8]:

```python
# A regression model to predict people's height based on their weight

import numpy as np
import matplotlib.pyplot as plt

# Weight (Kg)
X = np.array([63, 89, 78])
# Height (cm)
Y = np.array([155, 192, 171])

N = len(X)

# Step 1:
# Calculate the slope using Least-Squares Regression
slope = ((N * np.dot(X, Y)) - np.sum(X) * np.sum(Y)) / \
    ((N * np.sum(X ** 2)) - (np.sum(X) ** 2))
print(f"slope: {slope}")

# Step 2:
# Calculate the y-intercept
y_intercept = (np.sum(Y) - (slope * np.sum(X))) / N
print(f"y_intercept: {y_intercept}")

slope = round(slope, 3)
y_intercept = round(y_intercept, 3)

# Plot the regression line
plt.plot(X, Y, "o", label="Original data", markersize=10)
plt.plot(X, slope * X + y_intercept, "r", label="Fitted line")
plt.xlabel("Weight")
plt.ylabel("Height")

plt.legend()
plt.show()

# A sample input weight
x = 60

# Step 3:
# Predict a person's height based on his weight
y = slope * x + y_intercept
y = round(y, 2)

# The prediction for a person with 60Kg weight is 149.29cm height
print(f"The prediction for x = {x} is y = {y}")
```

```
slope: 1.4021526418786692
y_intercept: 65.1682974559687
```




```
The prediction for x = 60 is y = 149.29
```

Standard Deviation

Symbol

σ or s

In statistics, the standard deviation is a measure of the amount of variation or dispersion of a set of values. A low standard deviation indicates that the values tend to be close to the mean (also called the expected value) of the set, while a high standard deviation indicates that the values are spread out over a wider range.

Standard deviation may be abbreviated as **SD**. The σ or σ_x symbols are used to express a population's standard deviation. The s symbol is used to express the standard deviation of a sample of a population, which then can be referred as the *standard deviation of the sample* or *sample standard deviation*.

The standard deviation of a random variable, statistical population, data set, or probability distribution is the square root of its variance. It is algebraically simpler, though in practice less robust, than the average absolute deviation. A useful property of the standard deviation is that unlike the variance, it is expressed in the same unit as the data.

In addition to expressing the variability of a population, the standard deviation is commonly used to measure confidence in statistical conclusions. For example, the margin of error in polling data is determined by calculating the expected standard deviation in the results if the same poll were to be conducted multiple times.

This derivation of a standard deviation is often called the "standard error of the estimate", or "standard error of the mean" when referring to a mean. It is computed as the standard deviation of

all the means that would be computed from that population, if an infinite number of samples were drawn and a mean for each sample were computed.

To calculate standard deviation, add up all the data points and divide by the number of data points, calculate the variance for each data point and then find the square root of the variance.

Examples

To calculate a population's standard deviation, we divide by the total of elements in the dataset:

$$\sigma = \sqrt{\frac{\sum(X_i - \mu_x)^2}{n}}$$

To calculate the standard deviation of a sample of a population, we divide by the total of elements in the dataset minus one:

$$s = \sqrt{\frac{\sum(X_i - \bar{X})^2}{n-1}}$$

Summation

Symbol

$$\sum_{i=1}^{n} i$$

In mathematics, summation is the addition of a sequence of any kind of numbers, called addends or summands. The result is their sum or total. Beside numbers, other types of values can be summed as well: functions, vectors, matrices, polynomials and, in general, elements of any type of mathematical objects on which an operation denoted "+" is defined.

The mathematical notation uses an enlarged form of the upright capital Greek letter Sigma to compactly represent summation of many similar terms, naming it the "summation symbol".

The summation operation is defined as:

$$\sum_{i=m}^{n} a_i = a_m + a_{m+1} + a_{m+2} + \cdots + a_{n-1} + a_n$$

Where i is the index of summation, a_i is an indexed variable representing each term of the sum, m is the lower bound of summation, and n is the upper bound of summation. The "$i = m$" under the summation symbol means that the index i starts out equal to m. The index i is incremented by one for each successive term, stopping when $i = n$.

Examples

A sum of squares:

$$\sum_{i=3}^{6} i^2 = 3^2 + 4^2 + 5^2 + 6^2 = 86.$$

In [ ]:
```
import math

A = [2.12, 0.9, 1.98]

# Use math.fsum() instead of sum() to achieve
# higher precision with floating point numbers
math.fsum(A) == 5.0
```

Variables

In Statistics, depending on the context, an independent variable is sometimes called a "predictor variable", regressor, covariate, "manipulated variable", "explanatory variable", "exposure variable" (v. reliability theory), "risk factor" (v. medical statistics), "feature" (in machine learning and pattern recognition) or "input variable".

In econometrics, the term "control variable" is usually used instead of "covariate". From the Economics community, the independent variables are also called "exogenous".

Depending on the context, a dependent variable is sometimes called a "response variable", "regressand", "criterion", "predicted variable", "measured variable", "explained variable", "experimental variable", "responding variable", "outcome variable", "output variable", "target" or "label".

In economics, endogenous variables are usually referencing the target.

"Explanatory variable" is preferred by some authors over "independent variable" when the quantities treated as independent variables may not be statistically independent or independently manipulable by the researcher. If the independent variable is referred to as an "explanatory variable" then the term "response variable" is preferred by some authors for the dependent variable.

"Explained variable" is preferred by some authors over "dependent variable" when the quantities treated as "dependent variables" may not be statistically dependent. If the dependent variable is referred to as an "explained variable" then the term "predictor variable" is preferred by some authors for the independent variable.

Variables may also be referred to by their form: continuous or categorical, which in turn may be binary/dichotomous, nominal categorical, and ordinal categorical, among others.

A variable may be thought to alter the dependent or independent variables, but may not actually be the focus of the experiment. So that the variable will be kept constant or monitored to try to minimize its effect on the experiment. Such variables may be designated as either a "controlled variable", "control variable", or "fixed variable".

Extraneous variables, if included in a regression analysis as independent variables, may aid a researcher with accurate response parameter estimation, prediction, and goodness of fit, but are not of substantive interest to the hypothesis under examination. For example, in a study examining the effect of post-secondary education on lifetime earnings, some extraneous variables might be gender, ethnicity, social class, genetics, intelligence, age, and so forth.

A variable is extraneous only when it can be assumed (or shown) to influence the dependent variable. If included in a regression, it can improve the fit of the model. If it is excluded from the regression and if it has a non-zero covariance with one or more of the independent variables of interest, its omission will bias the regression's result for the effect of that independent variable of interest. This effect is called confounding or omitted variable bias; in these situations, design changes and/or controlling for a variable statistical control is necessary.

Extraneous variables are often classified into three types:

1. Subject variables, which are the characteristics of the individuals being studied that might affect their actions. These variables include age, gender, health status, mood, background, etc.
2. Blocking variables or experimental variables are characteristics of the persons conducting the experiment which might influence how a person behaves. Gender, the presence of racial discrimination, language, or other factors may qualify as such variables.
3. Situational variables are features of the environment in which the study or research was conducted, which have a bearing on the outcome of the experiment in a negative way. Included are the air temperature, level of activity, lighting, and the time of day.

In modelling, variability that is not covered by the independent variable is designated by e_I and is known as the "residual", "side effect", "error", "unexplained share", "residual variable", "disturbance", or "tolerance".

Variance

Symbol

σ^2 or s^2

In probability theory and statistics, variance is the expectation of the squared deviation of a random variable from its mean. Informally, it measures how far a set of numbers is spread out from their average value.

The variance, also known as "variation" or "sum of squares", has a central role in statistics, where some ideas that use it include descriptive statistics, statistical inference, hypothesis testing, goodness of fit, and Monte Carlo sampling.

Variance is also an important tool in the sciences, where statistical analysis of data is common. The variance is the square of the standard deviation, the second central moment of a distribution, and the covariance of the random variable with itself.

To express a population's variance, use the σ^2 symbol. To express the variance of a sample of a population, use the s^2 symbol.

The variance is the average of the squared differences from the mean. To figure out the variance, first calculate the difference between each point and the mean. Then, square and average the results.

Because of this squaring, the variance is no longer in the same unit of measurement as the original data. Taking the root of the variance means the standard deviation is restored to the original unit of measure and therefore much easier to interpret.

Examples

Calculate the variance in set Z:

$Z = \{6, 8, 10, 7, 8, 13, 9, 10\}$

Calculate the variance of a sample of the population:

$$\sigma_z^2 = \frac{1}{n-1} \sum_{i=1}^{n} (z_i - \mu_z)^2 = 4.696428571428571$$

Calculate the variance of the population: $\sigma_z^2 = \frac{\sum_{i=1}^{n} (z_i - \mu_z)^2}{n} = 4.109375$

In [ ]:
```python
import statistics

A = [6, 8, 10, 7, 8, 13, 9, 10]

# Variance of a sample of the population
sVar_A = statistics.variance(A)

# Output is 4.696428571428571
print(sVar_A)

# Variance of a population
pVar_A = statistics.pvariance(A)

# Output is 4.109375
print(pVar_A)
```

Venn Diagram

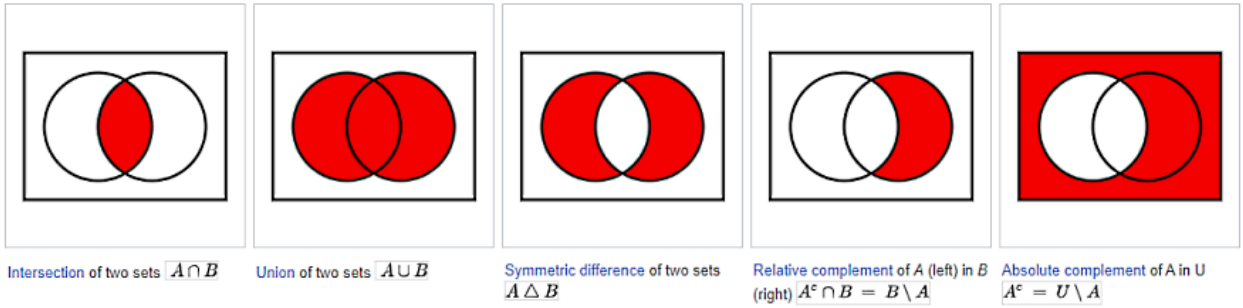
A Venn diagram, also called primary diagram, set diagram or logic diagram, is a diagram that shows all possible logical relations between a finite collection of different sets. These diagrams depict elements as points in the plane, and sets as regions inside closed curves.

A Venn diagram consists of multiple overlapping closed curves, usually circles, each representing a set. The points inside a curve labelled S represent elements of the set S, while points outside the boundary represent elements not in the set S. This lends itself to intuitive visualizations; for example, the set of all elements that are members of both sets S and T, denoted $S \cap T$ and read "the intersection of S and T", is represented visually by the area of overlap of the regions S and T.

In Venn diagrams, the curves are overlapped in every possible way, showing all possible relations between the sets. They are thus a special case of Euler diagrams, which do not necessarily show all relations. A Venn diagram in which the area of each shape is proportional to the number of elements it contains is called an area-proportional (or scaled Venn diagram).

Venn diagrams were conceived around 1880 by John Venn. They are used to teach elementary set theory, as well as illustrate simple set relationships in probability, logic, statistics, linguistics, and computer science.

Examples



Intersection of two sets $A \cap B$ | Union of two sets $A \cup B$ | Symmetric difference of two sets $A \triangle B$ | Relative complement of A (left) in B (right) $A^c \cap B = B \setminus A$ | Absolute complement of A in U $A^c = U \setminus A$

Complement Laws:

- $A \cup A^c = U$
- $A \cap A^c = \varnothing$
- $\varnothing^c = U$
- $U^c = \varnothing$
- If $A \subseteq B$, then $B^c \subseteq A^c$

Commutative Laws:

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

Associative Laws:

- $A \cup (B \cup C) = (A \cup B) \cup C = A \cup B \cup C$
- $A \cap (B \cap C) = (A \cap B) \cap C = A \cap B \cap C$

Distributive Laws:

- $A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$
- $A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$

De Morgan's Laws:

- $\overline{A \cup B} = \overline{A} \cap \overline{B}$
- $\overline{A \cap B} = \overline{A} \cup \overline{B}$
- $(A \cup B)^c = A^c \cap B^c$
- $(A \cap B)^c = A^c \cup B^c$

In [12]:
```
A = set({0, 2, 4, 6, 8})
B = set({1, 2, 3, 4, 5})

print("Intersection of two sets (using A & B):", A & B)
print("Intersection of two sets (using A.intersection(B)):",
```

```
    A.intersection(B))
print()
print("Union of two sets (using A | B:)", A | B)
print("Union of two sets (using A.union(B)):",
    A.union(B))
print()
print("Symmetric difference of two sets (using A ^ B):", A ^ B)
print("Symmetric difference of two sets (using A.symmetric_difference(B)):",
    A.symmetric_difference(B))
print()
print("Relative complement of A in B (using B - A):", B - A)
print("Relative complement of A in B (using B.difference(A)):",
    B.difference(A))
```

```
Intersection of two sets (using A & B): {2, 4}
Intersection of two sets (using A.intersection(B)): {2, 4}

Union of two sets (using A | B:) {0, 1, 2, 3, 4, 5, 6, 8}
Union of two sets (using A.union(B)): {0, 1, 2, 3, 4, 5, 6, 8}

Symmetric difference of two sets (using A ^ B): {0, 1, 3, 5, 6, 8}
Symmetric difference of two sets (using A.symmetric_difference(B)): {0, 1, 3, 5, 6, 8}

Relative complement of A in B (using B - A): {1, 3, 5}
Relative complement of A in B (using B.difference(A)): {1, 3, 5}
```

Y Hat

Symbol

$\hat{y}$

Y hat (written $\hat{y}$) is the predicted value of y (the dependent variable) in a regression equation. It can also be considered to be the average value of the response variable.

The regression equation is just the equation which models the data set. The equation is calculated during regression analysis. A simple linear regression equation can be written as:

$$\hat{y} = b_0 + b_1 x$$

Since b_0 and b_1 are constants defined by the analysis, finding $\hat{y}$ for any particular point simply involves plugging in the relevant value of x.

Example

Supposing we want to predict first grade reading abilities from the number of hours per week a child spends reading in preschool, we would have a set of data points: reading ability scores (assuming to use an unbiased scale) and survey data from homes which would tell us how many hours per day of preschool reading.

Having this information, we can use simple linear regression and the least squares method to find the regression equation that best would fit the data.

Assuming our line is:

$\hat{y} = 2.45x - 0.16$

Let's say $\hat{y}$ is the predicted average reading level for a child who has read $^1/_2$ an hour a day in preschool. To find this value, we just need to plug in $x = 0.5$:

$\hat{y} = 2.45(0.5) - 0.16 = 1.065$

The resulting regression line predicts that a child who reads to half an hour a day in preschool would have a 1.065 reading level in kindergarten (according to the above mentioned scale).