Mathematics Cheat Sheet

Friend

November 25, 2024

1 Numbers

1.1 Natural Numbers $\mathbb{N}$

The set of natural numbers represents the process of counting. Whether or not 0 is part of $\mathbb{N}$ depends on the definition and may vary. If 0 is not included, the set is defined as:

$$\mathbb{N} = \{1, 2, 3, 4, \dots, n, n+1, \dots\}. \tag{1.1}$$

How can we perform arithmetic with natural numbers? Addition and multiplication are unrestricted. We say that $\mathbb{N}$ is closed under addition and multiplication. Other operations, such as subtraction and division, are not universally applicable because negative numbers are not part of the natural numbers. A subset of $\mathbb{N}$ is the set of **prime numbers**, defined as:

$$\mathbb{P} = \{1, 2, 3, 5, 7, 11, 13, 17, 19, 23, 29, \dots\}. \tag{1.2}$$

Prime numbers are only divisible by 1 and themselves!

1.2 Integers $\mathbb{Z}$

The set of integers is obtained by extending the natural numbers to include negative numbers:

$$\mathbb{Z} = \{\dots, -3, -2, -1, 0, 1, 2, 3, \dots\}. \tag{1.3}$$

Now subtraction is also possible without restriction.

1.3 Rational and Irrational Numbers $\mathbb{Q}$, $\mathbb{R}\backslash\mathbb{Q}$

To perform unrestricted division, we need fractions:

- $$\mathbb{Q}_+ = \left\{ \frac{a}{b} \mid a, b \in \mathbb{N}, b \neq 0 \right\} \tag{1.4}$$

Including negative fractions, we get the set of rational numbers:

- $$\mathbb{Q} = \left\{ \frac{a}{b} \mid a, b \in \mathbb{Z}, b \neq 0 \right\} \tag{1.5}$$

- In $\mathbb{Q}$, all basic arithmetic operations are allowed.
- $\mathbb{Q}$ includes all positive and negative fractions, as well as all terminating decimal fractions (e.g., -3.75) and repeating decimal fractions (e.g., 0.6666...).

One operation is not fully allowed within the rational numbers: taking square roots, since it can lead to infinite numbers that cannot be expressed as fractions. These numbers are called **irrational numbers**, e.g.,

$$\sqrt{2} = 1.41421356\ldots \tag{1.6}$$

1.4 Real Numbers $\mathbb{R}$

By combining the rational and irrational numbers, we get the real numbers $\mathbb{R}$. However, taking square roots of negative numbers is not defined. For example,

$$\sqrt{-4} \tag{1.7}$$

is not defined, and such numbers are not included in $\mathbb{R}$.

1.5 Complex Numbers $\mathbb{C}$

A complex number z is represented as a pair of real numbers:

$$x + iy \mid x, y \in \mathbb{R}, \quad i = \sqrt{-1}. \tag{1.8}$$

The important feature of the imaginary unit i is that it allows us to take the square root of negative numbers. A complex number $z \in \mathbb{C}$ can also be written as the pair (x, y), where x is the real part and y is the imaginary part. Thus, the set of complex numbers $\mathbb{C}$ can be geometrically represented as pairs of real numbers (x, y) on the complex plane (also called the Gaussian plane), as shown in the figure below.
The addition of two complex numbers is defined as:

$$z_1 + z_2 = (x_1 + x_2) + (y_1 + y_2) \cdot i \tag{1.9}$$

and the subtraction as:

$$z_1 - z_2 = (x_1 - x_2) + (y_1 - y_2) \cdot i. \tag{1.10}$$

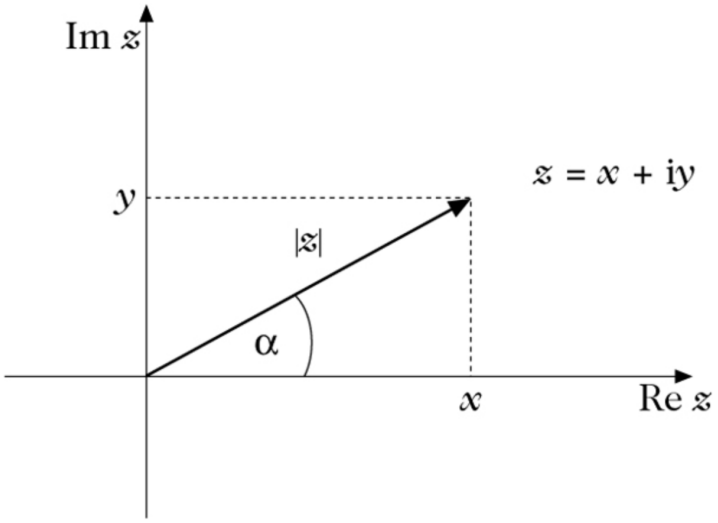

Figure 1.1: Gaussian Plane; $c \in \mathbb{C}$ as a real number pair (x, y)

Multiplication is defined as:

$$z_1 \cdot z_2 = (x_1 + x_2) \cdot (y_1 - y_2) \cdot i = x_1 y_1 + x_1 y_2 \cdot i + y_1 x_2 \cdot i + x_2 y_2 \cdot i^2, \tag{1.11}$$

which simplifies to:

$$z_1 \cdot z_2 = (x_1 y_1 - x_2 y_2) + (x_1 y_2 + y_1 x_2) \cdot i. \tag{1.12}$$

If z is a complex number, then z^* is its complex conjugate. In the representation below, the real part of z is reflected. In particular, we have

$$i^* = -i, \quad z^* = x - y \cdot i. \tag{1.13}$$

Multiplying by the complex conjugate gives the magnitude of z:

$$|z| = \sqrt{x^2 + y^2}. \tag{1.14}$$

The division of complex numbers is defined as:

$$\frac{z_1}{z_2} = \frac{z_1}{z_2} \cdot \frac{z_2^*}{z_2^*}. \tag{1.15}$$

This operation is a bit cumbersome and can be avoided when possible. If it cannot be avoided, multiply the numerator and denominator separately and simplify using the definition of i.
A different representation is possible using polar coordinates:

$$z = a + i \cdot b \quad a, b, r \in \mathbb{R}, \theta \in [0, 2\pi] | \Leftrightarrow z = r \cdot (\cos(\theta) + i \cdot \sin(\theta)) \Leftrightarrow r \cdot e^{i\theta}. \tag{1.16}$$

1.5.1 Special Numbers

π - The Circle Constant

3.1415926535... is the irrational number π[1]. Pi describes the ratio of the circumference to the diameter of a circle. Many formulas involve π:

$$\text{Circumference} \quad U = \pi \cdot d = 2 \cdot \pi \cdot r, \tag{1.17}$$

$$\text{Area} \quad A = \pi \cdot r^2, \tag{1.18}$$

$$\text{Volume} \quad V = \frac{4}{3} \cdot \pi \cdot r^3. \tag{1.19}$$

e - Euler's Number

Euler's number is the base of the natural logarithm $\ln(e) = 1$. Euler's number can be approximated as

$$e = 2.71828, \tag{1.20}$$

but like π, it does not have an exact solution. Named after the Swiss mathematician and physicist Leonhard Euler (1707-1783), e is crucial for exponential functions.

[1] Pi's digits omitted for brevity.

2 Fundamentals & Arithmetic Laws

A binary operation can be defined as a way in which two objects determine a third. The operation is abstractly expressed with '$\circ$'. The **law of closure states** that the result of an operation on two elements of a set is also an element of that set. That allows to define operations such as *addition* and *multiplication*. Depending on the **algebraic structure**, operations differ in outcome (e.g. $1+1 \neq 2 \in \mathbb{F}_2 = 0, 1$). Although operations may yield different outcomes, axioms are fundamental and true for most structures. *Note:* Having to prove a certain law holds for a specific case is often required during exercises in analysis.

2.1 Axioms

Axioms establish that algebraic structures have operations (addition and multiplication), and that these operations behave in specific, predictable ways. The most fundamental laws that apply to **most** structures are: See this article for more information.

2.1.1 Commutative Law

The order of the operation does not matter, e.g.,

$$2 \circ 3 = 3 \circ 2 \tag{2.1}$$

..

2.1.2 Associative Law

The grouping of three numbers does not affect the result of the operation, e.g.,

$$(2 \circ 3) \circ 4 = 2 \circ (3 \circ 4). \tag{2.2}$$

2.1.3 Distributive Law

The handling of parentheses depends on the number set and the type of operation. For addition and multiplication in the set of real numbers $\mathbb{R}$, both operations are distributive. Thus,

$$2 \odot (3 \oplus 4) = 2 \odot 3 \oplus 2 \odot 4. \tag{2.3}$$

2.1.4 Inequality Laws

Inequalities change depending on the operation:

- Adding/subtracting a constant:

$$a > b \Rightarrow a + c > b + c \tag{2.4}$$

- Multiplying by a positive constant:

$$a > b \Rightarrow a \cdot c > b \cdot c \tag{2.5}$$

- Multiplying by a negative constant reverses the inequality:

$$a > b \Rightarrow a \cdot (-c) < b \cdot (-c) \tag{2.6}$$

2.1.5 Identity Laws

These laws describe the neutral element in an operation:

- For addition:

$$a + 0 = a \tag{2.7}$$

- For multiplication:

$$a \cdot 1 = a \tag{2.8}$$

2.1.6 Inverse Laws

These laws describe how to reverse an operation:

- For addition:

$$a + (-a) = 0 \tag{2.9}$$

- For multiplication:

$$a \cdot a^{-1} = 1, \tag{2.10}$$

  where $a{-}1 = \frac{1}{a}$ and $a \neq 0$

2.1.7 Zero Laws

The number zero has special properties:

$$a \cdot 0 = 0 \tag{2.11}$$

Division by zero is undefined.

2.1.8 Absolute Value Properties

The absolute value represents the distance from zero:

- $$|a| \geq 0 \tag{2.12}$$

- $$|a| = a \tag{2.13}$$

- $$|a \cdot b| = |a| \cdot |b|. \tag{2.14}$$

2.2 The Binomial Formulas

(a)
$$(a+b)^2 = a^2 + 2ab + b^2 \tag{2.15}$$

(b)
$$(a-b)^2 = a^2 - 2ab + b^2 \tag{2.16}$$

(c)
$$(a+b) \cdot (a-b) = a^2 - b^2 \tag{2.17}$$

2.3 Set Operations

In set operations, the following laws hold:

- Commutative:

$$A \cup B = B \cup A \tag{2.18}$$

$$A \cap B = B \cap A \tag{2.19}$$

- Associative:

$$(A \cup B) \cup C = A \cup (B \cup C) \tag{2.20}$$

$$(A \cap B) \cap C = A \cap (B \cap C) \tag{2.21}$$

- Distributive:

$$A \cap (B \cup C) = (A \cap B) \cup (A \cap C) \tag{2.22}$$

2.4 Powers

A power is a shorthand notation for repeated multiplication by itself.

- $$a^0 = 1 \tag{2.23}$$

- $$a^1 = a \tag{2.24}$$

- $$a^{-1} = \frac{1}{a} \tag{2.25}$$

- $$a^{-n} = \frac{1}{a^n} \tag{2.26}$$
- $$a^{n} = \frac{1}{a^{-n}} \tag{2.27}$$
- $$a^p \cdot a^q = a^{p+q} \tag{2.28}$$
- $$a^p : a^q = a^{p-q} \tag{2.29}$$
- $$a^q \cdot b^q = (a \cdot b)^q \tag{2.30}$$
- $$a^q : b^q = (a : b)^q \tag{2.31}$$
- $$(a^p)^q = a^{p \cdot q} \tag{2.32}$$
- $$\frac{a^m}{a^n} = a^{m-n} \tag{2.33}$$
- $$\frac{a^n}{b^n} = \left(\frac{a}{b}\right)^n \tag{2.34}$$

- $$\left(\frac{a}{b}\right)^{-n} = \left(\frac{b}{a}\right)^{n} \tag{2.35}$$

Building on the basic exponent rules, we have:

- $$(a \cdot b)^n = a^n \cdot b^n \tag{2.36}$$

- $$\left(\frac{a}{b}\right)^{n} = \frac{a^n}{b^n} \tag{2.37}$$

2.5 Roots

The root of a number, when multiplied by itself, gives the number. By default, this refers to square roots, but higher roots (e.g., cubic roots $\sqrt[3]{x}$) are also possible. In terms of powers, the square root is expressed as

$$\sqrt{x} = x^{\frac{1}{2}} \tag{2.38}$$

and

$$\sqrt[n]{x} = x^{\frac{1}{n}}, \tag{2.39}$$

e.g.

$$\sqrt[3]{125} = 125^{\frac{1}{3}}. \tag{2.40}$$

If a power has a solution, then

$$x^n = a \Leftrightarrow x = \sqrt[n]{a} \tag{2.41}$$

as in

$$3^4 = 81 \equiv \sqrt[4]{81} = 3. \tag{2.42}$$

Roots also follow these additional rules:

- Nested roots:

$$\sqrt[m]{\sqrt[n]{a}} = \sqrt[m \cdot n]{a} \tag{2.43}$$

- For products:

$$\sqrt[m]{a \cdot b} = \sqrt[m]{a} \cdot \sqrt[m]{b} \tag{2.44}$$

Note: The nth root is the inverse function of the power function x^n.

2.6 Logarithms

Question: What number must I raise a to, to get y? Written as

$$\log_a(x) = y. \tag{2.45}$$

Note: The logarithms of zero and negative numbers are not defined!

The logarithm is the inverse function of the exponential function:

$$f(x) = a^x = y, \quad f^{-1}(y) = \log_a(y) = x. \tag{2.46}$$

Thus, the logarithm provides the exponent of the exponential function to the base a. For the exponential function $f(x) = e^x$ with e as the base, the natural logarithm (ln) is defined as

$$f^{-1} \tag{2.47}$$

.

Logarithms have specific rules:

- $$\log_a(1) = 0 \tag{2.48}$$

- $$\log_a(a) = 1 \tag{2.49}$$

- $$\log_a(p \cdot q) = \log_a(p) + \log_a(q) \tag{2.50}$$

- $$\log_a\left(\frac{p}{q}\right) = \log_a(p) - \log_a(q) \tag{2.51}$$

- $$\log_a(p^q) = q \cdot \log_a(p) \tag{2.52}$$

- $$\log_a\left(\sqrt[n]{p}\right) = \frac{\log_a(p)}{n} \tag{2.53}$$

- $$\log_a(p) = \frac{\log_b(p)}{\log_b(a)} \tag{2.54}$$

These additional rules expand on logarithmic behavior:

- Change of base:

$$\log_b(a) = \frac{\ln(a)}{\ln(b)} \tag{2.55}$$

- Reciprocal property:

$$\log_a\left(\frac{1}{b}\right) = -\log_a(b) \tag{2.56}$$

Logarithmic scaling is helpful when data varies significantly or when relative differences between values are important. Logarithmic scaling makes patterns easier to discern.

3 Probability

3.1 Introduction

3.1.1 Dependent and Independent Variables

The independent variable is the factor that a researcher manipulates, while the dependent variable is the outcome that is measured to see the effect of the independent variable.
In an experiment, the independent variable is the variable that is changed or controlled to test the effects on the dependent variable. The dependent variable is the variable being tested and measured. For example, in a study to determine the effect of different amounts of sunlight on plant growth, the amount of sunlight is the independent variable, and the growth of the plant is the dependent variable.
Mathematically, if X is the independent variable and Y is the dependent variable, the relationship can be expressed as:

$$Y = f(X), \tag{3.1}$$

where f is a function that describes how Y depends on X.

In statistical analysis, the relationship between the independent and dependent variables can be analyzed using various methods such as correlation and regression analysis. The correlation coefficient r measures the strength and direction of the linear relationship between two variables:

$$r = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum(X_i - \bar{X})^2 \sum(Y_i - \bar{Y})^2}}, \tag{3.2}$$

where X_i and Y_i are the individual sample points, and $\bar{X}$ and $\bar{Y}$ are the means of X and Y respectively.

Regression analysis can be used to model the relationship between the independent and dependent variables. In simple linear regression, the relationship is modeled as:

$$Y = \beta_0 + \beta_1 X + \epsilon, \tag{3.3}$$

where β_0 is the intercept, β_1 is the slope, and ϵ is the error term.

More on exact methods to determine the meaning of such relations is discussed later in *statistics*.

3.1.2 Kolmogorov's Axioms

Kolmogorov's three axioms are the most well-known description of the fundamental properties of probability theory. Let $\Omega = \{\omega_1, \omega_2, \omega_3, \omega_4, \omega_5, \omega_6\}$ be the sample space of a random experiment, A and B subsets of Ω, and P a function that assigns a real number between 0 and 1 to each A. $P(A)$ is called probability if the following three conditions (axioms) are met:

1. $P(A) \geq 0$: This condition states that the probability of any subset of Ω (event) is non-negative. This property is also called non-negativity.

2. $P(\Omega) = 1$: The second axiom further restricts the range of values of the function P. With axioms 1 and 2, $P(A)$ can take any value between 0 and 1.

3. $P(A \cup B) = P(A) + P(B)$ for disjoint sets A and B: This means that no outcome satisfies both events. A and B are called disjoint in this case.

3.1.3 Probability Concepts

Probability Formulas

In probability theory, there are various concepts that help us understand the relationships between different events or variables. These include joint probability, marginal probability, and conditional probability. The joint probability $P(X, Y)$ gives the probability that both event X and event Y occur. It can be calculated by multiplying the conditional probability of one event given the other by the probability of the other event:

$$P(X, Y) = P(X|Y) \cdot P(Y) = P(Y|X) \cdot P(X). \tag{3.4}$$

The marginal probability (in English **marginal** or **total probability**) $P(X)$ is the probability that event X occurs, regardless of other events. It can be calculated by summing the joint probabilities over all possible values of Y:

$$P(X) = \sum_y P(X, Y = y) = \sum_y P(X|Y = y) \cdot P(Y = y). \tag{3.5}$$

The conditional probability $P(X|Y)$ gives the probability that event X occurs given that event Y has occurred. It can be calculated by dividing the joint probability of X and Y by the probability of Y:

$$P(X|Y) = \frac{P(X, Y)}{P(Y)}. \tag{3.6}$$

If a third variable Z is present, these formulas can be extended to calculate conditional and joint probabilities considering Z. For example, the joint probability of X and Y given Z is defined as:

$$P(X, Y|Z) = P(X|Y, Z) \cdot P(Y|Z) = P(Y|X, Z) \cdot P(X|Z). \tag{3.7}$$

And the conditional probability of X given Y and Z is defined as:

$$P(X|Y, Z) = \frac{P(X, Y|Z)}{P(Y|Z)} \tag{3.8}$$

with Y and Z being variable.

Conditional Independence

Conditional independence occurs when the occurrence of event X has no effect on the probability of event Y given event Z. In other words, X and Y are conditionally independent if the conditional probability of Y given X and Z is equal to the conditional probability of Y given Z. This is mathematically expressed by the equation

$$P(Y|X,Z) = P(Y|Z). \tag{3.9}$$

If X and Y are independent given Z, it is written as

$$P(X,Y|Z) = P(X|Z) \cdot P(Y|Z). \tag{3.10}$$

Independence of Variables

Two variables X and Y are independent if the occurrence of one variable has no effect on the probability of the occurrence of the other variable. Mathematically, this means that the joint probability of X and Y is equal to the product of the individual probabilities of X and Y:

$$P(X,Y) = P(X) \cdot P(Y). \tag{3.11}$$

If X and Y are independent, it also holds that

$$P(X|Y) = P(X) \tag{3.12}$$

and

$$P(Y|X) = P(Y). \tag{3.13}$$

3.2 Probability Distributions

Probability distributions describe how probabilities are assigned to different outcomes of a random variable. They are categorized into discrete and continuous types.
A **Discrete Random Variable** has a finite or countable number of possible values, such as the outcomes of a coin toss or a die roll.
A **Continuous Random Variable** has an infinite number of possible values within a given range, such as hair length measured with increasing precision.
The **Probability Density Function** (PDF) describes the likelihood of a continuous random variable taking on a particular value. The total area under the PDF curve is 1, representing the total probability. Since the probability of a single value is effectively zero, probabilities are calculated over intervals using integration, resulting in the cumulative distribution function (CDF).
Key concepts include the expected value E, variance V (with standard deviation $\sqrt{V}$), and the PDF.

3.3 Continuous Probability Distributions

3.3.1 Normal Distribution

The normal distribution is also called the Gaussian bell curve. The two parameters (μ and σ) represent the mean and standard deviation of the normal distribution. The central limit theorem states that under certain general conditions, the sum of n independent, identically distributed random variables is again normally distributed.
As an example, let's consider rolling n fair dice: If you roll only one die, each face value is equally likely. However, if you roll n dice, the average face value is described by the normal distribution.
Therefore, the normal distribution is the most important, as natural phenomena with sufficiently large n approximate it. The formula for calculating the distribution is:

$$f(x, \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} \cdot e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}. \tag{3.14}$$

It holds that $E = \mu$ and $V = \sigma^2$ (variance V and standard deviation σ). The total area enclosed by the curve of the normal distribution is always 1. If $\mu = 0$ and $\sigma = 1$, it is called the standard normal distribution, which is described by a simplified equation (since $\mu = 0$ and $\sigma = 1$):

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2}. \tag{3.15}$$

The prefactor ensures that the total area under the curve (and thus the integral from $-\infty$ to ∞) is exactly 1. The $\frac{1}{2}$ in the exponent of the exponential function gives the normal distribution a unit variance. Every normal distribution is a variant of the standard normal distribution with scaled standard deviation ($\frac{1}{\sigma}$) and *z-transformed* $\frac{x-\mu}{\sigma}$. The normal distribution is usually written as: $\mathcal{N}(\mu, \sigma^2)$.

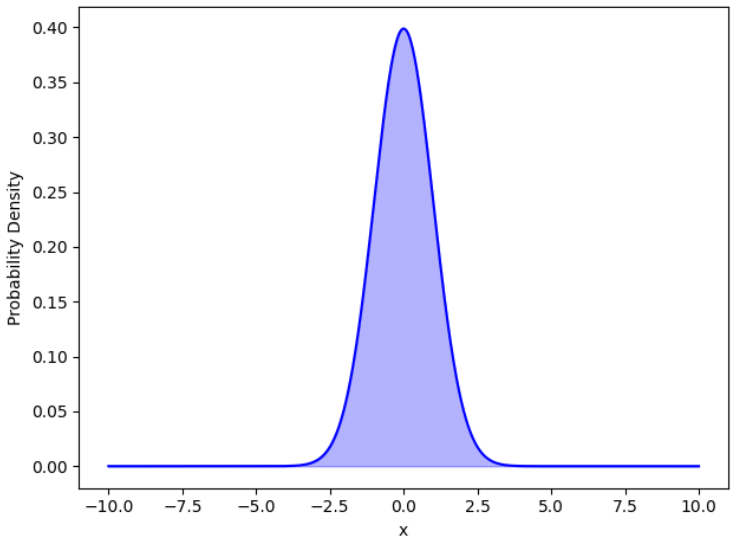

Figure 3.1: Normal Distribution with $\mu = 0$ and $\sigma = 1$

Mixed Distribution

A mixed distribution consists of several subsets that together form a larger distribution. Such an approach can be used to model a large population with different subpopulations, each of which has individual characteristics.

Formally, for each subpopulation z, a specific distribution $P(X \mid Z = z)$ is provided. These are mixed according to the probability $P(Z = z)$ of selecting an individual from this subpopulation, i.e.,

$$P(X = x) = \sum_z P(Z = z) \cdot P(X = x \mid Z = z). \tag{3.16}$$

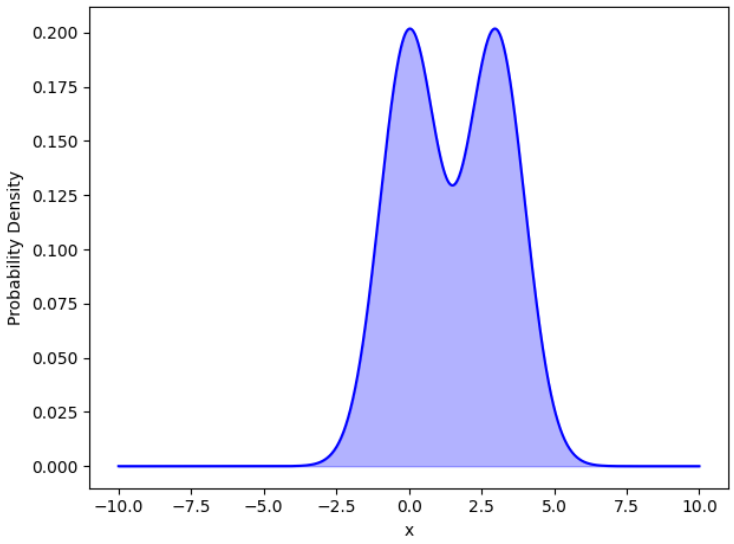

Figure 3.2: Mixed Normal Distribution

t-Distribution or Student's t-Distribution

The t-distribution, also known as Student's t-distribution, is used in statistical procedures when the sample size is small and the population standard deviation is unknown. It compensates for the underestimation of the standard deviation by the normal distribution in such cases. The t-distribution has wider tails than the normal distribution, which accounts for the increased variability in smaller samples. As the sample size increases, the t-distribution approaches the normal distribution.
The t-distribution is defined as:

$$T = \frac{Z}{\sqrt{V/\nu}}, \tag{3.17}$$

where Z is a standard normal variable, V is a chi-squared variable with ν degrees of freedom, and T has an expected value $E = 0$ and variance $V = \frac{\nu}{\nu-2}$.

Cauchy Distribution

The Cauchy distribution, also known as the Lorentz distribution, is a continuous probability distribution with heavy tails and undefined mean and variance, often used in physics and spectroscopy. Unlike the normal distribution, the Cauchy distribution does not have a finite mean or variance, which makes it distinct in its applications and properties.

3.3.2 Exponential Distribution

The exponential distribution is a continuous distribution used to model the duration of random time intervals. The parameter λ represents the number of expected *events* per time interval. Examples include the length of

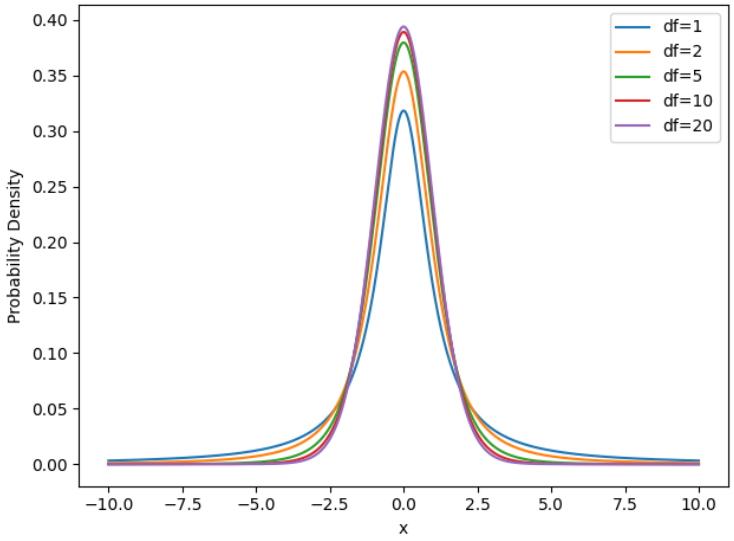

Figure 3.3: t-Distribution

a telephone call or radioactive decay. The distribution does not allow negative values, as negative times are meaningless. It is often abbreviated as $\exp(\lambda)$ in statistics. The density function is defined as follows:

$$f_\lambda(x) = \begin{cases} \lambda e^{-\lambda x} & \text{if } x \geq 0 \\ 0 & \text{if } x < 0 \end{cases} \tag{3.18}$$

The expected value E is defined as $\frac{1}{\lambda}$, the variance V as $\frac{1}{\lambda^2}$. The mode (the value at which the probability is highest) of this density function is at $x_{\text{mod}} = 0$. If you wish to calculate the probability of the occurrence of an event, it is ideal to use the cumulative distribution function $F(x)$, which forms the integral up to a value x. This creates an accumulated probability $P(X \leq x)$. Often, the actual distribution is not an exponential distribution, but the exponential distribution is easy to handle and is applied for simplification. It is applicable when a Poisson process is present, i.e., the Poisson assumptions are fulfilled. The exponential distribution is part of the much larger and more general exponential family, a class of probability measures characterized by ease of handling.

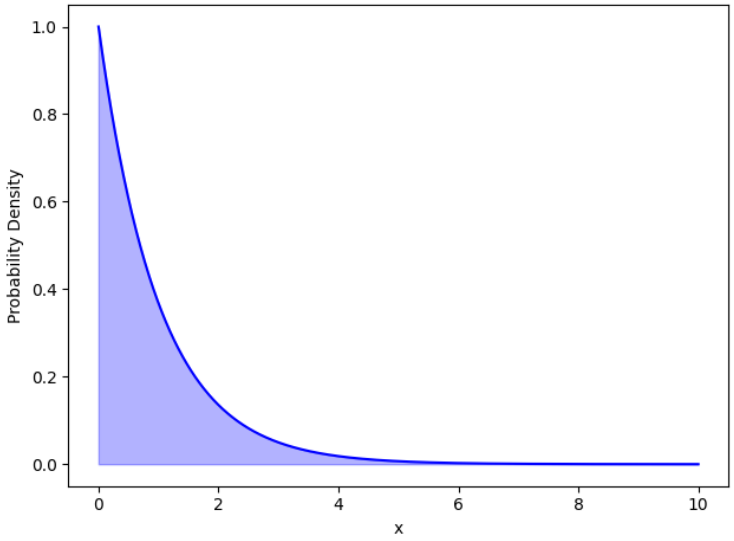

Figure 3.4: Exponential Distribution

Poisson Distribution

The Poisson distribution is a discrete probability distribution that describes the distribution of count data. In other words, how often does a certain countable event occur if it is repeated very often? The parameter here indicates the average event rate. The probability for the random variable X of the Poisson distribution is calculated using the following formula:

$$P(X = x) = \frac{\lambda^x}{x!} \cdot e^{-\lambda}, \quad x \in \mathbb{N}_0. \tag{3.19}$$

There is a connection between the exponential distribution and the Poisson distribution. Both consider the same phenomenon from different perspectives. The exponential distribution indicates how the probability of the duration of various processes is distributed. The Poisson distribution counts how often the counted events occur in a fixed interval. Starting from the exponential distribution, one wants to determine the probability that exactly n events occur in a time interval of t. As will be shown, the result is the Poisson distribution. Since the binomial coefficient for larger values can only be calculated with increased computational effort, one can use the Poisson distribution to approximate the binomial distribution. The Poisson distribution is generally used to approximate the binomial distribution when n is large and p is small. For the expected value $E = \mu$ of the Poisson distribution, we use $\mu = \lambda = n \cdot p$, which is identical to the variance. In general, the Poisson distribution approximates the binomial distribution very well for values of $n \geq 100$ and $\lambda \leq 10$. In addition to the speed advantages in calculation, the Poisson distribution has the advantage that it is countably infinite, thus extending into positive infinity.

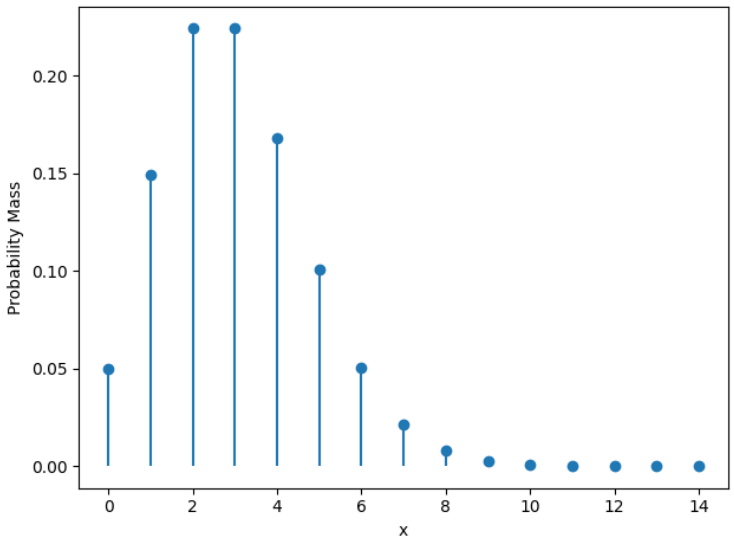

Figure 3.5: Poisson Distribution

Gamma Distribution

The gamma distribution is a continuous probability distribution that generalizes the exponential distribution by introducing a shape parameter α and a rate parameter β. It is often used to model waiting times for multiple events in a Poisson process. The gamma distribution is defined by the following probability density function (PDF):

$$f(x; \alpha, \beta) = \frac{\beta^\alpha x^{\alpha-1} e^{-\beta x}}{\Gamma(\alpha)}, \quad x > 0, \tag{3.20}$$

where $\Gamma(\alpha)$ is the gamma function, which generalizes the factorial function to non-integer values. The expected value E and variance V of the gamma distribution are given by:

$$E = \frac{\alpha}{\beta}, \quad V = \frac{\alpha}{\beta^2}. \tag{3.21}$$

The gamma distribution is particularly useful in Bayesian statistics and reliability engineering, where it is used to model the time until failure of systems with multiple components.

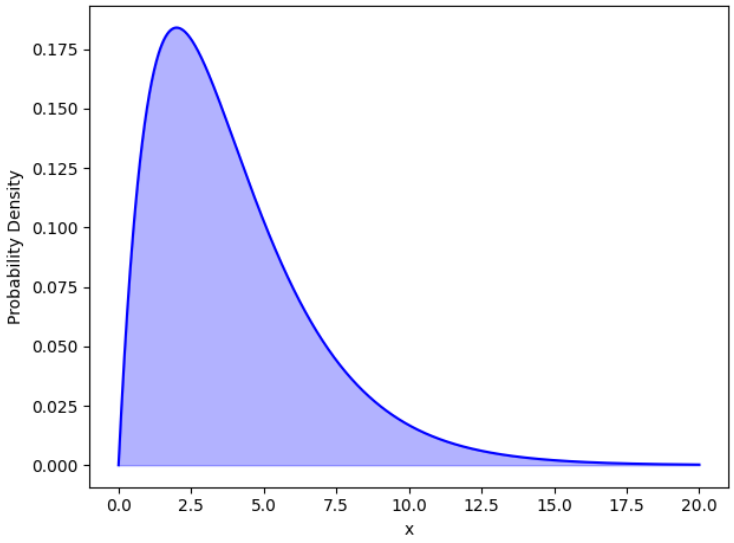

Figure 3.6: Gamma Distribution with $\alpha = 2$ and $\beta = 1$

Beta Distribution

The beta distribution is a continuous probability distribution defined on the interval [0, 1], making it suitable for modeling proportions and probabilities. It is characterized by two shape parameters, α and β, which determine the distribution's shape. The probability density function (PDF) of the beta distribution is given by:

$$f(x; \alpha, \beta) = \frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha, \beta)}, \quad 0 \leq x \leq 1, \tag{3.22}$$

where $B(\alpha, \beta)$ is the beta function, which serves as a normalization constant. The expected value E and variance V of the beta distribution are given by:

$$E = \frac{\alpha}{\alpha + \beta}, \quad V = \frac{\alpha\beta}{(\alpha + \beta)^2(\alpha + \beta + 1)}. \tag{3.23}$$

The beta distribution is widely used in Bayesian statistics, particularly as a prior distribution for binomial proportions. It is also used in project management for modeling task completion times and in various fields where probabilities and proportions are analyzed.

χ^2 Distribution

The χ^2 distribution is a continuous distribution often used for testing statistical independence or the validity of a hypothesis (goodness of fit), such as in the *Pearson's chi-square test* **??**. Very few real-world phenomena are well described by the χ^2 distribution. There is a parameter that defines the degrees of freedom n[1]. These degrees of freedom determine the distribution insofar as

[1]If one were to take random independent samples of n normally distributed quantities, these samples would sum to a χ^2 distribution with n degrees of freedom.

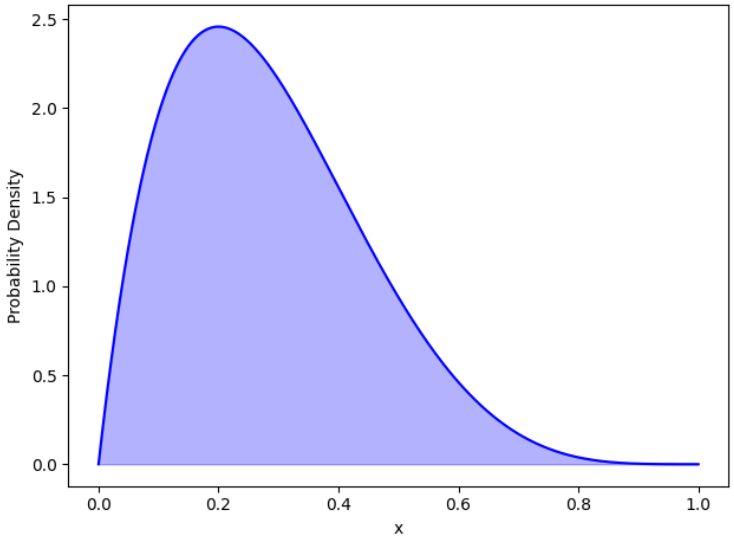

Figure 3.7: Beta Distribution with $\alpha = 2$ and $\beta = 5$

$$\chi_n^2 = Z_1^2 + \cdots + Z_n^2 \tag{3.24}$$

holds, meaning that n independent, squared, and standard normally distributed random variables are approximately equivalent to it, that is,

$$Z_k \sim \mathcal{N}(0,1) \quad \forall k = 1, \ldots, n \tag{3.25}$$

and

$$\chi^2 \sim \chi_n^2. \tag{3.26}$$

It also holds that

$$E_{\chi^2} = n \tag{3.27}$$

and

$$V_{\chi^2} = 2 \cdot n. \tag{3.28}$$

3.3.3 Uniform Distribution

The French mathematician Pierre Simon de Laplace (1749 to 1827) was one of the first to intensively study random experiments in which it can be assumed that each of its outcomes occurs with the same probability. Random experiments with uniform distribution are called Laplace experiments. The uniform distribution is a special case among probability distributions, as it exists both as a *continuous* and as a *discrete* distribution. Here are briefly the formulas for their calculation. First, for the case of a discrete distribution:

$$f(x) = \frac{1}{n}, \quad E(x) = \frac{n+1}{2}, \quad V(x) = \frac{1}{n}\sum_{i=0}^{n}(x_i - \mu)^2. \tag{3.29}$$

And for a continuous distribution:

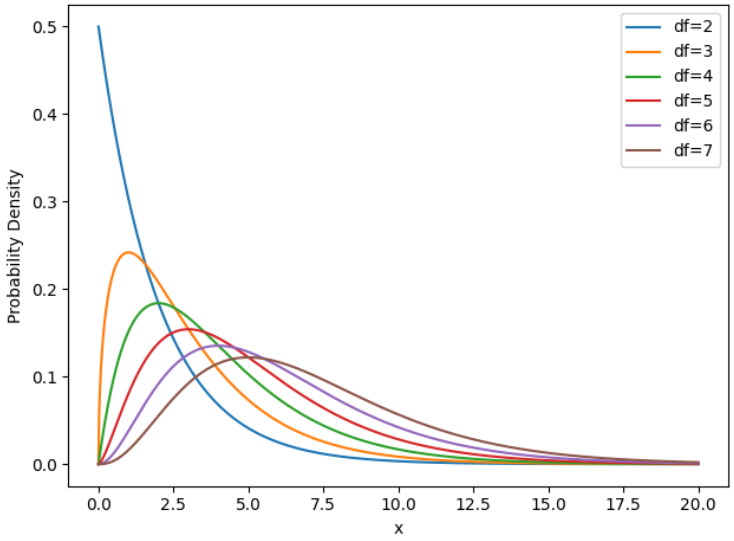

Figure 3.8: χ^2 Distribution

$$f(x) = \begin{cases} \frac{1}{b-a} & \text{if } a \leq x \leq b \\ 0 & \text{otherwise.} \end{cases} \tag{3.30}$$

Here, a and b are the boundaries of an interval that includes x. Since the same probability applies to all x, this depends on the boundaries of the interval

$$E(x) = \frac{a+b}{2}, \quad V(x) = \frac{1}{12}(b-a)^2. \tag{3.31}$$

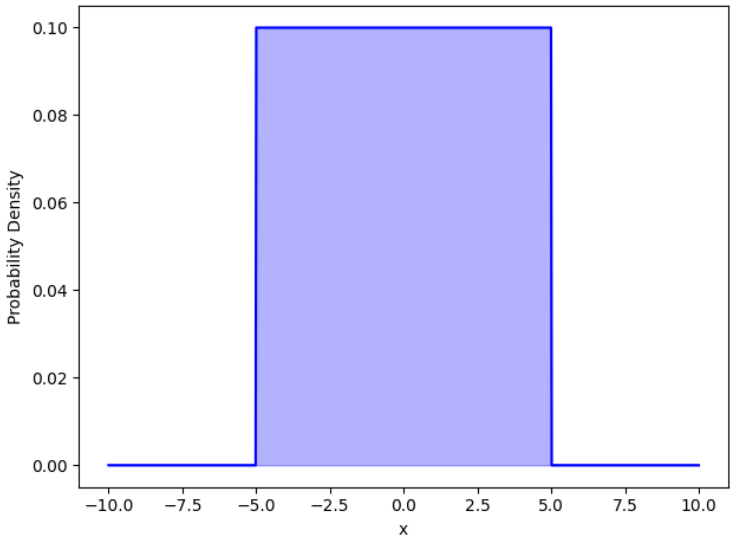

Figure 3.9: Uniform Distribution

3.4 Discrete Distributions

Binomial Distribution

Processes in which only two possible outcomes are conceivable (e.g., a coin toss) can be described with the binomial distribution. A prerequisite is that the experiment consists of identical and independent trials. The

parameters n and k suggest that this is a discrete probability distribution that answers questions about k successes in n trials. It holds:

Variable	Formula
$P(X = k)$	$\binom{n}{k} p^k (1-p)^{n-k}$
E	np
V	npq
σ	$\sqrt{npq}$
$\binom{n}{k}$	$\frac{n!}{k!(n-k)!}$

(3.32)

The binomial coefficient describes the number of ways in which k objects can be arranged in a group of n without repetition. The binomial distribution is left-skewed when $p > 0.5$[2], right-skewed when $p < 0.5$, and symmetric when $p = 0.5$. When n is sufficiently large, the normal distribution can be used as an approximation to the binomial distribution, as the skewness decreases with increasing n.

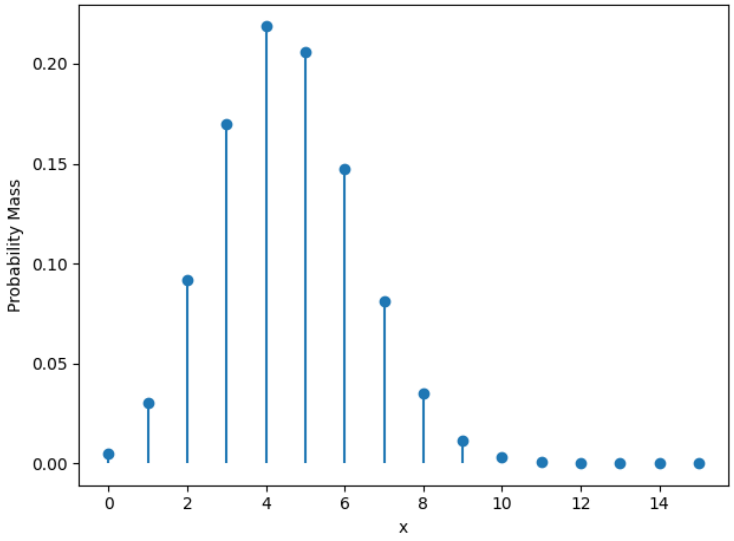

Figure 3.10: Binomial Distribution with $n = 15$ and $p = 0.3$

Bernoulli Distribution

The Bernoulli distribution is a discrete distribution whose random variable X takes only two values: 0 (failure) or 1 (success). It arises when performing a Bernoulli experiment (which has only two possible outcomes) exactly once. The Bernoulli distribution is therefore a special case of the binomial distribution for $n = 1$.
It holds:

$$E_{\text{Bernoulli}} = p, \tag{3.33}$$

$$V_{\text{Bernoulli}} = p(1-p), \tag{3.34}$$

and

[2]Greater than but not equal to! Symbol missing.

$$f_{\text{Bernoulli}}(x) = \begin{cases} 1-p & \text{if } x = 0 \\ p & \text{if } x = 1 \\ 0 & \text{otherwise.} \end{cases} \tag{3.35}$$

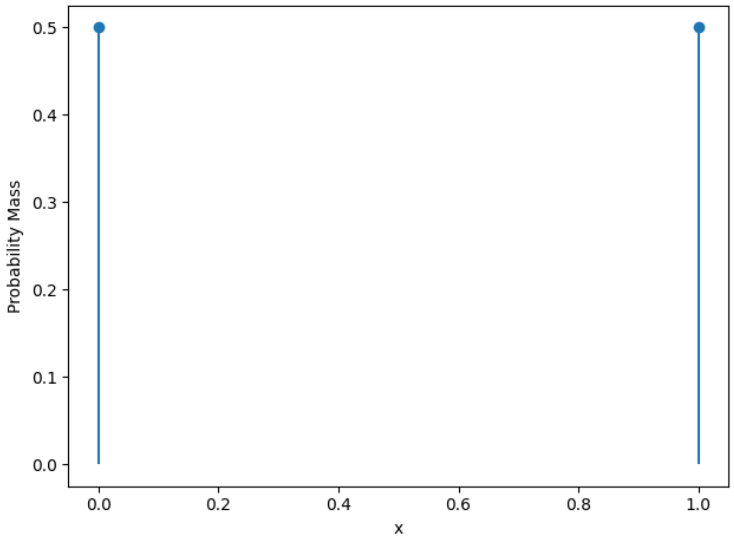

Figure 3.11: Bernoulli Distribution with $p = 0.5$

3.5 Maximum Likelihood Estimation

Maximum Likelihood Estimation (MLE) is a method used to estimate the parameters of a statistical model. It works by finding the parameter values that maximize the likelihood function, which measures how well the model explains the observed data. The purpose of MLE is to provide the most likely estimates of the model parameters based on the given data.
The likelihood function $L(\theta)$ for a set of observations $X = \{x_1, x_2, \ldots, x_n\}$ is defined as the joint probability of the observations given the parameters θ:

$$L(\theta) = P(X|\theta) = \prod_{i=1}^{n} P(x_i|\theta). \tag{3.36}$$

To find the MLE, we take the natural logarithm of the likelihood function, known as the log-likelihood function $\ell(\theta)$, and then find the parameter values that maximize it:

$$\ell(\theta) = \log L(\theta) = \sum_{i=1}^{n} \log P(x_i|\theta). \tag{3.37}$$

The MLE $\hat{\theta}$ is the value of θ that maximizes $\ell(\theta)$:

$$\hat{\theta} = \arg\max_{\theta} \ell(\theta). \tag{3.38}$$

To illustrate MLE, take a look at an animation that shows how the likelihood function changes as the parameter values are adjusted: execute the script found here: mle_viz.py

3.6 Bayes

Bayes' theorem makes statements about probabilities given that one knows data or observations, moving from events ($Data$) to their causes (parameters Θ). The more data there is, the more reliable the probability distribution of a variable becomes. Bayes' theorem states:

$$P(\Theta|Data) = \frac{P(Data|\Theta) \cdot P(\Theta)}{P(Data)}. \tag{3.39}$$

Here, $P(\Theta|Data)$ is the **posterior** probability of the parameters given the data. $P(Data|\Theta)$ is the **likelihood** of observing the data given the parameters. For $P(\Theta)$, one usually sets a **prior**, that is, its probabilities without prior knowledge.
Let us consider the so-called "Monty Hall Problem". In this game, there are three doors, behind one of which is a prize, and behind the other two are goats. The player selects a door, and the host opens another door behind which there is a goat. The player then has the option to stay with their original choice or to switch. The question is: Should the player stay with their original choice? Using Bayes' theorem, this question can be clearly answered.
Let A be the chosen door and C the door that the host opens. What is the probability that the prize is behind door A, given that C was opened, and what is it for door B? ***We need to calculate*** $P(A|C)$ ***and*** $P(B|C)$**.** If $P(B|C) > P(A|C)$, one should switch! We know that $P(A) = P(B) = P(C) = \frac{1}{3}$, and that only one door is opened behind which there is not the prize. It holds:

$$P(A|C) = \frac{P(C|A) \cdot P(A)}{P(C)} = \frac{P(C|A) \cdot P(A)}{P(C|A) \cdot P(A) + P(C|B) \cdot P(B) + P(C|C) \cdot P(C)}, \tag{3.40}$$

where $P(C)$ normalizes the probabilities according to the principle of total probability (the sum of all probabilities under which C is opened in our case). Substituting the values, we obtain:

$$P(A|C) = \frac{0.5 \cdot \frac{1}{3}}{0.5 \cdot \frac{1}{3} + 1 \cdot \frac{1}{3} + 0 \cdot \frac{1}{3}} = \frac{1}{3}. \tag{3.41}$$

To understand the logic, ask the question: *How likely is it to open door X, given that the prize is behind door Y?* On the other hand:

$$P(B|C) = \frac{P(C|B) \cdot P(B)}{P(C)} = \frac{P(C|B) \cdot P(B)}{P(C|B) \cdot P(B) + P(C|A) \cdot P(A) + P(C|C) \cdot P(C)} \tag{3.42}$$

$$= \frac{1 \cdot \frac{1}{3}}{1 \cdot \frac{1}{3} + 0.5 \cdot \frac{1}{3} + 0 \cdot \frac{1}{3}} = \frac{2}{3} \tag{3.43}$$

And thus greater for B, so one should switch.

Therefore, using Bayes' theorem, we find that the probability of winning by switching doors is $\frac{2}{3}$, while the probability of winning by staying is only $\frac{1}{3}$. This counterintuitive result demonstrates the power of Bayesian reasoning in updating probabilities based on new information.

Bayes' theorem is widely used in various fields such as statistics, machine learning, medicine, and engineering. It allows for the updating of beliefs in light of new evidence, making it a fundamental tool for probabilistic inference.

In practical applications, Bayes' theorem helps in situations where one needs to determine the probability of a hypothesis given observed data. For example, in medical diagnostics, Bayes' theorem can be used to calculate the probability of a disease given a positive test result, taking into account the prior probability of the disease and the accuracy of the test.

Bayesian methods also play a crucial role in modern data science and machine learning algorithms, such as Bayesian networks and Bayesian inference, which provide a probabilistic approach to reasoning under uncertainty.