Probability for Data Science
A Quick and Practical Handbook

Tommi Bimbato

February 20, 2025

For contact and more: **GitHub: tbimbato**

Preface

This handbook is derived from the lecture notes of the "Probability For Data Science" course, part of the Master's degree in Data Science at the University of Verona, taught by Professors Paolo Dai Pra and Francesca Collet during the Academic Year 2024/2025.

This file is open to all students who need it. Please note that this is an amateur handbook, and I am not responsible for any errors or inaccuracies. Any form of paid or non-free distribution is strictly prohibited. (Contact me if you became aware of commercial use of this handbook).

It is important to note that it is **not my intention to be precise or overly theoretical**. This is an **informal and amateur handbook** designed to support **real lectures** and **real study**. If you find any mistakes or inaccuracies, please feel free to contact me (refer to the source where you obtained this manual).

If you disagree with certain theoretical nuances or details, kindly understand that the purpose of this handbook is **not** to be rigorouslly precise but to serve as a simple and accessible guide for learning.

Contents

1 Introduction

In this handbook, we assume that the reader is familiar with the concept of a sample space S (or Ω), which is the set of all possible outcomes of a random experiment. We will build upon this foundation to introduce the basic principles of probability, random variables, and their properties.

2 Basic Probability Concepts

2.1 Axioms of Probability

- A probability of an event A is always a number between 0 and 1.
- The probability of the sample space S is 1.
- The probability of the union of two disjoint events is the sum of their individual probabilities.

2.1.1 Useful rules to be remebered

- $P(A^c) = 1 - P(A)$ or: the probability of the complement of an event is 1 minus the probability of the event itself.
- $P(A \cup B) = P(A) + P(B) - P(A \cap B)$ or: the probability[1] of the union of two events equals the sum of their probabilities minus the probability of their intersection (to avoid double counting). If the events are independent, their probabilities simply add up. If they are disjoint this P is 0.
- $P(A \cap B) = P(A)P(B|A)$ or: the probability of the intersection of two events equals the probability of the first event times the conditional probability of the second event given the first.

2.2 Conditional Probability

The concept of *conditional probability* is basically like saying: "hey, whats the probability of that event knowing that this one already happened?". It is denoted as $P(A|B)$ and is defined as:

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

You can read it as "the probability of A **given** B".

2.3 Bayes' Formula

Bayes' Formula is a very useful tool in probability theory and statistics. It is used to update the probability of a hypothesis given new evidence. Let's say we have this event, that has a certain P. Another event occur and now we want to *update* that P in order to take into account also the fact that the new event happend. The formula is:

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

Where:

- $P(A|B)$ is the probability of event A given event B.
- $P(B|A)$ is the probability of event B given event A. (the 'reverse' on the first up there)
- $P(A)$ and $P(B)$ are the probabilities of events A and B respectively.

But why it works? Bayes' formula works because it flips conditional probabilities by using known information.

It starts with what we know: how often a cause happens overall (prior) and how likely it is to see a specific effect if the cause is true (likelihood). Then, it updates this with how common the effect is overall (normalization). This gives us the probability of the cause given the effect. It's like updating your guess about a situation after seeing new evidence.

[1] From now on sometime called just 'P'.

2.4 independence

Two events are independent if the occurrence of one does not affect the occurrence of the other. Mathematically, two events A and B are independent if:

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

A practical example of indipendet event, taking into account a simple random experiment as tossing a dice can be the following: the event of getting a 6 on the first toss is independent of the event of getting a 6 on the second toss because the two tosses are unrelated and can't influence each other.

But it is important to understand that independence doesn't mean disjointedness. Two events can be independent and still have some outcomes in common. For example, the events of getting a 6 on the first toss and getting an odd number on the second toss are independent, even though they share the outcome of getting a 6.

2.4.1 Independence properties

Two independent events brings some useful properties:

- $P(A \cap B) = P(A) \cdot P(B)$ or: the probability of the intersection of two independent events is the product of their individual probabilities. No need to subtract the probability of the intersection, that actually is 0.
- In Markov chains, as we can see further in this handbook, the probability of a sequence of events is the product of the probabilities of the individual events because future is indipendent from the past.

3 Random Variables

A Random Variable[2] is a variable whose possible values are numerical outcomes of a random phenomenon. Let's say you have an experiment with n possible outcomes, you can instantiate a RV for describe the outcome of that experiment, but it must be a number, so in the case your outcome is not numerical you have to convert it (in a sort, a kind of pre-processing).

It is a function that assigns a real number to each outcome in the sample space. There are two types of random variables: discrete and continuous. Mathematically:

$$X : S \to \mathbb{R}$$

Where S is the sample space and $\mathbb{R}$ is the set of real numbers.

3.1 Distribution Functions

The point with RVs is to study the likelihood that a RV takes a certain value, and so the probability this RV takes a certain value is called $P(X = x)$ or: the probability that big X (the RV) takes the value little x

- **Discrete Time RVs**: the outcome is finite and numerable, like the number of heads in a series of coin flips, success in a series of some kind of trials. The probability this RV takes a certain value is called $P(x)$ or: the probability mass function of X.

Focus	
Discrete	Probability mass function: $P(X = x) = P(x)$

- **Continuos Time RVs**: the outcome is infinite and not numerable, like the height of a person, the time between two events. The probability this RV takes a certain value is called $f(x)$ or: the probability density function of X.

[2] From now on 'RV' or 'RVs'

Focus	
Continuos	Probability density function: $P(X = x) = f(x)$

3.2 Cumulative Distribution Function

When you are dealing with RVs you don't need always only the likelihood that a RV takes a certain value, but also the likelihood that a RV takes a value less than or equal to a certain value. This is where the Cumulative Distribution Function (CDF) comes in. It is defined as:

$$F(x) = P(X \leq x)$$

Or: the probability that our RV outcome stays less or equal to a certain value x. For instance you want to calculate what's the probability that the sum of 10 dice tosses is lesso or equal to 32? You can do it with it.

We call it 'cumulative' because it accumulates the probability of all the values less than or equal to a certain value x. And we write it as:

$$F(x) = P(X \leq x)$$

Or: big F of x describes the probability that the outcome of X stays under (but including) x.

- For **discrete** RVs, the CDF is the sum of the probabilities of all the values less than or equal to a certain value.
$$F(x) = \sum_{x_i \leq x} P(x_i)$$
- For **continuous** RVs, the CDF is the integral of the probability density function from negative infinity to a certain value (that is actually x to be clear).
$$F(x) = \int_{-\infty}^{x} f(x)dx$$

3.3 Expected Value

Well, thats really what it sounds like: the expected value of a RV is the average value of the RV over many trials. It is basically what you can expect that could happen on a "long-run" (maybe?[3]).

Let's say we toss two dice for a thousand time and watch the sum of their outcomes: after that for motivation that we will understand later the more likely value (ore expected value) will be 7. It is denoted as $E[X]$ or μ and is calculated as:

- For **discrete** RVs:
$$E[X] = \sum_{i} x_i P(x_i)$$
Or: multiply every possible outcome of X by the probability of that outcome and sum them up.
- For **continuous** RVs:
$$E[X] = \int_{-\infty}^{\infty} x f(x)dx$$
Or multiply every possible outcome of X by the probability density function of that outcome and integrate them. It's the balance point of the distribution (a sort of outcome baricenter).

So the expected value is just a pondered average of the possible outcomes of a RV, where the weights are the probabilities of the outcomes.

[3] Kind of...

3.4 Variance

The variance[4] of a RV is a measure of how much the values of the RV vary around the expected value. It is denoted as $Var(X)$ or σ^2.

Imagine throwing again two dice for a thousand times and look at the sum of the two number each toss: the variance in this example rappresent how much the sum of the two dice varies around the expected value of 7 (that is $E(X)$). The formula for the variance is:

- For **discrete** RVs:
$$Var(X) = \sum_i (x_i - E[X])^2 P(x_i)$$
Or: for each possible outcome of X, subtract the expected value from the outcome, square the result, multiply by the probability of that outcome, and sum them up.

- For **continuous** RVs:
$$Var(X) = \int_{-\infty}^{\infty} (x - E[X])^2 f(x) dx$$
Or: for each possible outcome of X, subtract the expected value from the outcome, square the result, multiply by the probability density function of that outcome, and integrate them.

Doesn't look nice to calculate, right? Well, there is a more useful formula that is:

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

Or: the variance of X is the expected value of X^2 minus the square of the expected value of X. This formula is more useful because it is easier to calculate the expected value of X^2 than the variance using the first formula.

Remark
To compute the expected value of X^2 you can use the formula: $$E[X^2] = \sum_i x_i^2 P(x_i)$$ Simply elevate small x to the square for every iteration.

3.4.1 Properties of Expected value and Variance

- Expected value properties:
  - $E[aX + b] = aE[X] + b$ or: the expected value of a constant times a RV plus another constant is the constant times the expected value of the RV plus the other constant. Meaning that the expected value is linear.
  - $E[aX + bY] = aE[X] + bE[Y]$ or: the expected value of a linear combination of two RVs is the linear combination of their expected values.
  - Spoiler: The expected value of a RV elevated at the power of n is said to be the n-th moment of the RV.
- Variance properties:
  - $Var(X) = E[X^2] - E[X]^2$ retrieved from the fact that $Car(X) = E[(X - E[X])^2]$
  - $Var(aX + b) = a^2 Var(X)$ or: the variance of a constant times a RV plus another constant is the constant squared times the variance of the RV.
  - If two RVs are **indipendent** then the variance of their sum is the sum of their variances. $Var(X + Y) = Var(X) + Var(Y)$

In the next pages we will see that some special RVs have special ways to compute Variance and Mean!

[4] We are more used to understand what it is called "standard deviation", because keeps the same unit of measure of the RV, well it is denoted as σ and is calculated as: $\sigma = \sqrt{Var(X)}$

4 Special discrete Random Variables

There are a lot of special RVs that have distinct properties and distributions. In this section, we will see some of the most common ones. When we are facing a special RV we can use some special formulas to calculate the expected value and the variance. We will describe special RVs with the following format:

$$X \sim \text{Distribution(parameters)}$$

Every special RV has a 'coded name' and a bunch of parameters that control the distribution. Where X is the RV, "Distribution" is the name of the distribution, and "parameters" are the parameters of the distribution. The symbol $\sim$ means "is distributed as". Here's a quick reference list of the special RVs we will see in this section:

Focus
• **Bernoulli**: the one that says "success" or "failure". • **Binomial**: the one that express the number of successes in a fixed number of trials. • **Poisson**: the one with the rate. • **Geometric**: the one that says "how many trials do I need to get the first success?". • **Hypergeometric**: the "what's the probability to find aces in a deck of cards drawing n cards?".

4.1 Bernoulli Random Variable

Imagine a random experiment with only two possible outcomes: success or failure. A Bernoulli Random Variable is a discrete RV that takes the value 1 if the outcome is a success and 0 if the outcome is a failure.

$$X \sim Bernoulli(p)$$

Meaning that X is distributed following a bernoullian distribution with parameter p that represent the probability of success. Bernoullian RVs has some special properties:

- Distribution function: $P(X = 1) = p$ and $P(X = 0) = 1 - p$
- Expected value: $E[X] = (1 \cdot p) + (0 \cdot (1 - p)) = p$
- Variance: $Var(X) = p(1 - p)$
- If we have multiple Bernoulli RVs and we sum them up, the result is a Binomial RV (see later).
- If we have multuple indipendent Bernoulli RVs, we can find the probability of an exact sequence of results by simply multiplying the probabilities of the individual results. For instance: we have 3 barnoullian RVs X_1, X_2, X_3 and we want to compute the P that the results will be exactly "success", "fail", "success" we obtain it simply by $P(X_1 = 1)P(X_2 = 0)P(X_3 = 1)$ or $p_1 \cdot (1 - p_2) \cdot p_3$.

4.2 Binomial Random Variable

Let's say that we have a bunch of bernoulli RVs that have the same parameter p[5]. A Binomial Random Variable is a discrete RV that represents the number of successes in a fixed number of independent Bernoulli trials. E.g.,You tossed a coin for 87 times?[6] The number of heads you get in those 87 trials is a Binomial RV.

$$X \sim Binomial(n, p)$$

You repeat an experiment n time and that experiment has probability p to be a success, here you have a binomial RV. Here's some properties:

- Distribution function: $P(X = k) = \binom{n}{k} p^k (1 - p)^{n-k}$

[5] It's clear as we will see that a Binomial random variable with parameter $n = 1$ is a bernoullian one.
[6] You have loads of free time.

- Expected value: $E[X] = np$
- Variance: $Var(X) = np(1-p)$
- Approximations:
  - **Poisson**: if n is large and p is small, the Binomial distribution can be approximated by a Poisson distribution with parameter $\lambda = np$ (we will see Poisson later but it's something like $X \approx Pois(\lambda = n \cdot p)$).
  - **Normalization**: if n is large, the Binomial distribution can be approximated by a Normal distribution with mean np and variance $np(1-p)$, in a more compact way: $X \approx N(\mu = np, \ \sigma^2 = np(1-p))$.
- Sum of Binomial RVs: if you sum up two Binomial RVs with the same parameter p, the result is a Binomial RV with the sum of the number of trials. For instance if you sum up two Binomial RVs $X \sim Binomial(n, p)$ and $Y \sim Binomial(m, p)$ the result is $J = X + Y \sim Binomial(n + m, p)$. Intuitively you can 'merge' two RVs that has the same parameter as you are 'adding' trials.

4.3 Poisson's Random Variable

Poisson's Random Variable is a discrete RV that represents the number of events that occur in a fixed interval of time or space. For instance, the number of customers that enter a store in an hour, the number of cars that pass through a toll booth in a day, the number of typos in a book. It has something to do with a concept of rate, that is the average number of events that occur in a unit of time or space.

$$X \sim Pois(\lambda)$$

So the only parameter is, in fact, λ, that is exactly that rate. Here's some properties:

- Distribution function: $P(X = k) = \frac{e^{-\lambda}\lambda^k}{k!}$
- Expected value: $E[X] = \lambda$
- Variance: $Var(X) = \lambda$
- If we sum up a number n of Poisson's independent RVs we come up with a new Poisson's RV with parameter $\lambda_1 + \lambda_2 + \ldots + \lambda_n$. E.g.,$X \sim Pois(\lambda_1)$, $Y \sim Pois(\lambda_2)$, $J = X + Y \sim Pois(\lambda_1 + \lambda_2)$.
- Poisson process: we will analyze it later, but it is a process that generates Poisson's RVs.

4.4 Geometric Random Variable

A geometric random variable is a discrete RV that represents the number of Bernoulli trials needed to get the first success. Let's say you are gambling[7] and you are playing a game where you have a probability p to win. The number of games you need to play to get your first win is a Geometric RV.

It responds to the question: "how many trials do I need to get the first success?".

$$X \sim Geom(p)$$

Where p is the probability of success. Here's some properties:

- Distribution function: $P(X = k) = (1-p)^{k-1}p$
- Expected value: $E[X] = \frac{1}{p}$
- Variance: $Var(X) = \frac{1-p}{p^2}$
- Geometric RVs has non memory. What does it mean? It means that the probability of getting the first success in exactly k trials is the same as the probability of getting the first success in exactly $k+1$ trials. It is like the past doesn't matter. Let's say you are searching the number of toss you have to do to obtain a tail for the first time. No matter how much head you head before the tail, the result is always the same. Formally: $P(X = k+1 | X > k) = P(X = k+1)$.

[7] please, don't..

4.5 Hypergeometric Random Variable

A hypergeometric random variable is a discrete RV that represents the number of successes in a fixed number of draws from a population, without replacement.

Imagine you have a deck of cards, and you want to know how many aces you will draw if you pick 5 cards at random. Since you're not putting the cards back into the deck, the probabilities change as you draw—this is what makes it hypergeometric.

It responds to the question: "how many successes will I get in a fixed number of draws without replacement?"

$$X \sim Hyper(N, K, n)$$

Where:

- N: Total size of the population.
- K: Number of successes in the population.
- n: Number of draws.
- X: Number of successes in those n draws.

Here's some properties:

- Distribution function: $P(X = k) = \frac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{n}}$ Where $\binom{a}{b}$ is the binomial coefficient, aka "how many ways can I choose b items from a."
- Expected value: $E[X] = n \cdot \frac{K}{N}$ (Makes sense: it's just the proportion of successes in the population times the number of draws.)
- Variance: $Var(X) = n \cdot \frac{K}{N} \cdot \frac{N-K}{N} \cdot \frac{N-n}{N-1}$ (That extra $\frac{N-n}{N-1}$ comes from the fact that we're not replacing items.)
- Key property: The hypergeometric distribution doesn't assume independence. Since you're not replacing what you draw, each draw affects the probabilities of future draws. This makes it fundamentally different from the binomial distribution, where trials are independent.

5 Special continuous Random Variables

In this section, we will see some special continuous RVs that have distinct properties and distributions. Here's a quick reference list of the special continuous RVs we will see in this section:

Focus
• **Uniform**: the one where every value in an interval is equally likely. • **Exponential**: the one that measures "waiting times" between events. • **Normal**: the famous bell-shaped one that appears everywhere. • **Gamma**: the one that generalizes the Exponential (sum of waiting times).

5.1 Uniform Random Variable

Uniform random variables describes the probability of a continuous RV taking a value in an interval. It is a continuous RV that has a constant probability density function over an interval. In this way every outcome within that interval is equally likely.

$$X \sim Unif(a, b)$$

Where a and b are the lower and upper bounds of the interval. Here's some properties:

- Distribution function:$f(x) = \begin{cases} \frac{1}{b-a}, & \text{for } a \leq x \leq b, \\ 0, & \text{otherwise.} \end{cases}$
- Expected value:$\mathbb{E}[X] = \frac{a+b}{2}$

- Cumulative Distribution function: $F(x) = \begin{cases} 0, & \text{for } x < a, \\ \frac{x-a}{b-a}, & \text{for } a \leq x \leq b, \\ 1, & \text{for } x > b. \end{cases}$
- Variance:$\text{Var}(X) = \frac{(b-a)^2}{12}$
- If $X \sim Unif(a, b)$ an $Y = mX + c$ then $Y \sim Unif(ma + c, mb + c)$.
- If you pick an interval within a and b, the probability that the outcome is contained in that interval is the same for every interval of the same lenght within a and b.
- If you sum up two Uniform RVs, the result is a Triangular RV.

  Let $X_1, X_2 \sim Unif(0, 1)$, then the sum $S = X_1 + X_2$ has a triangular distribution with $f_S(s) = \begin{cases} 2s, & \text{for } 0 \leq s \leq 1, \\ 2 - 2s, & \text{for } 1 < s \leq 2, \\ 0, & \text{otherwise.} \end{cases}$

5.2 Exponential Random Variable

Exponential RVs are crucial in probability theory because they model the time between events in a Poisson process. They are continuous RVs that represent the waiting time until the next event occurs. For example, the time between two customers entering a store, the time between two cars passing through a toll booth, the time between two typos in a book. It responds to the question "how much time do I have to wait until the next event?". Exp RVs are the continuous counterpart of the Geometric RV.

$$X \sim Exp(\lambda)$$

That lambda (λ) is the rate in which events occur. You can see it like the measure that says "this event occur at the rate of something every x unit of time". Here's some properties:

- Distribution function: $f(x) = \begin{cases} \lambda e^{-\lambda x}, & \text{for } x \geq 0, \\ 0, & \text{otherwise.} \end{cases}$ [8]
- Expected value:$\mathbb{E}[X] = \frac{1}{\lambda}$
- Variance: $\text{Var}(X) = \frac{1}{\lambda^2}$
- Memoryless property: the probability of waiting $t+s$ time units given that you have already waited t time units is the same as the probability of waiting s time units. Formally: $P(X > t + s | X > t) = P(X > s)$.
- if $x \sim Exp(\lambda)$ then $c \cdot X \sim Exp(\frac{\lambda}{c})$.
- Let $X_1, X_2, X_3, \ldots$ be independent exponential random variables with rates $\lambda_1, \lambda_2, \lambda_3, \ldots$, respectively. The minimum of these random variables, $\min(X_1, X_2, X_3, \ldots)$, is also an exponential random variable with parameter $\lambda = \lambda_1 + \lambda_2 + \lambda_3 + \ldots$. This means that to compute the probability of the first event occurring among a set of independent exponential random variables, you can sum up their rates and use the resulting parameter to determine the probability of the minimum.
- If you sum n independent exponential random variables with the same rate parameter λ, the resulting random variable follows a Gamma distribution with shape parameter n and rate parameter λ. Formally: Sum $\sim \Gamma(n, \lambda)$ where n is the number of summed exponential random variables, and λ is their common rate parameter.

[8] Yes, that's kinda immediate, if you are searching for the probability that an event occur at the time minus 34 hours, well, it's 0.

5.3 Gamma Random Variable

The Gamma Random Variable is a continuous RV often used to model waiting times or the sum of independent exponential random variables. It is a generalization of the exponential distribution and is widely applied in queuing theory, reliability analysis, and Bayesian statistics.

A Gamma random variable is denoted as:

$$X \sim \Gamma(k, \theta)$$

where:

- $k > 0$ is the shape parameter.
- $\theta > 0$ is the scale parameter (also called the inverse rate).

Alternatively, it can be parameterized using k (shape) and $\beta = \frac{1}{\theta}$ (rate).
Here are some key properties:

- **Probability Density Function (PDF)**:

$$f(x; k, \theta) = \frac{x^{k-1} e^{-\frac{x}{\theta}}}{\Gamma(k)\theta^k}, \quad x > 0,$$

  where $\Gamma(k)$ is the Gamma function, defined as:

$$\Gamma(k) = \int_0^\infty t^{k-1} e^{-t}\, dt.$$

- **Expected Value (Mean)**:

$$\mathbb{E}[X] = k\theta$$

- **Variance**:

$$\mathrm{Var}(X) = k\theta^2$$

- **Special Cases**:
  - When $k = 1$, the Gamma distribution reduces to the exponential distribution.
  - When k is a positive integer, the Gamma distribution describes the sum of k independent exponential random variables with the same rate.
- **Additivity Property**: If $X_1 \sim \Gamma(k_1, \theta)$ and $X_2 \sim \Gamma(k_2, \theta)$ are independent, then their sum:

$$X = X_1 + X_2$$

  is also Gamma distributed:

$$X \sim \Gamma(k_1 + k_2, \theta).$$

5.4 Normal Random Variable

The Normal Random Variable is a continuous RV that represents a distribution commonly found in nature and real-world phenomena. It is often used to model quantities such as heights, weights, test scores, or measurement errors. The distribution is symmetric, bell-shaped, and characterized by its mean and variance.

A normal random variable is denoted as:

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

Where:

- μ is the mean (the central value around which the distribution is centered).
- σ^2 is the variance (a measure of the spread or dispersion of the distribution).

Here are some key properties:

- **Probability Density Function (PDF)**:

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

where $x \in \mathbb{R}$.

- **Expected Value (Mean)**:

$$\mathbb{E}[X] = \mu$$

- **Variance**:

$$\text{Var}(X) = \sigma^2$$

- **Standard Normal Distribution**: A special case of the normal distribution occurs when $\mu = 0$ and $\sigma^2 = 1$. This is called the standard normal distribution:

$$Z \sim \mathcal{N}(0, 1)$$

Its PDF simplifies to:

$$f(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}$$

- **Linear Combination of Normal RVs**: If $X_1 \sim \mathcal{N}(\mu_1, \sigma_1^2)$ and $X_2 \sim \mathcal{N}(\mu_2, \sigma_2^2)$ are independent, then a linear combination $Y = aX_1 + bX_2 + c$ is also normally distributed:

$$Y \sim \mathcal{N}(a\mu_1 + b\mu_2 + c, a^2\sigma_1^2 + b^2\sigma_2^2).$$

- **Central Limit Theorem (CLT)**: The sum (or average) of a large number of independent and identically distributed random variables, regardless of their original distribution, approximates a normal distribution as the sample size increases.

6 More variable at the same time

6.1 Central Limit Theorem / CLT

Well, this is a big one. The Central Limit Theorem (CLT) is one of the most important theorems in probability theory.

How does it work? Imagine you have a large number of random variables, each with its own distribution. If you sum up these random variables, the distribution of the sum will be approximately normal, even if the original distributions are not normal. This is true for any distribution, no matter how weird or skewed it is. The more random variables you sum up, the closer the distribution of the sum will be to a normal distribution. It means that we can approximate any distribution with a normal one if we sum up a large number of random variables. This is why the normal distribution is so important in statistics: it is the distribution that emerges when we sum up a large number of random variables.

Example of practical application of CLT

- CLT makes us approximate discrete distibution to a normal one. For instance, if you sum up a large number of Bernoulli RVs, the distribution of the sum will be approximately normal. This is why the normal distribution is often used to approximate the Binomial distribution.
- Poisson distribution can be approximated by a normal distribution if the rate parameter is large. Say that your λ is greater than 50, you can approximate the Poisson distribution with a normal one with mean λ and variance λ.
- The mean of the sum is the sum of the means, and the variance of the sum is the sum of the variances.

6.1.1 Example: Sum of A LOT of bernoulli RVs and approximation to normal

We have a huge number of bernulli RVs, ok that could be modeled through a Binomial RV, but since the number of trials is huge we can approximate it to a normal distribution. In the end, what's the point on having a Binomial Rv with parameter $p =$ something and $n = 1000000$ if we can approximate it to a normal one with mean np and variance $np(1-p)$?

Focus

normalization of Binomial RVs

$$X \sim Binomial(n, p) \approx N(\mu = np,\ \sigma^2 = np(1-p))$$

Ok so

- The mean $\mu = n \cdot p$.
- The variance $\sigma^2 = n \cdot p \cdot (1-p)$.
- The standard deviation $\sigma = \sqrt{n \cdot p \cdot (1-p)}$.[a]

[a] I had problems remembering wether in the normalization you get σ or σ^2, well, it's σ

Now that we have our approximated normal distribution, to make it more accurate we can apply what's called the continuity correction.

Continuity correction It's a small adjustment to the boundaries of the interval to make the approximation more accurate. We have that big amount of discrete RVs, we approximated to a normal continuous distribution.

Now, if we want to calculate the probability that the sum of these RVs falls within a certain interval, we can use the normal distribution to approximate it. But since the normal distribution is continuous, we need to adjust the boundaries of the interval to account for the discreteness of the original distribution. This adjustment is called the continuity correction.

How and when to use Continuity Correction The continuity correction is used when approximating a discrete distribution with a continuous one. It is especially useful when the discrete distribution is a sum of many independent random variables.

For example we want to compute the probability that the sum of 1000 Bernoulli RVs with $p = 0.5$ falls within the interval $[480, 520]$.
We can approximate the sum to a normal distribution with mean $np = 500$ and variance $np(1-p) = 250$. Then we can use the normal distribution to calculate the probability that the sum falls within the interval $[480, 520]$.

But since the original distribution is discrete, we need to adjust the boundaries of the interval to make the approximation more accurate. This is where the continuity correction comes in.
To apply it simply adjust the boundaries of the interval by adding or subtracting 0.5. In this case, the adjusted interval would be $[479.5, 520.5]$.

Recap:

- Approximate the discrete distribution to a normal one.
- Adjust the boundaries of the interval by adding or subtracting 0.5 if you are working with discrete distributions.
- Optional: standardize the normal distribution to use standard normal values table and calculate the probability.
- Calculate the probability using the normal distribution.

Note that continuity correction is not always necessary, but it can make the approximation more accurate when working with discrete distributions. When computing something like $P(X = 12)$ and you are coming from a discrete world to a continuous one, you should use the continuity correction. That $P(X = 12)$ will be something like $P(11.5 < X < 12.5)$.[9]

[9] I forgot to write that in continuous RVs the probability that a variable is exactly a specified value is 0.

See, it's like "how likely that RV falls in the small interval around 12 that is $[11.5, 12.5]$?".

6.2 Sum of i.i.d. Random Variables, relevant examples

Sum of i.i.d. Bernoulli Random Variables Imagine you have a series of independent Bernoulli trials, each with the same success probability p. The number of successes in these trials is a Binomial random variable. But what if you sum up the outcomes of these trials? What kind of random variable do you get? If you sum n i.i.d Bernoulli RVs, you get a Binomial RV with parameters n and p. This is because the sum of Bernoulli RVs is a Binomial RV. Mathematically:

$$X_1, X_2, \ldots, X_n \sim Bernoulli(p)$$

$$Y = X_1 + X_2 + \ldots + X_n \sim Binomial(n, p)$$

Where:

- $X_1, X_2, \ldots, X_n$ are independent and identically distributed Bernoulli random variables with success probability p.
- Y is the sum of these Bernoulli random variables.

Sum of i.i.d. Poisson Random Variables Imagine you have a series of independent Poisson processes, each with the same rate parameter λ. The number of events in these processes is a Poisson random variable. But what if you sum up the outcomes of these processes? It actually happens that you get a Poisson RV with rate parameter $n\lambda$. This is because the sum of Poisson RVs is also a Poisson RV.

Example: For instance, take a Poisson RV that represents the number of customers entering a store in an hour. If you sum up the number of customers entering the store in 5 hours, you get a Poisson RV with a rate parameter 5 times the original rate parameter. No matter how many Poisson RVs you sum up, the result is always a Poisson RV if their λ is the same.

Sum of i.i.d. Exponential Random Variables If you sum n i.i.d. exponential RVs with the same rate parameter λ, the resulting random variable follows a Gamma distribution with shape parameter n and rate parameter λ.

Formally : let X_n be n exponenital RVs with the same parameter λ, $\sum_n X_n \sim \Gamma(n, \lambda)$ where n is the number of summed exponential random variables, and λ is their common rate parameter.

Sum of i.i.d. Gamma Random Variables If you sum n i.i.d. Gamma RVs with the same rate parameter λ, the resulting random variable follows a Gamma distribution with shape parameter n and rate parameter λ.

Formally : let X_n be n Gamma RVs with the same parameter λ, $\sum_n X_n \sim \Gamma(n, \lambda)$ where n is the number of summed exponential random variables, and λ is their common rate parameter.

Sum of i.i.d. Uniform Random Variables If you sum up two Uniform RVs, the result is a Triangular RV.

Example: Let $X_1, X_2 \sim Unif(0, 1)$, then the sum $S = X_1 + X_2$ has a triangular distribution with

$$f_S(s) = \begin{cases} 2s, & \text{for } 0 \leq s \leq 1, \\ 2 - 2s, & \text{for } 1 < s \leq 2, \\ 0, & \text{otherwise.} \end{cases}$$

Sum of i.i.d. normal Random Variables If you sum two normal RVs with the same mean and variance, the result is a normal RV with the sum of the means and variances.

Example: Let $X_1 \sim \mathcal{N}(\mu_1, \sigma_1^2)$ and $X_2 \sim \mathcal{N}(\mu_2, \sigma_2^2)$ be independent normal RVs. If you sum them up, $Y = X_1 + X_2$, then $Y \sim \mathcal{N}(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$.

6.3 Joint Probability Distribution

Imagine you work for a large *e-commerce company* like Amazon. The company wants to analyze two key quantities:

- X: The number of orders received in a day.
- Y: The number of those orders that are successfully delivered on the same day.

Here's what we know:

- On average, about 10 orders are placed per day (X).
- Not all orders are delivered successfully. Let's assume that 80% of the orders are successfully delivered (Y).

Question: What's the chance that on a given day, 12 orders are placed and 9 of them are delivered?

Explanation: First, think about how the orders work:

- The total number of orders (X) can vary daily, but it's usually around 10. Some days are busier, others quieter.
- Once you know how many orders you received in a day, say 12, the number of successful deliveries (Y) depends on the success rate. If 80% of orders are typically delivered, you'd expect around 80% of the 12 orders, which is roughly 9 or 10 deliveries.

So, the probability of receiving 12 orders and successfully delivering 9 depends on both the average daily orders (X) and the delivery success rate (Y).

Conclusion: This type of analysis helps the company predict busy days, optimize delivery resources, and ensure customer satisfaction.

6.4 Joint Probability Distribution

Mathematically:

$$f_{X,Y}(x, y) \text{ or } P_{X,Y}(x, y) = P(X = x, Y = y)$$

Or: the probability that in the same time X is equal to x and Y is equal to y.

7 Discrete Markov Chains

7.1 Introduction

Markov Chains [10] are a specific type of stochastic process that respond to the Markov property.

Stochastic process Is a collection of random variables that represent the evolution of a system over time.

Let say we have a stochastic process named S, it is a collection of random variables $S_1, S_2, S_3, \ldots$ that represent the state of the system at different points in time. So generally we can call it something like $S(t)$ where t is the time, so a random variable that mutates according to time passing. If we want to know what's going on regarding the system at time t, we can look at the random variable $S(t)$.

Moreover, a stochastic process is related to the concept of states, that is the possible values that the random variable can take. For example: if we are talking about the weather, the states could be "sunny", "cloudy", "rainy", etc.. and the random variable $S(t)$ represent the weather at time t, e.g., $S(t = tomorrow)$ = "sunny", paraphrasing: "what's the weather at time = tomorrow? Sunny!".

[10] From now on: "MC"

Markov property Now that we outlined the concepts behind stochastic processes, we can understand what does the markov property states. A stochastic process has the Markov property if the probability of moving to the next state depends only on the current state and not on the previous states.

What does it means? It means that the future is independent of the past given the present. So to make up a practical example, if today is rainy, the probability that tomorrow will be sunny depends only on today's weather, not on the weather of the day before yesterday.

We can outline this property in a more formal way using mathematical notation. A stochastic process S has the Markov property if:

$$P(S_{n+1} = s|S_1 = s_1, S_2 = s_2, \ldots, S_n = s_n) = P(S_{n+1} = s|S_n = s_n)$$

Paraphrasing:

- S Is the stochastic process iself. S can be a collection of random variables $S_1, S_2, S_3, \ldots$ that represent the state of the system at different points in time like "sunny", "cloudy", "rainy".
- S_n is the state of the system at time n, and to make it more understandable we can intend it as "time = today".
- S_{n+1} is the state of the system at time $n + 1$, again we can intend it as "time = tomorrow".
- Small s is a specific state, so s_1 is a specific state at time 1.
- Generally $S_{\text{day 'number'}}$ is the state of the system at the day 'number'.

That condition above says that: *The probability that the system will be in a specific state s at time = 'tomorrow' given what was the weather in all the day before, is the same as the probability that the system will be in a specific state s at time = 'tomorrow' given what was the weather JUST today.* Or, more simply: The next state depends only on the current state, not on the whole past states.

To summarize, if a stochastic process has this Markov property we can call it a Markov Chain. Or MC.

7.2 A brief overview of Markov Chains's component, characteristics and properties

Let's have a quick overview of the properties and the characteristics of Markov Chains:

- **State Space**: Every MC has a state space $\mathbb{S}$, which is the set of all possible states that the system can be in. For example, if we are modeling the weather, the state space might be $\mathbb{S}$ = {"Sunny", "Rainy"..}.
- **Transition Matrix**: A MC is related to something that we call a transition matrix. This matrix describes the probabilities of moving from one state to another. It is a square matrix where each entry P_{ij} represents the probability of moving from state i to state j. The rows of the matrix sum to 1 as this process must be Stochastic[11].
- **Stationary Distribution**: A stationary distribution is a probability distribution that remains unchanged after the system has evolved for a long time. To be more clear, the stationary distr. is that distribution that the system reaches after a generic long time[12]. A practical example of this concept can be expressed as "If you keep flipping a coin for a long time, the proportion of heads and tails will eventually stabilize at 0.5 each".
- Then there are some aspects of MC that are importand to be underlined:
  - **Irreducibility**: A MC is said to be irreducible if it is possible to reach any state from any other state. In other words, there are no "absorbing states" that the system can't escape from. No states can make it impossible to be 'escaped'.
  - **Periodicity**: When a MC is **aperiodic** it means that the system can return to a state at any time. There is no looping path that the system follows. If the system can return to a state only at multiples of a certain number, then the system is **periodic**. For example: i have a bunch of states and i go to state "6" only if i follow a specific number of steps that is multiple of 3, then the system is periodic with period 3.

[11] A characteristic of Stochastic processes is that the sum of the probabilities of all possible outcomes must be equal to 1.
[12] The longer the time the more accurate the stationary distribution is.

- **Ergodicity**: This name sounds frightening but it's not. A MC is said to be ergodic if it is both irreducible and aperiodic. In other words, the system can reach any state from any other state, and it can return to a state at any time. If a MC is aperiodic + irreducible, then it is ergodic. Easy peasy.
- States can also have their characteristics:
  * **Transient States**: When a state is said to be transient, there is some possibility that the system will never return to that state. Back to the weather example, if the state "sunny" is transient, it means that there is some possibility that the weather will never be sunny again[13].
  * **Recurrent States**: In the reality, the weather is recurrent. If a state is recurrent, the system will eventually return to that state. And actually the state "sunny" is recurrent, it means that the weather will eventually be sunny again with probability 1.
  * **Absorbing States**: The name is clear, if a state is absorbing, the system will never leave that state. It's like a black hole, once you are in, you can't escape. Probability of 'remain' in that state is 1. Weather example: let the state "black hole approaching earth" be an absorbing state, once the earth is in that state, it will never leave it[14].
  * **Periodic States**: When we have some kind of periodicity, as said above, few or all states can be part of a periodic loop. Those states are called periodic states.

7.3 Transition Matrix and State transition Graph

The transition matrix is a square matrix where each entry P_{ij} represents the probability of moving from state i to state j. The rows of the matrix sum to 1 as this process must be Stochastic.

Example: Let's say we have a MC with 3 states: "Sunny", "Cloudy", "Rainy". The transition matrix might look like this:

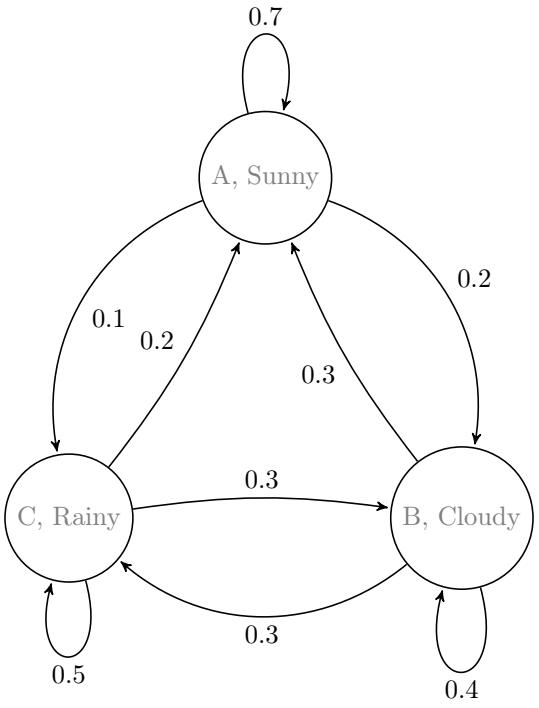
$$P = \begin{bmatrix} 0.7 & 0.2 & 0.1 \\ 0.3 & 0.4 & 0.3 \\ 0.2 & 0.3 & 0.5 \end{bmatrix}$$

This matrix tells us that:

- The probability of moving from "Sunny" to "Sunny" is 0.7.
- The probability of moving from "Sunny" to "Cloudy" is 0.2.
- The probability of moving from "Sunny" to "Rainy" is 0.1.
- And so on for the other states.

We can rapresent it also through a state transition graph. A state transition graph is a visual representation of the states and the transitions between them. Each state is represented by a node, and the transitions are represented by arrows between the nodes.

[13] Well..
[14] Well....



Following the example above, we can say, for instance, that the probability of moving from "Sunny" to "Sunny" is 0.7, the probability of moving from "Sunny" to "Cloudy" is 0.2, and so on..

7.4 State at time n

We have a beautiful MC that responds to Markov property, we have the transition matrix, we have the state transition graph, we have everything. Now we want to know, let's say, what's the state of the chain at time n?

Well, that's quite easy algebraically. We can use the transition matrix to calculate the state of the chain at any time n. Let's say we have a toy MC with

$$\mathbb{S} = \{0, 1, 2\}$$

and the transition matrix is:

$$P = \begin{bmatrix} 0.7 & 0.2 & 0.1 \\ 0.3 & 0.4 & 0.3 \\ 0.2 & 0.3 & 0.5 \end{bmatrix}$$

We also know something we call "initial distribution" that is the probability distribution of the states at time 0. Let's say that the initial distribution[15] is:

$$\alpha_0 = \begin{bmatrix} 0.3 & 0.4 & 0.3 \end{bmatrix}$$

Now we questioning: "What's the state of the chain at time $n = 3$?". We can calculate it using the transition matrix and the initial distribution.

The state of the chain at time n is given by:

$$\alpha_n = \alpha_0 \cdot P^n$$

Where:

- α_n is the probability distribution of the states at time n, it's a vector in which each entry represents the probability of being in a specific state at time n. For instance, if $\alpha_n[2] = 0.4$, it means that the probability of being in state 2 at time n is 0.4.
- α_0 is the initial distribution as said before.

[15] The initial distribution is a vector of the same dimension as $|\mathbb{S}|$ that represents the probability for the chain to start at time $t = 0$ in a specific state.

- P is the transition matrix.
- P^n is the transition matrix raised to the power of n. This represents the probability of moving from one state to another in n steps.

This is called the "N-step transition probability of the chain".

7.5 Stationary Distribution

As for random variables, MC follow a sort of law of large numbers. If you keep flipping a coin for a long time, the proportion of heads and tails will eventually stabilize at 0.5 each.

The same concept applies to MC. If you keep running a MC for a long time, the proportion of time spent in each state will eventually stabilize. This stable distribution is called the stationary distribution.

The stationary can be intended also as the long-term behavior of the chain. Or the limit of t that goes to infinity. Mathematically:

$$\lim_{n\to\infty} \alpha_n = \pi$$

Where:

- π is the stationary distribution.
- α_n is the probability distribution of the states at time n.

How do we compute this π? Well, we can compute the stationary distribution by solving a system of linear equations. The stationary distribution π is the solution to the equation:

$$\pi = \pi \cdot P$$

Where:

- π is the stationary distribution.
- P is the transition matrix.

And moreover, the sum of the entries of the stationary distribution must be equal to 1 as it represents a probability distribution.[16]

$$\sum_{i\in\mathbb{S}} \pi_i = 1$$

Now let's look it in a more practical way. We want to compute this π and for computing it we have to solve a system of two equations:

$$\begin{cases} \pi = \pi \cdot P \\ \sum_{i\in\mathbb{S}} \pi_i = 1 \end{cases}$$

- The first equation is the "balance equation", it represents the fact that the stationary distribution is unchanged by the transition matrix. You can see it as a proof that the stationary distribution is the distribution that the system reaches after a long time.
- The second equation is the normalization constraint, it ensures that the sum of the entries of the stationary distribution is equal to 1 as we want a probability distribution[17].

Coming back to out toy MC: The stationary distribution π is the solution to the equation:[18]

$$\begin{cases} \pi = \pi \cdot P \\ \pi_0 + \pi_1 + \pi_2 = 1 \end{cases}$$

[16] AKA the normalizaiton constraint.
[17] ..And we know that the sum of the probabilities of all possible outcomes must be equal to 1.
[18] Remember that $\pi \cdot P$ is a matrix multiplication.

7.6 Proportion of time spent in each state

The stationary distribution π tells us the proportion of time that the chain spends in each state in the long run. So after computing this system you actually have a vecton π composed by the proportion of time that the chain spends in each state.
Let's say that $\pi[4] = 0.324$.

It means that the state associated to the 4^{th} entry of the vector π is visited 32.4% of the time in the long run and moreover we have a probability of being in state 4 of 32.4%, again, in the long run.

8 Poisson Processes

A **Poisson process** is a stochastic process used to model the occurrence of random events over time. It's defined by a single rate parameter λ, which represents the average number of events per unit of time.

8.1 When do we use Poisson vs Exponential?

In a Poisson process:

- **Poisson random variables:** These are used to count **how many events** happen in a fixed time interval. For example, the number of events in a time interval of length t is:

  $$N(t) \sim Pois(\lambda t),$$

  where λ is the rate of events per unit time. So, how many events happened in t? That's a Poisson random variable with parameter t times λ.

- **Exponential random variables:** These are used to model **when events happen**, specifically the time between consecutive events (called "inter-arrival times"). The time between events follows an exponential distribution:

  $$T \sim Exp(\lambda).$$

  Where T is the time between events, and λ is the rate of events per unit time. So every event is separated by an amount of time that follows an exponential random variable with parameter λ, and never change, and has no memory!

In summary, Poisson counts **how many events**, while Exponential tells us **when events happen**.

Exponential RVs EXP RVs has cool properties:

- **Memoryless Property**: The time until the next event is independent of how much time has already passed.

  $$P(T > s + t | T > s) = P(T > t)$$

- **Additivity Property**: The sum of independent exponential random variables is Gamma distributed.

  $$X_1 \sim Exp(\lambda_1), X_2 \sim Exp(\lambda_2) \implies X_1 + X_2 \sim Gamma(2, \lambda_1 + \lambda_2)$$

  Where 2 is the number of summed exponential random variables and the second argument is the sum of their rate parameters.

- **Minimum Property**: The minimum of two exponential random variables is also exponential.

  $$X_1 \sim Exp(\lambda_1), X_2 \sim Exp(\lambda_2) \implies \min(X_1, X_2) \sim Exp(\lambda_1 + \lambda_2)$$

  So to find which event happens first, you actually come up with an exponential random variable with the sum of the rate parameters.

9 WIP CTM

...
...