Probability for Data Science
A Quick and Practical Handbook

Tommi Bimbato

September 30, 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

Part I

The Foundations

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. Maybe, in a non-specified future, i will add some sections or appendix that will cover set theory, combinatorics and other useful concepts that are not included in this informal handbook.

1 Basic Probability Concepts

1.1 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.

Useful rules to be remembered:

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

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

1.2 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".

1.3 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.

1.4 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.

1.4.1 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.

2 Random Variables

A random variable[2] (RV) assigns a real number to each outcome of a random experiment. If the original outcome is not numeric, we first map it to a number (simple pre-processing).

Mathematically, an RV is a function:

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

where S is the sample space and $\mathbb{R}$ the real line.

There are two broad types: **discrete** and **continuous**.

[2] From now on, we use "RV" or "RVs".

2.1 Probability Laws (PMF and PDF)

We study RVs through their probability laws:

- **Discrete RVs**: take values in a countable set (e.g., number of heads). Their law is the *probability mass function* **(PMF)**:

$$p_X(x) = \mathbb{P}(X = x)$$

.

- **Continuous RVs**: take values on intervals (e.g., heights, distance..). Their law is the *probability density function* **(PDF)**:

$$f_X(x), \text{ with } \mathbb{P}(a \leq X \leq b) = \int_a^b f_X(t)\, dt$$

What we are saying here is quite important. For discrete RVs the PMF gives you the probability that the RV takes a specific value. For continuous RVs, the PDF does not give you the probability that the RV takes a specific value (in fact, this probability is always 0), but rather it gives you a density that you can integrate over an interval to find the probability that the RV falls within that interval. Integrating between two points gives you the probability that the RV falls within that range. Set this range small enough, and you can approximate the probability of the RV being close to a specific value.

2.2 Cumulative Distribution Function (CDF)

The CDF captures "how much probability has accumulated up to x":

$$F_X(x) = \mathbb{P}(X \leq x).$$

For common cases:

- **Discrete**: $F_X(x) = \sum_{x_i \leq x} p_X(x_i)$.
- **Continuous**: $F_X(x) = \int_{-\infty}^{x} f_X(t)\, dt$.

Note that the CDF for discrete random variables is a step function. What does it means? Basically you sum up the probabilities of all the countable values until you reach your desired parameter x. In fact, Intuitively, the

probability to be at most x is the sum of the probabilities of all the values less than or equal to x.

The continuous case is a little less intuitive but the core idea remains the same. We have a PDF that says whats the probability for that RV to fall in a specified range. If we integrate from $-\infty$ to x, we get the probability that our RV is less than or equal to x.

2.3 Expected Value

The expected value is the long-run average value of an RV. Denoted $\mathbb{E}[X]$ or μ:

- **Discrete**: $\mathbb{E}[X] = \sum_i x_i \, p_X(x_i)$.
- **Continuous**: $\mathbb{E}[X] = \int_{-\infty}^{\infty} x \, f_X(x) \, dx$.

We can draw a parallel between the PMF, the PDF, and the expected value. As you can notice, $\mathbb{E}[X]$'s formula is composed in both cases by the sum (or integral) of a product between the actual implementation of a density function and the value of the RV itself. This is because the expected value is a kind of "average" of the possible values of the RV, weighted by how likely each value is to occur (given by the PMF or PDF).

2.4 Variance

Variance[3] measures how much X varies around its mean. Denoted $\mathrm{Var}(X)$ or σ^2:

- **Discrete**: $\mathrm{Var}(X) = \sum_i (x_i - \mathbb{E}[X])^2 \, p_X(x_i)$.
- **Continuous**: $\mathrm{Var}(X) = \int_{-\infty}^{\infty} (x - \mathbb{E}[X])^2 \, f_X(x) \, dx$.

Useful identity:

$$\mathrm{Var}(X) = \mathbb{E}[X^2] - (\mathbb{E}[X])^2.$$

[3]The *standard deviation* is $\sigma = \sqrt{\mathrm{Var}(X)}$ and has the same units as X.

Remark

To compute the expected value of X^2 you can use the formula:

$$\mathbb{E}[X^2] = \sum_i x_i^2 \, p_X(x_i) \quad \text{(discrete)} \qquad \text{or} \qquad \mathbb{E}[X^2] = \int x^2 f_X(x)\, dx \quad \text{(continuous)}.$$

It's just like before, but with x^2 instead of x.

But.. hey, can you see the similarity? Expected value and variance are also called respectively the first and second moment of a RV.

2.4.1 Moments

The n-th moment of a random variable X is defined as:

$$\mathbb{E}[X^n] = \begin{cases} \sum_i x_i^n \, p_X(x_i) & \text{(discrete)} \\ \int x^n f_X(x)\, dx & \text{(continuous)} \end{cases}$$

What do they tell us? A moment (of a RV) is like a summary statistic that captures certain aspects of the distribution of the RV. The first moment (expected value) gives us the average or central tendency, while the second moment (variance) tells us about the spread or variability around that average. Higher-order moments can provide insights into the shape and characteristics of the distribution, such as skewness and kurtosis.

2.5 Properties of Expected Value and Variance

- Expected value properties:
  - Linearity: $\mathbb{E}[aX + b] = a\,\mathbb{E}[X] + b$.
    The expected value of a linear transformation of a RV is the linear transformation of the expected value. We can decompose the expected value of a linear transformation of a RV into the expected value of the RV times something plus something else.
  - More generally: $\mathbb{E}[aX + bY] = a\,\mathbb{E}[X] + b\,\mathbb{E}[Y]$.
    Yes, it works also if you are searching the expected value of a linear combination of multiple RVs.
  - Spoiler: The expected value of a RV elevated at the power of n is said to be the n-th moment of the RV. (See above..)
- Variance properties:

- $\mathrm{Var}(X) = \mathbb{E}[X^2] - (\mathbb{E}[X])^2$ (from $\mathrm{Var}(X) = \mathbb{E}[(X - \mathbb{E}[X])^2]$).
  As you can see we can manipulate the formula of the variance in order to obtain a more useful one. The first moment interact with the second moment. Why? Because the variance is a measure of how much the RV varies around its mean. So, it makes sense that the mean (first moment) would be involved in the calculation of variance. The first parameter a get raised to the power of 2 because the variance is based on squared deviations from the mean.
- $\mathrm{Var}(aX + b) = a^2 \mathrm{Var}(X)$.
  The constant b disappears because $\mathrm{Var}(c) = 0$ for any constant c. Adding a constant shifts the distribution but doesn't change the spread.
- If $X \perp Y$ $\mathrm{Var}(X + Y) = \mathrm{Var}(X) + \mathrm{Var}(Y)$[4]
  If two RVs are **independent** then the variance of their sum is the sum of their variances.

3 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:

[4] I just decided, as I'm following a lecture in Uppsala right now, to copy what Benny Avelin is doing at the blackboard. The symbol $\perp$ means "independent". We can use it to denote independence between random variables without writing the entire word or using other awkward notation.

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?".

3.1 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 $p_1 \cdot (1 - p_2) \cdot p_3$.

3.2 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: $P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$
- 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: $X \approx N(\mu = np, \quad \sigma^2 = np(1-p))$.
- 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.

3.3 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

[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.

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.

3.4 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: $Var(X) = \frac{1-p}{p^2}$

[7] please, don't..

- 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: $P(X = k+1|X > k) = P(X = k+1)$.

3.5 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.

4 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).

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

4.2 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: Var$(X) = \frac{1}{\lambda^2}$

[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.

- 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.

4.3 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**:

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

- **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:

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

4.4 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.

5 More variable at the same time

5.1 Random Vectors

A rabdin vector is a mapping

$$X : \Omega \to \mathbb{R}^n$$

Where Ω is the sample space and $\mathbb{R}^n$ is the n-dimensional real space. It is a vector of n random variables.

$$X = (X_1, X_2, \ldots, X_n)$$

Example:

$$Z = \text{lenght}$$

$$W = \text{weight}$$

$$X(w) = (Z(w), W(w))$$

Where $w \in \Omega$ and $n = 2$.

5.2 Joint Distribution Function - JDF

$$F_X(x) := P(\{X_1(w)) \leq x_1\} \cap .. \cap \{X_m(x) \leq x_m\})$$

And two RV are indipendent[9] if:

$$F_{XY}(x, y) = F_X(x) \cdot F_Y(y)$$

[10]

And if we are dealing with conditional probabilities:

We have that:

$$P(X_1 = x_1 | X_2 = x_2) = \frac{P\{X_1 = x_1\} \cap \{X_2 = x_2\}}{P\{X_2 = x_2\}} \quad (1)$$

and also: (2)

$$f_{X_1}(x_1) = P(X_1 = x_1), \quad (3)$$

$$f_{X_2}(x_2) = P(X_2 = x_2) \quad (4)$$

We know that the joint distibution is: (5)

$$f_{X_1X_2}(x_1, x_2) = P\{X_1 = x_1\} \cap \{X_2 = x_2\} \quad (6)$$

(7)

We can merge all and find out that: (8)

$$f_{X_1|X_2} := P(X_1 = x_1 | X_2 = x_2) = \frac{f_{X_1X_2}(x_1, x_2)}{f_{X_2}(x_2)} \quad (9)$$

[9] like.. X $\perp$ Y

[10] This is true for both discrete and continuous RVs and it is due to the fact that the probability of the intersection of two indipendent events is the product of their probabilities.

Exercise:

$$Y = 2X + \epsilon, \quad X \perp \epsilon$$

[11]

What is the distribution of $Y|X$?

$$P(Y = y|X = x) = P(2X + \epsilon = y|X = x) \tag{10}$$
$$= P(\epsilon = y - 2x|X = x) \tag{11}$$
$$= P(\epsilon = y - 2x) \tag{12}$$

5.3 Central Limit Theorem (CLT)

The Central Limit Theorem is perhaps the most beautiful and important result in all of probability theory. It explains why the normal distribution appears everywhere in nature and why it's so central to statistics.[12]

Here's the idea: take ANY distribution (weird, skewed, discrete, whatever), and if you:

1. Sample many values from it
2. Add them up (or take their average)
3. Repeat this process many times

The sums (or averages) will be approximately normally distributed, no matter how weird your original distribution was!

Intuitive Example: Imagine rolling a single die. The distribution is completely uniform (flat) - each outcome from 1 to 6 is equally likely. Now:

- Roll 2 dice and sum them $\rightarrow$ slightly bell-shaped distribution
- Roll 10 dice and sum them $\rightarrow$ more bell-shaped
- Roll 100 dice and sum them $\rightarrow$ almost perfectly normal!

This is CLT in action - the flat distribution becomes bell-shaped purely through summation.

[11] X is $\perp$ to ϵ means that they are independent.
[12] Spoiler: don't use CLT, especially at the exam!

5.3.1 The Formal Statement

Focus

Central Limit Theorem: Let $X_1, X_2, \ldots, X_n$ be independent and identically distributed random variables with:

- Mean: $E[X_i] = \mu$
- Variance: $\text{Var}(X_i) = \sigma^2 < \infty$

Then, as $n \to \infty$:

$$\frac{\sum_{i=1}^{n} X_i - n\mu}{\sqrt{n\sigma^2}} \xrightarrow{d} \mathcal{N}(0,1)$$

Or equivalently, for the sample mean $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$:

$$\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \xrightarrow{d} \mathcal{N}(0,1)$$

What does this mean in plain English?

- $\sum_{i=1}^{n} X_i - n\mu$ is "how far the sum is from its expected value"
- $\sqrt{n\sigma^2}$ is the standard deviation of the sum
- The fraction is the "standardized" sum
- $\xrightarrow{d}$ means "converges in distribution"
- The result approaches a standard normal distribution $\mathcal{N}(0,1)$

5.3.2 Three Equivalent Ways to State CLT

Version 1: Standardized Sum

$$Z_n = \frac{\sum_{i=1}^{n} X_i - n\mu}{\sqrt{n\sigma^2}} \xrightarrow{d} \mathcal{N}(0,1)$$

Version 2: Sample Mean

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow{d} \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$$

Version 3: Approximation for Large n For large n:

$$\sum_{i=1}^{n} X_i \approx \mathcal{N}(n\mu, n\sigma^2)$$

All three versions say the same thing - they just focus on different aspects!

Think about what happens when you add random variables: When you add many random variables, extreme values in opposite directions tend to cancel out. If one variable is unusually large, others are likely to be smaller, so the sum becomes more "average." The sample mean $\bar{X}_n = \frac{1}{n}\sum X_i$ becomes more concentrated around μ as n increases. The variance of the sample mean is $\frac{\sigma^2}{n}$, which shrinks as n grows.

The beautiful mathematical reason involves moment generating functions and the fact that products of MGFs (which correspond to sums of random variables) converge to the MGF of a normal distribution.

5.3.3 Practical Applications of CLT

1. Approximating Discrete Distributions

Focus
Binomial Approximation: If $X \sim$ Binomial(n, p) with large n: $$X \approx \mathcal{N}(np, np(1-p))$$ **Rule of thumb:** Use this when $np \geq 5$ and $n(1-p) \geq 5$.

Example: You flip a coin 1000 times. Instead of computing $P(X = 520)$ using the exact binomial formula (which is computationally intensive), you can use:

$$X \approx \mathcal{N}(500, 250)$$

and compute $P(519.5 < X < 520.5)$ using normal probabilities.

2. Poisson Approximation

Focus
If $X \sim$ Poisson(λ) with large λ (typically $\lambda \geq 30$): $$X \approx \mathcal{N}(\lambda, \lambda)$$

3. Confidence Intervals and Hypothesis Testing CLT is the foundation for:

- Confidence intervals for population means
- t-tests and z-tests
- Most of classical statistics!

5.3.4 The Continuity Correction

When approximating discrete distributions with continuous normal distributions, we need a small adjustment.

The Problem: - Discrete: $P(X = k)$ can be positive - Continuous: $P(X = k) = 0$ always

The Solution: Replace $P(X = k)$ with $P(k - 0.5 < X < k + 0.5)$.

Focus
Continuity Correction Rules: • $P(X = k) \to P(k - 0.5 < X < k + 0.5)$ • $P(X \leq k) \to P(X < k + 0.5)$ • $P(X \geq k) \to P(X > k - 0.5)$ • $P(a \leq X \leq b) \to P(a - 0.5 < X < b + 0.5)$

Detailed Example: Coin Flipping You flip a coin 1000 times. What's P(exactly 520 heads)?

Step 1: Set up the approximation

$$X \sim \text{Binomial}(1000, 0.5) \approx \mathcal{N}(500, 250)$$

Step 2: Apply continuity correction

$$P(X = 520) \approx P(519.5 < X < 520.5)$$

Step 3: Standardize

$$Z = \frac{X - 500}{\sqrt{250}} \approx \frac{520 - 500}{\sqrt{250}} = \frac{20}{15.81} \approx 1.265$$

Step 4: Calculate

$$P(519.5 < X < 520.5) = P(1.233 < Z < 1.297) \approx 0.0156$$

5.3.5 When Does CLT Work and When Doesn't It?

CLT Works When:

- Variables are independent
- Variables are identically distributed
- The variance is finite
- Sample size is "large enough" (typically $n \geq 30$)

CLT Struggles When:

- Heavy-tailed distributions (infinite variance)
- Strong dependencies between variables
- Very small sample sizes
- Extremely skewed distributions (need very large n)

5.3.6 "Speed" of Convergence: Berry-Esseen Th.

How fast does CLT kick in? The Berry-Esseen theorem gives us a quantitative answer:

Focus
If $E[\|X\|^3] < \infty$, then: $$\left\| P\left(\frac{\sum X_i - n\mu}{\sqrt{n\sigma^2}} \leq x \right) - \Phi(x) \right\| \leq \frac{C \cdot E[\|X - \mu\|^3]}{\sigma^3 \sqrt{n}}$$ where $C \approx 0.4748$ and Φ is the standard normal CDF.

What this means: The error in the normal approximation decreases like $\frac{1}{\sqrt{n}}$. To cut the error in half, you need 4 times as much data.

5.3.7 CLT Variants and Extensions

Lindeberg-Lévy CLT: The standard version we've discussed.

Lyapunov CLT: For independent but not identically distributed variables.

Lindeberg-Feller CLT: The most general version.

Functional CLT (Donsker's Theorem): For stochastic processes.

5.3.8 Why CLT is So Important

1. Universality: Works for ANY distribution with finite variance.

2. Foundation of Statistics: Enables inference about populations from samples.

3. Explains Nature: Why measurement errors, heights, and many natural phenomena are normally distributed.

4. Computational Power: Allows us to replace difficult calculations with easy normal probability computations.

5. Quality Control: Manufacturing processes rely on CLT for quality assurance.

The Central Limit Theorem is truly one of the most elegant and powerful results in mathematics, connecting the abstract world of probability theory with the practical world of data analysis and scientific inference.

5.4 Sum of i.i.d. Random Variables: Special Cases and 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 $Y \sim \mathcal{N}(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$.

5.5 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.

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.

Part II

Learning from Data

Question: What is the average height of the population in a specific country?

Experiment: Pick a random person and measure the height. The set of all the *human in a specific country* is Ω.
Doing this, we made a continuos random variable:

$$F_X(x) = \int_{-\infty}^{x} f_X(t)ds$$

Where $f_X(t)$ is the probability density function of the height of a random person in that country.

We could also made it discrete saying that X is the height in cm rounded to the nearest integer, so $X \in \{140, 141, 142, \ldots, 200\}$.

We can also make some assumptions:

$$P(X < 0) = 0 \quad \iff \quad P(X \geq 0) = 1$$

$$P(X > 300) = 0 \quad \iff \quad P(X \leq 300) = 1$$

Now we want to estimate the average height of the population as said above.

$$E[X]$$

Next step: we collect data. $X_1, X_2, \ldots, X_n$ are the heights of n random people in that country. We can use this data to estimate the average height of the population. Then we can intuitively say that:

$$E[X] \approx \frac{1}{n} \sum_{i=1}^{n} X_i$$

But, saying this we are assuming the following things:

- The height of the people are independent each other. There is not correlation (that, for instance, is not true, see genetics).
- The heights of people are identically distributed. They all follow the same distribution ($FX_X(x_i) = FX_X(x_j)$ for all i, j).

This is called the **independent and identically distributed** assumption, or i.i.d. assumption.[13]

Now let's talk about error threshold. We want to be sure that our estimate is accurate enough. So we define an error threshold ε and we want to be

[13] We already see this in this handbook but, i mean..

sure that the probability that our estimate is within this threshold is high enough.

$$P(|\frac{1}{n}\sum_{i=1}^{n} X_i - E[X]| > \varepsilon)$$

We observe that at the rise of n the probability that our estimate is within the error threshold ε increases. That yells that the more data we collect, the more accurate our estimate is.

6 Concentration Inequalities

When we collect data and make estimates, we want to know: "How close is our estimate to the true value?" Concentration inequalities help us answer this question by giving bounds on how far our estimates can deviate from their expected values.

6.1 Why Do We Need Concentration Inequalities?

Let's go back to our height measurement example. We want to estimate the average height $E[X]$ of a population using sample data $X_1, X_2, \ldots, X_n$. Our estimate is:

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

The big question is: "How confident can we be that $\hat{\mu}$ is close to the true mean $E[X]$?"

This is where concentration inequalities come in. They tell us things like:

- "With 95% probability, our estimate is within ε of the true value"
- "If we collect more data, our estimate gets more accurate"
- "Here's how much data we need to achieve a certain accuracy"

6.2 Essential Tools: Indicator Functions and Moment Generating Functions

Before diving into inequalities, let's understand two key tools:

6.2.1 Indicator Functions

An indicator function is like a simple yes/no question turned into math:

$$\mathbb{I}_A(x) = \begin{cases} 1, & \text{if } x \in A \\ 0, & \text{if } x \notin A \end{cases}$$

For example, $\mathbb{I}_{X \geq 5}$ equals 1 if $X \geq 5$ and 0 otherwise. The cool thing is:

$$E[\mathbb{I}_{X \geq 5}] = P(X \geq 5)$$

This connection between expectations and probabilities is super useful in proofs.

6.2.2 Moment Generating Functions (MGF)

A moment generating function is like a "fingerprint" for a random variable:

$$M_X(t) = E[e^{tX}]$$

Think of it as encoding all the information about a random variable in a single function. The MGF is useful because:

- It helps us prove concentration inequalities
- Different distributions have different MGFs
- If we know the MGF, we can find all the moments (mean, variance, etc.)

6.3 Fundamental Concentration Inequalities

6.3.1 Markov's Inequality: basic one

Focus
Markov's Inequality: For any non-negative random variable X and any $\varepsilon > 0$: $$P(X \geq \varepsilon) \leq \frac{E[X]}{\varepsilon}$$

Intuition: If the average value is small, it's unlikely to see very large values. For example, if the average grade on an exam is 60, it's impossible for more than 60% of students to score above 100.

Proof: Here's the elegant proof using indicator functions:

$$E[X] = E[X \cdot \mathbb{I}_{X \geq \varepsilon} + X \cdot \mathbb{I}_{X < \varepsilon}] \tag{13}$$
$$\geq E[X \cdot \mathbb{I}_{X \geq \varepsilon}] \quad \text{(since } X \geq 0\text{)} \tag{14}$$
$$\geq E[\varepsilon \cdot \mathbb{I}_{X \geq \varepsilon}] \quad \text{(since } X \geq \varepsilon \text{ when } \mathbb{I}_{X \geq \varepsilon} = 1\text{)} \tag{15}$$
$$= \varepsilon \cdot P(X \geq \varepsilon) \tag{16}$$

Rearranging gives us the inequality.

Example: If you know that people spend an average of \$50 per month on coffee, then at most 10% of people spend more than \$500 per month (since $50/500 = 0.1$).

6.3.2 Chebyshev's Inequality: (Variance)

Focus
Chebyshev's Inequality: For any random variable X with mean μ and variance σ^2, and any $k > 0$: $$P(\lvert X - \mu \rvert \geq k\sigma) \leq \frac{1}{k^2}$$

Intuition: Most of the probability mass is concentrated around the mean. The more standard deviations you go away from the mean, the less likely you are to find values there.

Key Insight: Chebyshev is much stronger than Markov because it uses variance information. It tells us:

- At least 75% of values are within 2 standard deviations of the mean
- At least 89% of values are within 3 standard deviations of the mean
- At least $(1 - 1/k^2) \times 100\%$ of values are within k standard deviations

Proof Sketch: Apply Markov's inequality to $(X - \mu)^2$:

$$P(|X - \mu| \geq k\sigma) = P((X - \mu)^2 \geq k^2\sigma^2) \leq \frac{E[(X - \mu)^2]}{k^2\sigma^2} = \frac{\sigma^2}{k^2\sigma^2} = \frac{1}{k^2}$$

6.3.3 Variance of Sums: A Crucial Building Block

Before we get to Hoeffding's inequality, we need to understand how variance behaves with sums:

For independent random variables X and Y:

$$\begin{aligned} \text{Var}(X + Y) &= E[(X + Y - E[X + Y])^2] && (17) \\ &= E[((X - E[X]) + (Y - E[Y]))^2] && (18) \\ &= \text{Var}(X) + \text{Var}(Y) + 2\text{Cov}(X, Y) && (19) \end{aligned}$$

Since X and Y are independent, $\text{Cov}(X, Y) = 0$[14], so:

$$\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y)$$

This extends to n independent variables:

$$\text{Var}\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} \text{Var}(X_i)$$

6.3.4 Hoeffding's Inequality: (unbounded)

Focus

Hoeffding's Inequality: Let $X_1, X_2, \ldots, X_n$ be independent random variables with $X_i \in [a_i, b_i]$. Then for any $\varepsilon > 0$:

$$P\left(\left|\frac{1}{n}\sum_{i=1}^{n} X_i - E\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right]\right| \geq \varepsilon\right) \leq 2\exp\left(-\frac{2n\varepsilon^2}{(b-a)^2}\right)$$

where $b = \max_i b_i$ and $a = \min_i a_i$.

Intuition: If your random variables are bounded (like survey responses on a 1-10 scale), then your sample average concentrates very tightly around the true average. The bound decays exponentially fast as you collect more data.

[14] Covariance measures how two variables change together. Independence means knowing one variable tells you nothing about the other.

Why is this amazing?

- **Exponential decay**: The probability of large deviations drops super fast
- **Dimension-free**: Works for any number of variables
- **Non-asymptotic**: Valid for any sample size n

Practical Application - Confidence Intervals: If we want our estimate to be within ε of the true value with probability at least $1 - \delta$, we need:

$$2\exp\left(-\frac{2n\varepsilon^2}{(b-a)^2}\right) \leq \delta$$

Solving for ε:

$$\varepsilon = (b-a)\sqrt{\frac{\ln(2/\delta)}{2n}}$$

So our confidence interval is:

$$\left[\frac{1}{n}\sum_{i=1}^{n} X_i - \varepsilon, \frac{1}{n}\sum_{i=1}^{n} X_i + \varepsilon\right]$$

Example: You're surveying customer satisfaction on a 1-10 scale. With $n = 1000$ responses and wanting 95% confidence ($\delta = 0.05$):

$$\varepsilon = 9 \times \sqrt{\frac{\ln(2/0.05)}{2 \times 1000}} = 9 \times \sqrt{\frac{3.69}{2000}} \approx 0.39$$

Your sample average will be within 0.39 points of the true average with 95% probability.

6.4 Sub-Gaussian Random Variables: Beyond Normal Distributions

Normal distributions are great, but many real-world variables behave "almost normal" without being exactly normal. Sub-Gaussian random variables capture this idea.

6.4.1 Definition and Intuition

Focus

A random variable X is **sub-Gaussian** with parameter σ^2 if for all $t > 0$:

$$P(|X - E[X]| \geq t) \leq 2\exp\left(-\frac{t^2}{2\sigma^2}\right)$$

We write: $X \sim \mathsf{SG}(\sigma^2)$

Intuition: A sub-Gaussian random variable has "tails that are at most as heavy as a Gaussian." In other words:

- Extreme values are rare (exponentially rare)
- The variable is "well-behaved" like a normal distribution
- But it doesn't have to be exactly normal

Examples of Sub-Gaussian Variables:

- Any bounded random variable (like Bernoulli or Uniform)
- Normal distributions (obviously)
- Many other "nice" distributions

6.4.2 The Sub-Gaussian Notation Explained

The notation $X \sim \mathsf{SG}(C\sigma^2)$ means:

- X is sub-Gaussian
- C is a constant that controls how "sub-Gaussian" it is
- σ^2 is like a variance parameter
- The tails decay at least as fast as a Gaussian with variance $C\sigma^2$

6.4.3 Why Do We Care About Sub-Gaussian Variables?

Sub-Gaussian variables are incredibly useful because:

- They give us concentration inequalities similar to Hoeffding's
- Many machine learning algorithms work well with sub-Gaussian data
- They provide a unified framework for analyzing different types of random variables

6.4.4 Sub-Gaussian Concentration

If $X_1, \ldots, X_n$ are independent sub-Gaussian variables with parameter σ^2, then:

$$P\left(\left|\frac{1}{n}\sum_{i=1}^{n} X_i - E\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right]\right| \geq \varepsilon\right) \leq 2\exp\left(-\frac{n\varepsilon^2}{2\sigma^2}\right)$$

This is very similar to Hoeffding's inequality but applies to a broader class of distributions.

6.5 Sub-Exponential Random Variables: When Variance Matters More

Sometimes we deal with random variables that have heavier tails than sub-Gaussian ones, but are still "controlled." These are sub-exponential variables.

6.5.1 Definition

Focus

A centered random variable X (meaning $E[X] = 0$) is **sub-exponential** with parameter σ if:

$$E[e^{tX}] \leq e^{\frac{t^2\sigma^2}{2}}$$

for all $|t| \leq \frac{1}{\sigma}$.
We write: $X \sim \mathrm{SE}(\sigma^2)$

Intuition: Sub-exponential variables are "in between" sub-Gaussian and really heavy-tailed distributions. Think of them as having:

- Exponential-like tails (heavier than Gaussian)
- But still controlled enough for concentration inequalities
- Often arise when dealing with squares of sub-Gaussian variables

Examples:

- Exponential distributions
- Squares of sub-Gaussian variables: if $X \sim \text{SG}(\sigma^2)$, then X^2 is often sub-exponential
- Chi-squared distributions

6.5.2 Why Sub-Exponential Variables Matter

In machine learning and statistics, we often encounter:

- Squared errors: $(Y - \hat{Y})^2$
- Loss functions that involve squares
- Variables that are products or ratios of other variables

These naturally lead to sub-exponential distributions, so understanding them is crucial for theoretical analysis.

6.6 Summary: The Concentration Inequality Hierarchy

Here's how our inequalities compare in terms of strength and requirements:

Inequality	**Requirements**	**Bound Quality**	**Rate**
Markov	$X \geq 0$	Weak	$1/\varepsilon$
Chebyshev	Finite variance	Medium	$1/\varepsilon^2$
Hoeffding	Bounded variables	Strong	$\exp(-n\varepsilon^2)$
Sub-Gaussian	Light tails	Strong	$\exp(-n\varepsilon^2)$

The Big Picture:

- **Markov**: Works for any non-negative variable, but gives weak bounds
- **Chebyshev**: Uses variance info, gives polynomial bounds
- **Hoeffding**: Requires boundedness, gives exponential bounds
- **Sub-Gaussian**: Generalizes to more distributions, still exponential bounds

The key insight is that as we add more assumptions about our random variables (boundedness, light tails, etc.), we get much stronger concentration results. This is the fundamental trade-off in probability theory: more assumptions $\rightarrow$ stronger conclusions.

7 Discrete Markov Chains

7.1 Introduction

Markov Chains [15] 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!".

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:

[15] From now on: "MC"

- 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[16].

[16] A characteristic of Stochastic processes is that the sum of the probabilities of all possible outcomes must be equal to 1.

- **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[17]. 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.

  - **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[18].

    * **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.

[17] The longer the time the more accurate the stationary distribution is.
[18] Well..

* **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[19].
* **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:

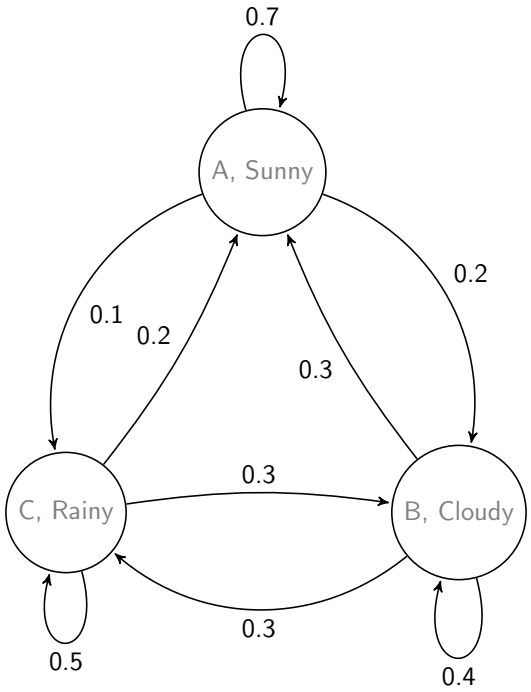
$$P = \begin{bmatrix} 0.7 & 0.2 & 0.1 \\ 0.3 & 0.4 & 0.3 \\ 0.2 & 0.3 & 0.5 \end{bmatrix}$$

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.

[19] 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[20] 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.
- 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:

[20] 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.

- 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.[21]

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

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[22].

Coming back to out toy MC: The stationary distribution π is the solution to the equation:[23]

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

[21] AKA the normalizaiton constraint.

[22] ..And we know that the sum of the probabilities of all possible outcomes must be equal to 1.

[23] 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 Continuos Markov Chains – WIP

...
...

10 Risk and statistical models

Definition of Statistical Model: A statistical model is a set of "admissible" distributions.

$$\mathcal{F} := \{f(x;\theta) : \theta \in \Theta\}$$

Where:

- $\mathcal{F}$ is the set of all possible distributions in the model
- $f(x;\theta)$ is a probability density function (pdf) or probability mass function (pmf) parameterized by θ
- Θ is the parameter space, which is the set of all possible values for θ

Example:

$$N(0, sigma^2)$$

$$\mathcal{F} := \{\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{x^2}{2\sigma^2}}; \sigma \in \mathbb{R}\}$$

Trivial model:

$\mathcal{F} := \{$all Distribution Functions F increasing $\lim_{x\to-\infty} F(x) = 0, \lim_{x\to+\infty} F(x) = 1\}$

A statistical model represents our assumptions and beliefs about the data-generating process.

In simple terms: we believe that there exists some true, unknown distribution $f^* \in \mathcal{F}$ that generated our observed data $X_1, X_2, \ldots, X_n$. These data points are independent and identically distributed (i.i.d.) samples from f^*.

The fundamental challenge: We don't know what f^* is exactly - we only know it belongs to our model family $\mathcal{F}$.

10.0.1 Types of Statistical Models

Statistical models can be classified based on how many parameters they have:

Focus

Parametric vs Non-Parametric Models:

1. **Parametric Model**: If the parameter space Θ has finite dimension d (i.e., $\Theta \subset \mathbb{R}^d$ for some fixed d), we call $\mathcal{F}$ a d-dimensional parametric model.
2. **Non-Parametric Model**: If the parameter space is infinite-dimensional, the model is non-parametric.

Examples: Parametric Model - Normal Distribution:

$$\mathcal{F} = \left\{ \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} : (\mu, \sigma^2) \in \mathbb{R} \times \mathbb{R}^+ \right\}$$

This is a 2-dimensional parametric model since $\Theta = \mathbb{R} \times \mathbb{R}^+$ has dimension 2.

Non-Parametric Model - All Distributions:

$\mathcal{F} = \{$all distribution functions $F : \mathbb{R} \to [0, 1]$ that are non-decreasing, right-continuous, with $\lim_{x \to -\infty} F(x) =$

This includes infinitely many possible distributions, making it non-parametric.

10.1 Risk Functions and Loss

In statistics, we want to estimate something about the true distribution f^*. But how do we measure how "good" our estimate is? This is where risk functions come in.

10.1.1 Loss Functions

A loss function $L(\theta, \hat{\theta})$ measures the "cost" of estimating the true parameter θ with our estimate $\hat{\theta}$.

Common Loss Functions:

- **Squared Loss**: $L(\theta, \hat{\theta}) = (\theta - \hat{\theta})^2$

  Heavily penalizes large errors

  Mathematically convenient (differentiable)

  Most common in regression problems

- **Absolute Loss**: $L(\theta, \hat{\theta}) = |\theta - \hat{\theta}|$

  Less sensitive to outliers than squared loss

  Used in robust statistics

- **0-1 Loss**: $L(\theta, \hat{\theta}) = \mathbf{1}_{\{\theta \neq \hat{\theta}\}}$

  Used in classification problems

  Only cares about getting the answer exactly right

10.1.2 Risk Function

The risk function measures the expected loss when using estimator $\hat{\theta}$ to estimate the true parameter θ:

$$R(\theta, \hat{\theta}) = E_\theta[L(\theta, \hat{\theta})]$$

Interpretation:

- $R(\theta, \hat{\theta})$ tells us how well our estimator $\hat{\theta}$ performs on average when the true parameter is θ
- Lower risk = better estimator
- The expectation $E_\theta[\cdot]$ is taken over all possible datasets that could be generated from the true distribution with parameter θ

Example - Squared Error Risk: If we use squared loss, the risk becomes:

$$R(\theta, \hat{\theta}) = E_\theta[(\theta - \hat{\theta})^2]$$

This is called the Mean Squared Error (MSE) and can be decomposed as:

$$\text{MSE} = \text{Bias}^2 + \text{Variance}$$

where:

- Bias $= E_\theta[\hat{\theta}] - \theta$ (systematic error)
- Variance $= \text{Var}_\theta(\hat{\theta})$ (variability of estimates)

10.2 The Estimation Problem

Given data $X_1, \ldots, X_n$ from some $f^* \in \mathcal{F}$, we want to:

1. Choose an estimator $\hat{\theta}(X_1, \ldots, X_n)$
2. Minimize the risk $R(\theta^*, \hat{\theta})$ where θ^* is the true parameter

The Challenge: We don't know θ^*, so we can't directly compute the risk!

This leads us to different approaches:

- **Minimax approach**: Choose $\hat{\theta}$ that minimizes $\max_{\theta \in \Theta} R(\theta, \hat{\theta})$
- **Bayes approach**: Put a prior distribution on θ and minimize expected risk
- **Asymptotic approach**: Study behavior as $n \to \infty$

10.3 Connection to Machine Learning

Everything we've discussed connects directly to modern machine learning:

- **Statistical Model $\mathcal{F}$ $\leftrightarrow$ Model Class** (e.g., neural networks, decision trees)
- **True Distribution f^* $\leftrightarrow$ Data Generating Process**
- **Parameter θ $\leftrightarrow$ Model Weights/Parameters**
- **Risk Function $\leftrightarrow$ Expected Loss/Generalization Error**
- **Empirical Risk $\leftrightarrow$ Training Loss**

The fundamental questions remain the same: How do we choose good models? How do we measure performance? How do we balance bias and variance? These concepts form the theoretical foundation for all of modern data science and machine learning.

11 Supervised learning

11.1 Classic Supervised Learning Framework

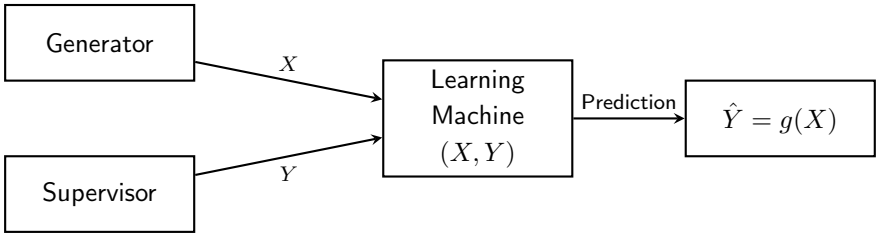


Figure 1: Supervised Learning Framework with Generator and Supervisor

We have a generator and a supervisor:

- The **generator** has a distribution f_X and F_X.
- The **supervisor** provides the target values by trying to learn something about $f_{Y|X}$.

For the classical example: iris dataset:

- The generator is the distribution of features f_X, probably a person that selected the flowers.
- The supervisor is the set of target names (species).

11.2 Loss function - basic idea

What makes a good prediction? Let's say our learning machine has 2 possible outputs: $\hat{Y} \in \{a, b\}$, and a true answer y. This is usually done with a loss function $L(y, \hat{y})$.

$$L : \mathbb{R}^2 \to \mathbb{R}$$

Higher value of L means worse prediction. We have to make a choice:

$$L(y, a) \overset{?}{<} L(y, b)$$

And we should choose the smallest one.

Example of loss function:

- Least squares: $L(y, \hat{y}) = (y - \hat{y})^2$
- Absolute error: $L(y, \hat{y}) = |y - \hat{y}|$
- 0-1 loss: $L(y, \hat{y}) = \mathbb{I}(y \neq \hat{y})$
- Log loss (for probabilistic outputs): $L(y, \hat{p}) = -\log(\hat{p}_y)$ where $\hat{p}_y$ is the predicted probability for the true class y.
- Exponential loss (used in boosting): $L(y, \hat{y}) = e^{-y\hat{y}}$ where $y \in \{-1, 1\}$.
- Quadratic loss: $L(y, \hat{y}) = (y - \hat{y})^2$.
- ....

11.3 Risk function - basic idea

We pick a loss, then we can define a risk:

$$R(g) := \mathbb{E}[L(Y, g(X))]$$

Where:

- g is our prediction function (learning machine).
- L is the loss function.
- The expectation is over the joint distribution of (X, Y).

Basically the risk takes our prediction function g, and tells us how well it performs on average, according to the loss function L. For example let's say that in a linear regression problem we have two models: $g^1(X)$ and $g^2(X)$. We can compute the risk for both models:

$$R(g^1) = \mathbb{E}[L(Y, g^1(X))]$$

$$R(g^2) = \mathbb{E}[L(Y, g^2(X))]$$

We can decide which model is the best by comparing their risks, and on a practical side we are actually deciding whic *line* on our graph to choose. We want to minimize the risk:

$$g^* = \arg\min_g R(g)$$

The Learning Machine does not know the *real* $f_{X,Y}$:

$$((X,Y),...(X_n,Y_n)) \sim f_{X,Y}$$

Risk can be calculated in an empirical way:

Empirical Risk:

$$\hat{R}(g) = \frac{1}{n}\sum_{i=1}^{n} L(Y_i, g(X_i))$$

We can say that **the real goal of the learning machine is to minimize the empirical risk**.

$$\hat{g_n} := \arg\min_g \hat{R}(g) \quad = \quad \arg\min_g \frac{1}{n}\sum_{i=1}^{n} L(Y_i, g(X_i))$$

This process is conceptually what happens when we write:

```
model.fit(X, y)
```

The `fit` method searches through the space of possible functions g to find the one that minimizes the empirical risk $\hat{R}(g)$ on our training data.

11.3.1 $\mathcal{M}$ - the Model Space

The model space $\mathcal{M}$ is the set of all possible functions g that our learning machine can choose from. But if $\mathcal{M}$ is the space of all possible functions, does there really exist a function that minimizes the risk? Does the "best" function actually exist?

`Spoiler:` The choice of $\mathcal{M}$ involves the fundamental bias-variance tradeoff

11.3.2 Generalization Gap

The generalization gap is the difference between the empirical risk $\hat{R}(g)$ and the true risk $R(g)$ for a given function g.

$$\text{Generalization Gap} = R(\hat{g}_n) - \hat{R}_n(\hat{g}_n)$$

In simple words is the difference (the gap) between the training loss[24] and the testing loss[25].

How could we estimate the generalization gap? We may divide our dataset into two parts: training set and test set. We use the training set to fit our model (minimize empirical risk), and then we evaluate the performance of the fitted model on the test set to estimate the true risk

$$
\begin{aligned}
D_{Tr} &:= \{(X_1, Y_1), ..., (X_n, Y_n)\} \sim f_{X,Y} \\
D_{Te} &:= \{(X_{n+1}, Y_{n+1}), ..., (X_{n+m}, Y_{n+m})\} \sim f_{X,Y} \\
\hat{g}_n &:= \arg\min_g \hat{R}_n(g) = \arg\min_g \frac{1}{n} \sum_{i=1}^{n} L(Y_i, g(X_i))
\end{aligned}
$$

Where:

- D_{Tr} is the training dataset, used to fit the model.
- D_{Te} is the test dataset, used to evaluate the model's performance.
- $\hat{g}_n$ is the function that minimizes the empirical risk on the training data.

Going back to our question: *how do we estimate the generalization gap?* We can estimate the generalization gap in an empirical way:

$$\frac{1}{m} \sum_{i=1}^{m} L\left(Y_{n+i}, \hat{g}_n(X_{n+i})\right)$$

This should be a good proxy for the true risk $R(\hat{g}_n)$, and we can use it to estimate the generalization gap.

And, what is the true testing error? The testing is a conditional expectation:

$$R(\hat{g}_n) = \mathbb{E}[L(Y, \hat{g}_n(X))|D_{Tr}]$$

$$R(\hat{g}_n) \neq \mathbb{E}[L(Y, \hat{g}_n(X))]$$

26

[24] Training loss is the empirical risk, that is the performance of the model on the training data, computed using the empirical risk $\hat{R}_n(g)$.

[25] Testing loss is the true risk. Testing loss is the performance of the model on unseen data, computed using the true risk $R(g)$.

[26] The testing error is not the expectation over the joint distribution of (X, Y) in its entirety, but rather the expectation conditioned on the training data D_{Tr}.

11.4 Statistics (the functions on data)

A statistic is just a function on data. For example the common T-Statistic is a... *statistic*.

Let say we have this T statistic:

$$T : X_n \to \mathbb{T}$$

An estimator is a statistic intended to estimate "something". For example, the sample mean is an estimator of the population mean:

$$\frac{1}{n}\sum_{i=1}^{n} X_i \sim \mathbb{E}[X]$$

The empirical mean, is a statistic, and an estimator of the population expected value. We can estimate the expected value of X by computing the empirical mean on a sample of X and assume that it is a good approximation of the true expected value.

We have now what is called a **bias**:

$$\text{Bias}(\hat{\theta}) = \mathbb{E}[hat\theta] - \theta$$