Probability and Statistics – Lecture Notes

Han-Miru Kim

July 16, 2021

About

Lecture notes taken from the Wahrscheinlichkeit und Statistik lecture given by Dr. Josef Teichman during Spring Semester 2021 at ETH Zürich.

Material

There is a jupyter notebook which can be found at Teichman's Github.
The slides will be uploaded at the end of every week on his website which are accessed via the password sent per mail at the beginning of the semester.
A new scriptum will be published at the end of the semester. Until then, the old script can be used as they shouldn't differ too much.

1 Probability theory

1.1 Introduction

Although even the greeks and 17-th century mathematicians a mathematical framework to discuss probability and statistics was only developed in the 19-th and 20-th century.
One reason this might have been the case could be the empirical nature of statistics. Probabily theory was dependent on data and was difficult to axiomatize and whose proofs often relied on human intuition over formal derivation.
One difficulty was that in order to find a suitable notion of probability on an infinite dimensional vector space. For example, to think about the probability that a particle takes any given Brownian motion path is to find a notion of measure on the set $C([0,\infty),\mathbb{R})$.
As a consequence, the first fields medal that was awarded for a contribution to probability was in 2002 and 2006. Almost 70 years after the first medal was awarded.
So the axiomatisation was only done in the early 20-eth century and reached the center of mathematics in the 21-st century.

1.2 Probability Space

Definition 1.1 (Probability space)**.** A **Probability space** (or P-space) is a triple $(\Omega,\mathcal{A},\mathbb{P})$ consisting of a non-empty set Ω, a σ-Algebra $\mathcal{A}\subseteq 2^{\Omega}$, and a σ-additive probability measure $\mathbb{P}:\mathcal{A}\to[0,1]$. This means that the following are satisfied:

(a) $\Omega \in \mathcal{A}$

(b) $\forall A \in \mathcal{A} : A^c \in \mathcal{A}$

(c) $\forall (A_i)_{i\in\mathbb{N}}, A_i \in \mathcal{A}$, $\bigcup_{i\in\mathbb{N}} A_i \in \mathcal{A}$

(d) $\mathbb{P}[\emptyset] = 0, \mathbb{P}[\Omega] = 1$

(e) $\forall (A_i)_{i\in\mathbb{N}}, A_i \in \mathcal{A}$, if all A_i are disjoint, then $\mathbb{P}\left(\bigsqcup_{i\in\mathbb{N}} A_i\right) = \sum_{i\in\mathbb{N}} \mathbb{P}[A_i]$

It is useful (and sometimes necessary) to be able to talk about the "smallest measureable sets".

Definition 1.2. A subset $A \in \mathcal{A}$ of Ω is **atomic**, if

$$B \in \mathcal{A}, B \subseteq A \implies B = \emptyset \vee B = A$$

We denote the set of all atoms as $\text{Atom}(\mathcal{A})$. In particular, if $\mathcal{A} = \mathcal{P}(X)$, then the atoms are exactly the singletons. We can uniquely decompose any element $B \in \mathcal{A}$ into a disjoint union of atoms.

Remark 1.3. Often, when Ω is finite, it is useful to set $\mathcal{A} = \mathcal{P}(\Omega)$ i.e to let all subsets of Ω be measureable. In particular, the atoms are then exactly the singletons $\omega \in \Omega$. If that is the case, we can define a **weight** function

$$p : \Omega \to [0, 1], \quad p(\omega) := \mathbb{P}[\{\omega\}]$$

Then for any subset $A \subseteq \Omega$, its probability measure can be calculated as follows

$$\mathbb{P}[A] = \sum_{\omega\in A} p(\omega)$$

or in the general case, for $B \in \mathcal{A} \neq \mathcal{P}(\Omega)$:

$$\mathbb{P}[B] = \sum_{A\in\text{Atom}(\mathcal{A})} \mathbb{P}[A \cap B]$$

It can easily shown that the following axioms are satified:

Lemma 1.4. *Let $(\Omega, \mathcal{A}, \mathbb{P})$ be a P-space and $A, B \in \mathcal{A}$. Then*

(a) $\mathbb{P}[A^c] = 1 - \mathbb{P}[A]$

(b) $\mathbb{P}[A \cup B] = \mathbb{P}[A] + \mathbb{P}[B] - \mathbb{P}[A \cap B]$

(c) $\mathbb{P}\left(\bigcup_{i=1}^n A_i\right) = \sum_{k=1}^n (-1)^{k+1} \sum_{1\le i_1<\ldots<i_k\le n} \mathbb{P}[A_{i_1} \cap \ldots \cap A_{i_k}]$

(d) $A \subseteq B \implies \mathbb{P}[A] \le \mathbb{P}[B]$

Proof. We only show the proof for $\mathcal{A} = \mathcal{P}(\Omega)$. For generalized $\mathcal{A}$, the proof is more or less the same.

(a) By σ-additivity of $\mathbb{P}$, and $\mathbb{P}[\Omega] = 1$, we have

$$\mathbb{P}[\Omega] = \mathbb{P}[A \sqcup A^c] = \mathbb{P}[A] + \mathbb{P}[A^c] = 1$$

(b) With the decomposition $A \cup B = (A \setminus B) \sqcup (A \cap B) \sqcup (B \setminus A)$, we find

$$\mathbb{P}[A \cup B] + \mathbb{P}[A \cap B] = \mathbb{P}[A \setminus B] + \mathbb{P}[A \cap B] + \mathbb{P}[B \setminus A] + \mathbb{P}[A \cap B] = \mathbb{P}[A] + \mathbb{P}[B]$$

(c) We use induction over n. We find

$$\begin{aligned}\mathbb{P}\left[\bigcup_{i=1}^{n+1} A_i\right] &\overset{(b)}{=} \mathbb{P}\left[\bigcup_{i=1}^{n} A_i\right] + \mathbb{P}[A_{n+1}] - \mathbb{P}\left[\bigcup_{i=1}^{n} (A_i \cap A_{n+1})\right] \\ &\overset{\text{ind}}{=} \underbrace{\left(\sum_{k=1}^{n} (-1)^{k+1} \sum_{1 \leq i_1 < \ldots < i_k \leq n} \mathbb{P}[A_{i_1} \cap \ldots \cap A_{i_k}]\right)}_{=:\Sigma_1} + \mathbb{P}[A_{n+1}] \\ &\quad - \underbrace{\left(\sum_{m=1}^{n} (-1)^{m+1} \sum_{1 \leq i_1 < \ldots < i_m \leq n} \mathbb{P}[A_{n+1} \cap A_{i_1} \cap \ldots \cap A_{i_m}]\right)}_{=:\Sigma_2}\end{aligned}$$

Comparing this with the predicted result

$$\sum_{k=1}^{n+1} (-1)^{k+1} \sum_{1 \leq i_1 < \ldots < i_k \leq n+1} \mathbb{P}[A_{i_1} \cap \ldots \cap A_{i_k}]$$

we can see that they are the same:

- For $k = 1$, the summands $\mathbb{P}[A_1], \ldots, [A_n]$ are present in the first sum Σ_1 and $\mathbb{P}[A_{n+1}]$ is also present.
- For $k = n + 1$, there is only one term $\mathbb{P}[A_1 \cap \ldots \cap A_{n+1}]$, which is present in Σ_2 with sign $(-1)^{(n+1)+1}$.
- For $2 \leq k \leq n$, we can either chose k elements from $A_1, \ldots, A_n$ (a summand in Σ_1), or chose always chose A_{n+1} and $k - 1$ other elements from $A_1, \ldots, A_n$ (a summand in Σ_2).
  The negative sign before Σ_2 ensures that the term $\mathbb{P}[A_{n+1} \cap A_{i_1} \cap \ldots \cap A_{i_{k-1}}]$ has the correct sign $(-1)^{k+1}$

(d) This follows directly from $B = (B \setminus A) \sqcup A$.

□

Example 1.5 (Laplace Model)**.** To model many situations, we can set the space Ω such that every singleton has equal probability to occur. In the **Laplace Model**, we set

$$p(\omega) = \frac{1}{|\Omega|}$$

This obviously only works for finite Ω. For a subset $A \subseteq \Omega$, then $\mathbb{P}[A] = \frac{|A|}{|\Omega|}$.

Example 1.6 (Bernoulli-Experiment)**.** Consider the experiment where we measure the number of calls during a given time at a call center. So $\Omega = \{0, 1, 2, \ldots\}$. We can set the weights to be

$$p(\omega) = e^{-\lambda} \frac{\lambda^\omega}{\omega!}, \quad \text{for} \quad \omega \in \Omega$$

, where $\lambda > 0$ is some parameter. This model is called the **Poisson-Distribution**. Note that the above defintiion is well defined, as

$$\mathbb{P}[\Omega] = \sum_{n \in \mathbb{N}} e^{-\lambda} \frac{\lambda^n}{n!} = e^{-\lambda} \cdot e^{\lambda} = 1$$

If we let $A = \{1, 2, \ldots\}$ to be the outcome, where we get at least one call, then we see that

$$\mathbb{P}[A] = 1 - \mathbb{P}[A^c] = 1 - \mathbb{P}[\{0\}) = 1 - e^{-\lambda}$$

What is often very useful is to associate to each possible outcome $\omega \in \Omega$ a real value. This gives us the following definition.

Definition 1.7. A **random variable** is a real-valued function $X : \Omega \to \mathbb{R}$.
If the image $X(\Omega)$ is also countable, then we can turn the probability measure $\mathbb{P}$ on Ω into a probability measure on the image $X(\Omega)$ where for $x \in X(\Omega)$ we define

$$\mathbb{P}[X = x] := \mathbb{P}\left[\{\omega \in \Omega \big| X(\omega) = x\}\right] = \mathbb{P}[X^{-1}(x)]$$

which reads as: "The probability that the random variable X obtains the value $x \in X(\Omega)$ is the probability measure of the preimage of x under the function X."
In the case for general $\mathcal{A} \neq \mathcal{P}(\Omega)$, we must require that X be constant on every atom, or equivalently, that the preimage of singletons under X are $\mathcal{A}$-measureable.

$$\mathbb{P}[X = x] := \mathbb{P}\left[\{A \in \mathrm{Atom}(\Omega) \big| X(A) = x\}\right] = \mathbb{P}[X^{-1}(x)]$$

Other common notation for $\mathbb{P}[X = x]$ is $\mu_X(x), P(X = x), P(x), P_x$ etc.

Definition 1.8. For a random variable $X : \Omega \to \mathbb{R}$ we define the **expectation value** of X to be

$$\mathbb{E}[X] := \sum_{\omega \in \Omega} X(\omega)p(\omega) \qquad \left(\text{or} \quad \mathbb{E}[X] := \sum_{A \in \mathrm{Atom}(\Omega)} X(A)\mathbb{P}[A]\right)$$

If X can take on negative values, the right hand side can be problematic because we can run into non-absolute convergence. To remedy this, we can divide X into its positive and negative parts

$$X(\omega) = X^+(\omega) - X^-(\omega) = \max\{X(\omega), 0\} + \{\min\{0, X(\omega)\}$$

and then change the definition of the expectation as follows:

$$\mathbb{E}[X] := \mathbb{E}[X^+] - \mathbb{E}[X^-] = \sum_{X(\omega)>0} X(\omega)p(\omega) - \sum_{X(\omega)<0} \underbrace{-X(\omega)p(\omega)}_{\geq 0}$$

as long as both sums aren't infinite.

The expectation value can also be expressed in terms of the distribution of X:

$$\boxed{\mathbb{E}[X] = \sum_{x \in X(\Omega)} x \cdot \mathbb{P}[X = x]}$$

which is often easier to compute when we know how X is distributed.

Example 1.9. In the call-center example from before, let X be the number of calls, so $X(\omega) = \omega$. Then

$$\mathbb{E}[X] = \sum_{k \in \mathbb{N}} k\mathbb{P}[X = k] = \sum_{k=0}^{\infty} ke^{-\lambda}\frac{\lambda^k}{k!} = \lambda e^{-\lambda} \underbrace{\sum_{k=1}^{\infty} \frac{\lambda^{k-1}}{(k-1)!}}_{=e^{\lambda}} = \lambda$$

so the paramter λ gives us the expected number of calls.

Example 1.10. In an insurance contract the cost to the insurance company is

$$X = \begin{cases} c & \text{if the event } A \text{ occurs} \\ 0 & \text{else} \end{cases}$$

The premiums the insured have to pay should then be the expectation value

$$\mathbb{E}[X] = c \cdot \mathbb{P}[A] + 0 \cdot \mathbb{P}[A^c] = c \cdot \mathbb{P}[A]$$

More generally, we can define the **indicator function**

$$I_B(\omega) := \begin{cases} 1 & \text{for } \omega \in B \\ 0 & \text{for } \omega \notin B \end{cases}$$

for subsets $B \subseteq \Omega$, then we can write $X = cI_B$, so the expectation value is

$$\mathbb{E}[cI_B] = c\mathbb{P}[B] \quad \text{for} \quad c \in \mathbb{R}, B \in \mathcal{A}$$

This immediately gives us the following lemma

Lemma 1.11 (Linearity of the expectation value)**.** *Let $X, Y : \Omega \to \mathbb{R}$ be random variables and $a, b \in \mathbb{R}$. Then*

$$\mathbb{E}[aX + bY] = a\mathbb{E}[X] + bE[Y]$$

Proof. This follows directly from the definition of the expectation value:

$$\mathbb{E}[aX + bY] = \sum_{\omega \in \Omega} (aX + bY)(\omega)p(\omega) = a \sum_{\omega \in \Omega} X(\omega)p(\omega) + b \sum_{\omega \in \Omega} Y(\omega)p(\omega) = a\mathbb{E}[X] + b\mathbb{E}[Y]$$

□

Another useful result when calculating the expectation value is

Lemma 1.12. *If X only takes values in $\mathbb{N}$, then*

$$\mathbb{E}[X] = \sum_{j \in \mathbb{N}} \mathbb{P}[X > j]$$

Proof. We can use the representation using the distribution to find

$$\begin{aligned} \mathbb{E}[X] &= \sum_{k=1}^{\infty} k\mathbb{P}[X = k] \\ &= \sum_{k=1}^{\infty} \sum_{j=0}^{k-1} \mathbb{P}[X = k] \\ &= (\mathbb{P}[X = 1] + \mathbb{P}[X = 2] + \ldots) + (\mathbb{P}[X = 2] + \mathbb{P}[X = 3] + \ldots) + \ldots \\ &= \sum_{j=0}^{\infty} \sum_{k=j+1}^{\infty} \mathbb{P}[X = k] \\ &= \sum_{j=0}^{\infty} \mathbb{P}[X > j] \end{aligned}$$

□

1.3 Laplace Models

In this section, we look at examples where we can apply the Laplace models.
Recall that in the Laplace model, for a subset $A \subseteq \Omega$ we have

$$\mathbb{P}[A] = \frac{|A|}{|\Omega|}$$

Example 1.13 (Hat Problem)**.** We distribute n hats to n people. What is the probability that nobody receives their own hat. We set our event space Ω to be the set of all permutations S_n. Note that $|S_n| = n!$ For each $i \in \{1, \dots, n\}$, let A_i be the set of permutations that have i as a fixed point.

$$A_i = \{\omega \in \Omega | \omega(i) = i\}$$

If we set A to be the permutations that have any fixed point, then by Lemma 1.4

$$\begin{aligned}\mathbb{P}[A] = \mathbb{P}\left[\bigcup_{i=1}^n A_i\right] &= \sum_{k=1}^n (-1)^{k+1} \sum_{1 \le i_1 < \ldots < i_k \le n} \mathbb{P}\left[A_{i_1} \cap \ldots \cap A_{i_k}\right] \\ &= \sum_{k=1}^n (-1)^{k+1} \underbrace{\binom{n}{k} \frac{(n-k)!}{n!}}_{=\frac{1}{k!}} = -\sum_{k=1}^n \frac{(-1)^k}{k!}\end{aligned}$$

where we used the fact that there are $\binom{n}{k}$ ways to chose k indices $i_1, \dots, i_k$ and the probability that a random permutation fixes these k points is $\frac{(n-k)!}{n!}$.
So in the event that $n \to \infty$, the probability that there are no fixed points is

$$\mathbb{P}[A^c] = 1 - \mathbb{P}[A] = 1 + \sum_{k=1}^n \frac{(-1)^k}{k!} \overset{n\to\infty}{\longrightarrow} e^{-1}$$

Example 1.14 (Urn problems)**.** In an Urn we have N numbered balls, K of which are coloured red and $N - K$ white. We take a sample of n balls (with or without putting them back). If we set ω_i to be the number that is picked at the i-th step, then the event space Ω is

- With putting back: $\Omega_1 = \{(\omega_1, \dots, \omega_n) | 1 \le \omega_i \le N\}$
- Without putting back: $\Omega_2 = \{(\omega_1, \dots, \omega_n) | 1 \le \omega_i \le N, \omega_i \ne \omega_j \text{ for } i \ne j\}$

Then set $\mathbb{P}_i$ corresponding to an equal distribution on Ω_i. We're interested in the distribution of the random variable X, which measures the number of picked balls that are red. And let $A_{i,k}$ to be the number of samples that have exactly k red balls.

$$A_{i,k} = \{\omega \in \Omega_i || \{1 \le \omega_j \le K\}| = k\}$$

then the probability is simply $\mathbb{P}_i[X = k] = |A_{i,k}|/|\Omega_i|$ and we just have to find out the size of these sets.

- If the balls are returned into the Urn ($i = 1$) we have that $|\Omega_1| = N^n$, so if we set $p = \frac{K}{N}$ to be the proportion of red balls we have

$$|A_{1,k}| = K^k (N-K)^{n-k} \binom{n}{k}$$

$$\implies \mathbb{P}_1[X = k] = \binom{n}{k} p^k (1-p)^{n-k}$$

This is called the **Binomial distribution** with parameters n and p.

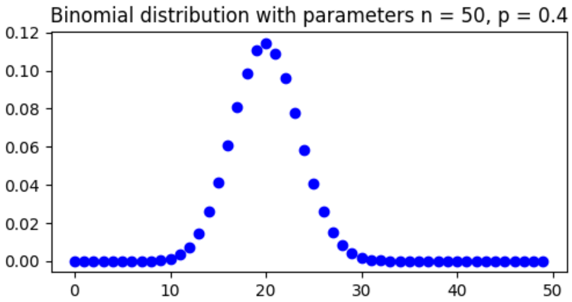

Figure 1: Plot of the Binomial distribution with parameters $n = 50, p = 0.4$. `./src/binomial.py`

- For $i = 2$ we have

$$|\Omega_2| = N \cdot (N-1)\dots(N-n+1) = \binom{N}{n} n!$$

$$|A_{2,k}| = K(K-1)\dots(K-k+1) \cdot (N-K)(N-K-1)\dots(N-K-(n-k)+1)\binom{n}{k}$$

$$= \binom{K}{k}\binom{N-K}{n-k} n!$$

and it follows that

$$\mathbb{P}_2[X=k] = \frac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{n}}$$

which is called the **hypergeometric** distribution with parameters n, N, K

Notice that for $p = \frac{K}{N}$ constant and n fixed, the hypergeometric distribution converges to the binomial distribution for N, K large enough because the removal of each ball has less and less effect on the probability to chose a red ball at any step.

1.4 Conditional Probability

For now, let $(\Omega, \mathcal{A}, \mathbb{P})$ be a discrete P-space and set $\mathcal{A} = \mathcal{P}(\Omega)$.

Definition 1.15. For $A, B \in \mathcal{A}$, the **conditional probability** of A given B is given by

$$\mathbb{P}[A|B] := \frac{\mathbb{P}[A \cap B]}{\mathbb{P}[B]}$$

In the "frequentist" interpretation, where we take n trials and set $\mathbb{P}[C] = \frac{n_c}{n}$ to be the relative frequency that C occurs, the conditional probability $\mathbb{P}[A|B]$ is equivalent to the relative probability that A occurs under those, where B occured, i.e

$$\mathbb{P}[A|B] \sim \frac{n_{A\cap B}/n}{n_B/n} = \frac{n_{A\cap B}}{n_B}$$

hence the name.

For a fixed $B \subseteq \Omega$, we obtain a new probability measure $\mathbb{P}[\cdot|B]$ with weights

$$p_B(\omega) = \begin{cases} C \cdot p(\omega) & \omega \in B \\ 0 & \omega \notin B \end{cases}$$

where C is some constant depending on B.

Example 1.16 (Two dice). Set $\Omega = \{(i,j)|1 \leq i,j \leq 6\}$ and $\mathbb{P}$ to be the equal distribution. If A_i is the event that the first dice shows i and B_k is the event that the sum of the die is k, then

$$\mathbb{P}[A_i|B_7] = \frac{\mathbb{P}[\text{sum is 7 and first throw is } i]}{\mathbb{P}[\text{sum is 7}]} = \frac{1/36}{1/6} = \frac{1}{6} = \mathbb{P}[A_i]$$

Using the formula for conditional probability, we immediately obtain the following theorem, which allows us to calculate the total probability of any event B if we understand the conditional probabilites given B.

Theorem 1.17 (Total proability theorem). *Let $(A_i)_{i\in I}$ be a partition of Ω, (i.e. $\bigcup_{i\in I} A_i = \Omega$ and $A_i \cap A_j = \emptyset$ for $i \neq j$). Then for any $B \subseteq \Omega$ we have*

$$\mathbb{P}[B] = \sum_{i\in I} \mathbb{P}[B \cap A_i] = \sum_{i \in I} \mathbb{P}[B|A_i]\mathbb{P}[A_i]$$

In the special case where our partition of Ω is $A \sqcup A^c$ we get

$$\mathbb{P}[B] = \mathbb{P}[B|A]\mathbb{P}[A] + \mathbb{P}[B|A^c](1 - \mathbb{P}[A])$$

By repeatedly using the definition of conditional probability, we get the following proposition for finite set of events

Proposition 1.18. *For a finite set of events $A_1, \ldots, A_n$ we have*

$$\mathbb{P}[A_1 \cap \cdots \cap A_n] = \mathbb{P}[A_1]\mathbb{P}[A_2|A_1]\mathbb{P}[A_3|A_1 \cap A_2] \ldots \mathbb{P}[A_n|A_1 \cap \cdots \cap A_{n-1}]$$

as long as $\mathbb{P}[A_1 \cap \cdots \cap A_n] > 0$.

Another consequence of our definition of conditional probability is Baye's forumla.

Corollary 1.18.1 (Bayes' Formula). *Given two events with* non-zero *probability, the following formula holds*

$$\mathbb{P}[B|A] = \frac{\mathbb{P}[A|B]\mathbb{P}[B]}{\mathbb{P}[A]}$$

and with the formula of total probability, we get

$$\mathbb{P}[B|A] = \frac{\mathbb{P}[A|B]\mathbb{P}[B]}{\mathbb{P}[A|B]\mathbb{P}[B] + \mathbb{P}[A|B^c](1 - \mathbb{P}[B]}$$

or, for a general partition $\Omega = \bigsqcup_{i\in I} B_i$ we have

$$\mathbb{P}[B_i|A] = \frac{\mathbb{P}[A|B_i]\mathbb{P}[B_i]}{\sum_{j\in I} \mathbb{P}[A|B_j]\mathbb{P}[B_j]}$$

We will see that Bayes formula is the simplest formula that generates a lot of unintuitive results.
If we are given a random variable, we can of course talk about the *conditional* expectation value.

Definition 1.19. Let $(\Omega, \mathcal{A}, \mathbb{P})$ be a discrete P-space. For an event $B \in \mathcal{A}$ with non-zero probabilty we define the **conditional expectation value** to be

$$\mathbb{E}[X|B] = \frac{\mathbb{E}[\mathbb{1}_B X]}{\mathbb{P}[B]} = \sum_{\omega\in\Omega} X(\omega)\mathbb{P}[\{\omega\}|B] = \sum_{x\in X(\Omega)} x\mathbb{P}[X = x|B]$$

where $\mathbb{1}_B$ is the characteristic function.
Given a partition $\mathcal{G} = (B_i)_{i\in I}$ of Ω into pariwse disjoint non-empty subsets, where I is at most countable, we can define a random variable by

$$\mathbb{E}[X|\mathcal{G}](\omega) := \sum_{i\in I} \mathbb{E}[X|B_i]\mathbb{1}_{B_i}(\omega)$$

which is the conditional expectation value of X given the partition $\mathcal{G}$.

We can assert that this definition is useful in the sense that it does what it is supposed to do. We can show that the conditional expectation value *is* the best approximation of X on the subset X.

Theorem 1.20. *Let X be a random variable on $(\Omega, \mathcal{A}, \mathbb{P})$, such that $\mathbb{E}[X^2] < \infty$ and $\mathcal{G} = (B_i)_{i\in I}$ a partition of Ω. Then*

$$\mathbb{E}\left[\left(X - \sum_{i\in I} c_i \mathbb{1}_{B_i}\right)^2\right]$$

is minimal for $c_i = \mathbb{E}[X|B_i]$[1]

Proof. First note that since the B_i are disjoint, we have $\mathbb{1}_{B_i}\mathbb{1}_{B_j} = \delta_{ij} I_{B_i}$. By linearity of the expectation value we get for any c_i

$$\begin{aligned}\mathbb{E}\left[X\sum_{i\in I} c_i\mathbb{1}_{B_i}\right] &= \mathbb{E}\left[\sum_{i\in I} c_i X\mathbb{1}_{B_i}\right] = \mathbb{E}\left[\sum_{i\in I} c_i \frac{\mathbb{E}[X\mathbb{1}_{B_i}}{\mathbb{P}[B_i]}\mathbb{1}_{B_i}\right]\\ &= \mathbb{E}\left[\sum_{i,j\in I} c_i\mathbb{E}[X|B_i]\mathbb{1}_{B_i}\mathbb{1}_{B_j}\right] = \mathbb{E}\left[\mathbb{E}[X|\mathcal{G}]\sum_{i\in I} c_i\mathbb{1}_{B_i}\right]\end{aligned}$$

### missing 5 mins

□

The conditional probability can also be seen as the orthogonal projection with respect to the dot product on random variables

$$\langle X, Y\rangle := \sum_{\omega\in\Omega} X(\omega)Y(\omega)p(\omega)$$

Definition 1.21. A collection of subsets $(A_i)_{i\in I}$, $A_i \subseteq \Omega$ is said to be (stochastically) **independent** (with respect to $\mathbb{P}$), if for *all* finite subsets $J \subseteq I$

$$\mathbb{P}\left[\bigcap_{j\in J} A_j\right] = \prod_{j\in J} \mathbb{P}[A_j]$$

[1] Implicitly, the sum doesn't go over all i, but only over those where B_i has non-zero probability.

Remark 1.22.

- For two events A, B with positive probability we have

$$A, B \text{ independent } \iff \mathbb{P}[A|B] = \mathbb{P}[A] \iff \mathbb{P}[B|A] = \mathbb{P}[B]$$

- Pairwise independence is *not* enough to show that a collection is independent as a whole. The condition must hold for *all* finite subsets J. As a counter example, consider the two coin throws where

$$\begin{aligned} A &= \text{“first throw is Head”} \\ B &= \text{“second throw is Head”} \\ C &= \text{“The outcomes of both throws are different”} \end{aligned}$$

these all have non-zero probability individually and are pairwise disjoint, but $A \cap B \cap C = \emptyset$.

Lemma 1.23. *If $(A_i)_{i \in I}$ are independent, then for $B_i = A_i$ or $B_i = A_i^c$, the collection $(B_i)_{i \in I}$ is also independent.*

Proof. We let $J \subseteq I$ be the indices, where $B_i = A_i$ and $K \subseteq I$ the ones where $B_i = A_i^c$ and use induction on $|K| = k$:
Since the condition holds for all $|J|$, we can let $\tilde{K} = K \cup \{l\}$. Then

$$\begin{aligned} \mathbb{P}\left[\bigcap_{j \in J} A_j \cap \bigcap_{k \in K} A_k^c \cap A_l^c\right] &= \mathbb{P}\left[\bigcap_{j \in J} A_j \cap \bigcap_{k \in K} A_k^c\right] - \mathbb{P}\left[\bigcap_{j \in J} A_j \cap A_l \cap \bigcap_{k \in K} A_k^c\right] \\ &= \prod_{j \in J} \mathbb{P}[A_j] \prod_{k \in K} \mathbb{P}[A_k^c](1 - \mathbb{P}[A_l]) \end{aligned}$$

□

Just like we could pull the probability measure onto a random variable to obtain the expectation value, we can pull the independence of events to define independence of random variables.

Definition 1.24. A collection of discrete random variables $(X_i)_{i \in I}$ is said to be **independent**, if the set of all possible events $(\{X_i = y\})_{i \in I}$ are independent for any selection of $y \in \mathbb{R}$. An equivalent categorisation is that for any finite subset $J \subseteq I$

$$\mathbb{E}\left[\prod_{j \in J} X_j\right] = \prod_{j \in J} \mathbb{E}[X_j] \iff \mathbb{P}[X_j = x_j, \forall j \in J] = \prod_{j \in J} \mathbb{P}[X_j = x_j]$$

1.5 Random Walks

The random walk is a model for the random motion of a particle in an n-dimensional grid $\mathbb{Z}^n$ starting at the origin. At every *period*, the particle can move in any direction.
Let's first check out the one-dimensional case with N periods: Let Ω to be the set of all binary sequences of length N

$$\Omega = \{\omega = (x_1, \ldots, x_N) | x_i \in \{\pm 1\}\}$$

and let $X_k(\omega)$ to the the k-th component of $\omega \in \Omega$. The position after the n-th period will then be

$$S_n(\omega) = \sum_{k=1}^{n} X_k(\omega)$$

For arbitrary periods ($N \to \infty$) we run into a problem. The set Ω is then bijective to $\mathcal{P}(\mathbb{N})$, which is an uncountable set. It is therefore difficult to find a good probability measure on Ω.
For finite N however, the set Ω has cardinality $|\Omega| = 2^N$ so we can just use the equal distribution $\mathbb{P}$ given by

$$\mathbb{P}[A] = \frac{|A|}{|\Omega|} = 2^{-N}|A|$$

It is easy to show the following

(a) $\mathbb{P}[X_k = +1] = \frac{1}{2}$

(b) $\mathbb{P}[X_{k_1} = x_{k_1}, \dots X_{k_l} = x_{k_l}] = 2^{-l}$ for any $1 \leq k_1 < \ldots < k_l \leq N$. In particular, the random variables $X_1, \dots, X_N$ are all independent (in the sense of definition 1.24)

(c) $\mathbb{E}[X_k] = (+1)\mathbb{P}[X_k = +1] + (-1)\mathbb{P}[X_k = -1] = 0$, so by linearity of the expectation value, $\mathbb{E}[S_n] = 0$.

Proposition 1.25. *For a given n, the random variable S_n takes on values $\{-n, -n+2, \dots, n-2, n\}$ with probality*

$$\mathbb{P}[S_n = 2k - n] = \binom{n}{k} 2^{-n} = \binom{n}{k} \frac{1}{2}^k \left(1 - \frac{1}{2}\right)^{n-k}, \quad k = 0, 1, \dots, n$$

The distribution of S_n is called a linearly transformed binomial distribution *with $p = \frac{1}{2}$*

Proof. Let define U_n the number of $+1$ steps up to period n. So

$$U_n = \sum_{k=1}^{n} \mathbb{1}_{\{X_k = +1\}}$$

Then $S_n = U_n - (n - U_n) = 2Un - n$, so by calculating the cardinality

$$|\{S_n = 2k - n\}| = |\{U_n = k\}| = \binom{n}{k} 2^{N-n}$$

the proability is just that divided by $|\Omega| = 2^N$:

$$\mathbb{P}[S_n = 2k - n] = \binom{n}{k} 2^{-n}$$

□

Using Stirling's formula

$$n! \sim \left(\frac{n}{e}\right)^n \sqrt{2\pi n}$$

We can calculate the probability that the particle returns to the origin after $2n$ steps:

$$\mathbb{P}[S_{2n} = 0] = \frac{\mathbb{P}[|S_{2n-1}| = 1]}{2} = \mathbb{P}[S_{2n-1} = 1] = \binom{2n}{n} 2^{-2n} \sim \frac{1}{\sqrt{\pi n}}$$

1.5.1 Reflection principle

For $a \in \mathbb{Z}$, consider the time until the particle first reaches a. This can be done with the random variable T_a given by

$$T_a(\omega) = \min\{n > 0 : S_n(\omega) = a\}$$

where for $a = 0$ we are interested in the first *return* to the origin and where we set $\min \emptyset = N + 1$.
We might ask ourselves, what is the probability that the particle reaches position $-a$ at some point and then travels to b? The following lemma simplifies this by reflecting the trajectory from $-a$ to b around $-a$. To reach $-2a - b$ the particle obviously has to pass $-a$ at some point, so the "extra" condition drops.

Lemma 1.26 (Reflection principle)**.** *For $a > 0$ and $b \geq -a$*

$$\mathbb{P}[T_{-a} \leq n, S_n = b] = \P[S_n = -2a - b]$$

Although this lemma seems rather simple, it has some nice consequences. It allows us to find the probability distribution for T_a.

Theorem 1.27. *For $a \neq 0$ we have*

$$\mathbb{P}[T_{-a} \leq n] = 2\mathbb{P}[S_n < -a] + \mathbb{P}[S_n = -1] = \mathbb{P}[S_n \notin (-a, a]]$$

Proof. Using the additivity $\mathbb{P}$ on disjoint sets, we have

$$\begin{aligned}\mathbb{P}[T_{-a} \leq n] &= \sum_{b=-\infty}^{\infty} \mathbb{P}[T_{-a} \leq n, S_n = b] \\ &= \sum_{b=-\infty}^{-a} \mathbb{P}[S_n = b] + \sum_{b=-a+1}^{\infty} \mathbb{P}[S_n = -2a - b] \\ &= \mathbb{P}[S_n \leq -a] + \P[S_n \leq -a - 1]\end{aligned}$$

□

Corollary 1.27.1. *For ever $a \neq 0$ we have*

(a) $\mathbb{P}[T_a > N] \to 0$ *for* $N \to \infty$

(b) $\mathbb{E}[T_a] = \sum_{k=1}^{N+1} k\mathbb{P}[T_a = k] \to \infty$ *for* $N \to \infty$*.*

Proof. Let $a > 0$. Using the previous theorem and Stirling's formula we get

$$\mathbb{P}[T_{-a} > N] = \mathbb{P}[S_n \in (-a, a]] \leq \frac{C}{\sqrt{\pi N}} \to 0$$

for some constant C. For the expectation value, we get

$$\begin{aligned}\mathbb{E}[T_{-a}] &= \sum_{k=0}^{N} \mathbb{P}[T_{-a} > k] \\ &= \sum_{k=0}^{N} \mathbb{P}[S_k \in (-a, a]] \\ &\geq \sum_{k=1}^{N} \mathbb{P}[S_k \in \{0, 1\}] \to \infty \quad \text{for} \quad N \to \infty\end{aligned}$$

□

In the frequentist interpretation, this corollary says that we every position will be reached eventually, but we might need infinitely many trials $\omega \in \Omega$ to get there.
If we have a path of length $2n$ that *stay above* 0, then by omitting the very first step, we uniquely obtain a paths of length $2n-1$ that *never reaches* -1. The two types of paths are in bijection, so we get the theorem

Theorem 1.28.

$$\mathbb{P}[T_0 > 2n] = \mathbb{P}[S_{2n} = 0]$$

Proof. The comment above is pretty much the proof of the theorem. Written out, it says

$$\mathbb{P}[T_0 > 2n] = \frac{1}{2}\mathbb{P}[T_{-1} > 2n-1] + \frac{1}{2}\mathbb{P}[T_1 > 2n-1] = \mathbb{P}[T_{-1} > 2n-1]$$
$$= \mathbb{P}[S_{2n-1} \in (-1,]] = \mathbb{P}[S_{2n-1} = 1] = \mathbb{P}[S_{2n} = 0]$$

□

One property of random walks is that they are random. This might seem like an obvious fact but it turns out that (we) humans are bad at generating random numbers.
When people are tasked to draw random walks of length 100 and we plot a histogram that measures the number of longest runs, the data does not fit what we expect if the walks were random.
In particular, almost all truly random walks had a longest run length of 7 or above. But humans see a sequence of 7 identical outcomes back-to-back and think that it doesn't seem random. A professor tested their students with flipping 100 coins and from their results was able to tell when they actually used random events or made it up by seeing what the longest run length was.
Random events show some structure, not *despite* their inherent randomness, but *because* of it.

1.5.2 The Arcsine Law

Let

$$L(\omega) = \max\{0leqn \leq 2N | S_n(\omega) = 0\}$$

be the time of the last visit of the origin of a random walk ω.
What does the distribution of L look like?

Theorem 1.29 (Arcsine Law)**.** *The distribution of L is the* ***discrete Arcsine*** *distribution*

$$\mathbb{P}[L = 2n] = \mathbb{P}[S_{2n} = 0]\mathbb{P}[S_{2N-2n} = 0] = 2^{-2N}\binom{2n}{n}\binom{2N-2n}{N-n}$$

which is symmetric around N and looks like a U-shape. So if we had to guess at what time the last return to monke was, it should be either in the beginning or at the end.
It is called the Arcsine law because for $f(x) = \frac{1}{\pi\sqrt{x(1-x)}}$ we have

$$\mathbb{P}\left[\frac{L}{2N} \leq z\right] \sim \sum_{k:\frac{k}{N}\leq z} \frac{1}{N} f(\frac{k}{N}) \sim \int_0^z f(x)dx = \frac{2}{\pi}\arcsin\sqrt{z}$$

Proof. If we let Z_m be time of the last return to origin in the first m moves, then we can use independence of the random variables to write

$$\begin{aligned}\mathbb{P}[Z_{2n} = 2k] &= \mathbb{P}[S_{2k} = 0, S_{2k+1} \neq 0, \dots S_{2n} \neq 0] \\ &= \mathbb{P}[S_{2k} = 0] \cdot \mathbb{P}[S_1 \neq 0, \dots, S_{2n-2k} \neq 0]\end{aligned}$$

To calculate the right hand side we use the maximum principle t get

$$\mathbb{P}[S_1 \geq 0, S_2 \geq 0, \dots S_{2n} \geq 0] = \binom{2n}{n} 2^{-2n}$$

and by substituting the weak inequality with a strict inequality, we obtain

$$\begin{aligned}\mathbb{P}[S_1 > 0, \dots S_{2j} > 0] &= \mathbb{P}[S_1 = 1, S_2 \geq 1, \dots, S_{2j} \geq 1] \\ &= \mathbb{P}[S_1 = 1] \cdot \mathbb{P}[S_2 \geq 1, \dots, S_{2j} \geq 1] \\ &= \mathbb{P}[S_1 = 0]\mathbb{P}[S_2 \geq 0, \dots, S_{2j} \geq 0] \\ &= \binom{2j}{j} 2^{-2j}\end{aligned}$$

so combining the first and last result, we get

$$\mathbb{P}[Z_{2n} = 2k] = \binom{2k}{k}\binom{2j}{j} 2^{-2j}$$

□

Note that 0 and $2n$ are the two most probable values for Z_{2n}.

1.5.3 Game systems

Recall that the expected value of the endpoint is zero

$$\mathbb{E}[S_n] = 0$$

in terms of game systems, this means that for a fair game the “gain” after n rounds is zero.

Definition 1.30. An event $A \subseteq \Omega$ is called **observable** until cycle n, if it is of the Form

$$\left\{\omega \in \{\pm 1\}^N | (X_1(\omega), \dots X_n(\omega)) \in C \quad \text{for some} \quad C \subseteq \{\pm 1\}^n\right\}$$

Denote the set of all until cycle n observable events as $\mathcal{F}_n$.

Here we use the convention $\mathcal{F}_0 = \{\emptyset, \Omega\}$. It is then clear that $\mathcal{F}_n \subseteq \mathcal{F}_{n+1}$.
Assume we know how the game will play out in the future. This information could then tell us whether we should continue to play or to stop.
A stopping time should only make use of the information up to the n-th round. We should not be able to say stop at round n *after* the n-th round is played. To formalize this, we use the following definition

Definition 1.31. A map $T : \Omega \to \{0, \dots, N\}$ is a **stopping-time**,if

$$\{\omega : T(\omega) \leq n\} \in \mathcal{F}_n, \quad \forall n \in \{0, \dots, N\}$$

Example 1.32. A non-example is to say stop when at cycle n we get the maximum value obtainable from outcome ω, i.e.

$$T(\omega) = \min \left\{ 0 \leq n \leq N | S_n(\omega) = \max_{0 \leq k \leq N} S_k(\omega) \right\}$$

this is obviously not a stopping-time as for example $\{\omega : T(\omega) = 0\}$ must require that $S_k \leq 0$ at all times k, which requires knowledge of all future rounds and is not a part of $\mathcal{F}_0$.

We said earlier that unless we had some extra information about the game is going to play out, we should not be able to tell whether to stop or not. This can be formalised as follows:

Theorem 1.33. *For every stopping-time T,*

$$\mathbb{E}[S_T] = 0$$

where $S_T(\omega) = S_{T(\omega)}(\omega)$ is the accumulated win when using the stopping rule T.

This is telling us that no blind strategy is going to make (or lose) us money in a fair game. The game's fairness cannot be chated.
While we're already looking at random walks as the outcomes of a game, let's generalize this to get a definition to model general games that come in discrete cycles.

Definition 1.34. A **game system** is a sequence of random variables $V = (V_k)_{k \in \mathbb{N}}, V_k : \Omega \to \mathbb{R}$ such that V_1 is constant and for $k \geq 2$ there exist functions

$$\varphi_k : \{+1, -1\}^{k-1} \to \mathbb{R} \quad \text{with} \quad V_k(\omega) = \varphi_k \left(X_1(\omega), \ldots, X_{k-1}(\omega) \right)$$

Now let's consider random walks, where the random variables X_i takevalue in $\{0, 1\}$ instead of $\{+1, -1\}$. Ω is then the set of $0-1$ sequences of length n.
Recall that in the Laplace model, $\mathbb{P}$ is the equal distribution on Ω.
This had the "consequence" that the random variables X_i for the random walk were independent and had probability

$$\mathbb{P}[X_i = 1] = \frac{1}{2}, (i = 1, \ldots, n)$$

on the other hand, we can set $\mathbb{P}[X_i = 1]$ to $\frac{1}{2}$ and then recover our probability distribution $\mathbb{P}$ on Ω by requiring that the random variables were independent.
This allows us to consider a varied system in which the probability above is given by

$$\mathbb{P}[X_i = 1] := p, \quad \text{for} \quad p \in [0, 1]$$

which uniquely determines a probability distribution $\mathbb{P}$ on Ω, which associates to an outcome $\omega = (x_1, \ldots, x_n)$ the probability

$$\mathbb{P}[\{\omega\}] = \mathbb{P}\left[\bigcap_{i=1}^{n} \{X_i = x_i\} \right] = \prod_{i=1}^{n} \mathbb{P}[X_i = x_i] = p^k (1-p)^{n-k}$$

where k depends on ω and measures the number of x_i such that $x_i = 1$.
This gives us the distribution of the random variable S_n, which is given by

$$\mathbb{P}[S_n = k] = \binom{n}{k} p^k (1-p)^{n-k}$$

a more analytical way to prove this is to decompose $S_n = X_1 + \ldots + X_n$ and use the property of the exponential function to get

$$\begin{aligned}\mathbb{E}[e^{i\lambda S_n}] &= \mathbb{E}[e^{i\lambda(X_1+\ldots+X_n}] \\ &= \mathbb{E}[e^{i\lambda X_1}]\ldots\mathbb{E}[e^{i\lambda X_n}] \\ &= \Big((1-p)+e^{i\lambda}p\Big)^n \\ &= \sum_{k=0}^{n}\binom{n}{k}p^k(1-p)^{n-k}e^{i\lambda k}\end{aligned}$$

but by doing comparison of coefficients with the expression

$$\mathbb{E}[e^{i\lambda S_n}] = \sum_{k=0}^{n}\P[S_n = k]e^{i\lambda k}$$

we recover the binomial distribution.
For

$$p_n(k) = \binom{n}{k}p^k q^{n-k} \quad \text{with} \quad q = 1-p$$

the binomial distribution allows for a recursive formulation:

$$p_n(k+1) = \frac{n-k}{k+1}\frac{p}{q}p_n(k) = \frac{np-kp}{k-kp+1-p}p_n(k)$$

, or at least when $q \neq 0$. In that case, we get a delta distribution where the only singleton event with positive probability is $S_n = n$.
We can ask whether there exists an analytic function describing $p_n(k)$ as a function of k?

Theorem 1.35 (de Moivre-Laplace)**.** *For $q = 1-p$ we have the analytic expression*

$$\boxed{p_n(k) = \binom{n}{k}p^k q^{n-k} = \frac{1}{\sqrt{2\pi npq}}\exp\left(-\frac{(k-np)^2}{2npq}\right)(1-r_n(k)) =: \varphi_{n,p}(k)\big(1+r_n(k)\big)}$$

where the remainder term $r_n(k)$ converges such that

$$\sup\{|r_n(k)|||k-np| \leq A\sqrt{n}\} \to 0 \quad \textit{for all} \quad A > 0 \quad \textit{for} \quad n \to \infty$$

Proof. Because we have the factorial in the expression p_n, we can use the Stirling Formula $m! \sim \sqrt{2\pi m}(\frac{m}{e})^m$ to get

$$\begin{aligned}p_n(k) &\simeq \frac{\sqrt{2\pi n}n^n p^k q^{n-k}}{\sqrt{(2\pi)^2 k(n-k)}k^k(n-k)^{n-k}} \\ &= [\ldots] \\ &= \frac{1}{\sqrt{2\pi n\frac{k}{n}(1-\frac{k}{n})}}\exp(ng_p(k/n))\end{aligned}$$

where $g_p(x)$ is given by

$$g_p(x) = x(\log p - \log x) + (1-x)\big(\log(1-p) - \log(1-x)\big)$$

Then, by expanding g_p in a taylor expansion at $x = p$ we first get

$$g_p(p) = 0 = g_p'(p), \quad \text{and} \quad g_p''(p) = \frac{-1}{p(1-p)}$$

[missing 5 mins] □

Although the proof is quite obtuse, the formula itself is rather intuitive. Notice that since the expectation value for $X_i = p$, we get that $\mathbb{E}[S_n] = np$. So the term $(k-np)^2$ measure the deviatin from the expected value.
Moreover, the variance of S_n is given by

$$\begin{aligned}\mathrm{Var}[S_n] &= \mathbb{E}\left[(S_n - \mathbb{E}[S_n])^2\right] \\ &= \mathbb{E}\left[\left(\sum_{i=1}^{n}(X_i - p)\right)^2\right] \\ &= \sum_{i,j\le n}^{k} \mathbb{E}[(X_i-p)(X_j-p)] \\ &= \sum_{i=1}^{n}\text{[missing 1 min]} \\ &= \end{aligned}$$

so the term npq is the variance.

Theorem 1.36 (Poisson approximation)**.** *For k fix and $n \to \infty, p \to \infty$ such that $np \to \lambda$ we have*

$$\boxed{\binom{n}{k}p^k(1-p)^{n-k} \to \frac{\lambda^k}{k!}e^{-\lambda}}$$

Proposition 1.37. *Let X_1, X_2 be independent random variables such that*

$$\mathbb{P}[X_1 = k | X_1 + X_2 = n] = \binom{n}{k}2^{-n} \quad \textit{for all} \quad n \in \mathbb{N}, 0 \le k \le n$$

then, X_1, X_2 follow the Poisson distribution with the same parameter λ.

Proof. By assumption, we can write

$$\begin{aligned}\frac{1}{n} = \frac{\binom{n}{n}2^{-n}}{\binom{n}{n-1}2^{-n}} &= \frac{\mathbb{P}[X_1 = n | X_1 + X_2 = n]}{\mathbb{P}[X_1 = n-1 | X_1 + X_2 = n} \\ &= \frac{\mathbb{P}[X_1 = n, X_2 = 0]}{\mathbb{P}[X_1 = n-1, X_2 = 1]} \\ &= \frac{\mathbb{P}[X_1 = n]\mathbb{P}[X_2 = 0]}{\mathbb{P}[X_1 = n-1]\mathbb{P}[X_2 = 1]}\end{aligned}$$

By setting $\lambda := \frac{\mathbb{P}[X_2=1]}{\mathbb{P}[X_2=0]}$ we get

$$\mathbb{P}[X_1 = n] = \frac{\lambda}{n}\mathbb{P}[X_1 = n-1] = \frac{\lambda^n}{n!}\mathbb{P}[X_1 = 0]$$

since the probabilities have to add up to 1, it follows that $\mathbb{P}[X_1 = 0] = e^{-\lambda}$. □

As a result, we can show that the sum of Poisson distributions is again a Poisson distribution.

Proposition 1.38. *If X_1, X_2 are independent and have Poission distributions for parameters λ_1, λ_2, then their sum $X = X_1 + X_2$ has Poission distribution with paramter $\lambda = \lambda_1 + \lambda_2$.*

2 Continuous Models

We now want to move from discrete probability spaces and consider *continuous* ones, where the base space Ω is uncountable.
A generalisation of the defintion of a P-space is given as follows

Definition 2.1. A (continuous) P-space is a tuple $(\Omega, \mathcal{A}, \mathbb{P})$ if

(a) $\mathcal{A}$ is a σ-Algebra.

(b) $\mathbb{P} : \mathcal{A} \to [0, \infty]$ is σ-additive and normed ($\mathbb{P}[\Omega] = 1$)

From measure theory, we know that we can extend $\mathcal{A}$ to the class of Lebesgue-measurable sets.
Although $\mathbb{P}$ often cannot be extended to the complete powerset $\mathcal{P}(\Omega)$, that is often not necessary.
Instead of starting with the probabilty measure $\mathbb{P}$ and analyze random variables as we did in the previous section, we can also go the other way around and introduce random variables X_i and look for a $\mathbb{P}$ that satisfies some properties with respect to the X_i.
For this, let's first consider the $0-1$ experiments, where

$$\Omega = \{0,1\}^{\mathbb{N}} = \{\omega = (x_1, x_2, \dots) : x_i \in \{0,1\}\}$$

and set $X_i(\omega) = x_i$ be the i-th component of the outcome. Then, let $\mathcal{A}$ be the σ-Algebra generated by sets of the form $\{\omega : X_i(\omega) = 1\}, i = 1, 2, \dots$.

Theorem 2.2. *Given a parameter $p \in [0,1]$, there exists a unique probability measure $\mathbb{P}_p$ on $\mathcal{A}$ such that*

- *$\mathbb{P}[X_i = 1] = p$ for all i*
- *The events $\{X_i = 1\}$ for all independent with respect to $\mathbb{P}$.*

Proof. It follows from the requirements on $\mathbb{P}$ that for any choice of $x_1, \dots, x_n$ we must have for $k = \sum_{i=1}^n x_i$

$$\mathbb{P}[X_1 = x_1, \dots, X_n = x_n] = \prod_{i=1}^n \mathbb{P}[X_i = x_I] = p^k(1-p)^{n-k}$$

which automatically shows that $\mathbb{P}$ is well defined on $\mathcal{A}$ and uniquely determins the events generated by finite union of the Form $\{X_1 = x_i, \dots, X_n = x_n\}$. By the Carathéodory-Hahn theorem from measure theory, such an extension exsists and is unique. □

Lemma 2.3 (Borel-Cantelli). *Let $A_1, \dots, A_2, \dots$ be a sequence of events in $\mathcal{A}$. And let*

$$A_\infty = \limsup_{n\to\infty} A_n = \bigcap_{n=1}^{\infty} \bigcup_{k=n}^{\infty} A_k$$

be the set where infinitely many of the A_k occur. Then

(a) $\sum_{k=1}^{\infty} \mathbb{P}[A_k] < \infty$ *implies* $\mathbb{P}[A_\infty] = 0$

(b) If the events $(A_k)_{k\in\mathbb{N}}$ *are independent, then* $\sum_{k=1}^{\infty} \mathbb{P}[A_k] = \infty$ *implies* $\mathbb{P}[A_\infty] = 1$.

Proof. Since the sequence $B_n = \bigcup_{k\geq n} A_k$ is decreasing, The proof follows from the subadditivity that

$$\mathbb{P}[A_\infty] = \lim_{n\to\infty} \mathbb{P}[B_n] \leq \lim_{n\to\infty} \sum_{k\geq n} \mathbb{P}[A_k] = 0$$

On the other hand, if they are independent, then we can write

$$\mathbb{P}\left[\bigcap_{k\geq n} A_k^c\right] = \prod_{k\geq n} \mathbb{P}[A_k^c] = \prod_{k\geq n} (1 - \mathbb{P}[A_k])$$

and since the exponential function is convex, $1 - x \leq e^{-x}$, so

$$\mathbb{P}\left[\bigcap_{k\geq n} A_k^c\right] \leq \exp\left(-\sum_{k\geq n} \mathbb{P}[A_k]\right) = 0$$

and since the sequence $\bigcap_{k\geq n} A_k^c$ is increasing in n it follows that

$$\mathbb{P}[A_\infty^c] = \lim_{n\to\infty} \mathbb{P}\left[\bigcap_{k\geq n} A_k^c\right] = 0$$

□

An interesting application of the Borel-Cantelli lemma is the following "experiment"

Example 2.4. Let $N \in \mathbb{N}$ and $\{x_1, \dots, x_n\}$ be a "binary text" of length N. Then the probability that the text appears in any outcome is 1.
To show this, we can consider the events A_k for $k = 1, 2, \dots$ given by

$$A_k := \{X_{(k-1)N} = x_1, \dots X_{kN} = x_N\}$$

these are all independent and have non-zero probability $\mathbb{P}[A_k] > 0$. Then we can apply the Borel-Cantelli Lemma.

2.1 Transformation of P-spaces

In the following, let $(\Omega, \mathcal{A}, \mathbb{P})$ be a P-space, $\tilde{\Omega} \neq \emptyset$, and $\tilde{\mathcal{A}} \subseteq \mathcal{P}(\tilde{\Omega})$ a σ-Algebra on $\tilde{\Omega}$.
Given a map $\varphi : \Omega \to \tilde{\Omega}$, we wish to induce a probability measure on $\tilde{\Omega}$ respecting φ. But that is not always possible for every φ.

Definition 2.5. A map $\varphi : \Omega \to \tilde{\Omega}$ is **measurable** (with respect to $\mathcal{A}$ and $\tilde{\mathcal{A}}$) if $\varphi^{-1}(\tilde{\mathcal{A}}) \subseteq \mathcal{A}$, i.e.

$$\varphi^{-1}(\tilde{A}) \in \mathcal{A}, \quad \text{for all} \quad \tilde{A} \in \mathcal{A}$$

Note that if $\tilde{\mathcal{A}}$ is the σ-Algebra generated by a collection of subsets $\tilde{\mathcal{A}}_0 \subseteq \tilde{\mathcal{A}}$, then it is sufficient to check if

$$\varphi^{-1}(\tilde{A}) \in \mathcal{A}, \quad \text{for all} \quad \tilde{A} \in \tilde{\mathcal{A}}_0$$

because the collection $\{\tilde{A} \subseteq \tilde{\Omega} | \varphi^{-1}(\tilde{A}) \in \mathcal{A}\}$ is a σ-Algebra containing $\tilde{\mathcal{A}}_0$.

Proposition 2.6. *If $\varphi : \Omega \to \tilde{\Omega}$ is measuarble, then we can obtain a probability measure $\tilde{\mathbb{P}}$ on $\tilde{\mathcal{A}}$ given by*

$$\tilde{\mathbb{P}}[\tilde{A}] := \mathbb{P}\left[\varphi^{-1}(\tilde{A})\right] \quad \textit{for all} \quad \tilde{A} \in \tilde{\mathcal{A}}$$

we call $\tilde{\mathbb{P}}$ the image of $\mathbb{P}$ under φ and write $\tilde{\mathbb{P}} = \mathbb{P} \circ \varphi^{-1}$.

Using the notion of measurable maps, we can re-define what a random variable is and obtaina nicer definition of a distribution of a random variable.

Definition 2.7. Let $(\Omega, \mathcal{A}, \mathbb{P})$ be a P-space. A **random variable** is a measurable map

$$X : (\Omega, \mathcal{A}) \to (\mathbb{R}, \mathcal{B})$$

, where $\mathcal{B}$ is the Borel σ-Algebra on $\mathbb{R}$ generated by intervals of the Form $(-\infty, b]$, for $b \in \mathbb{R}$, which contains all open and closed sets.
The **distribution** μ of X is the pushforward $\mathbb{P} \circ X^{-1}$ given by

$$\mu(A) = \mathbb{P}[X^{-1}(A)] = \mathbb{P}[\{\omega \in \Omega : X(\omega) \in A\}] \quad \text{for} \quad A \in \mathcal{B}$$

Example 2.8. In the case where $(\Omega, \mathcal{F}, \mathbb{P})$ is the model of $0-1$ experiments and X is the time until the first 1:

$$X(\omega) = \min\{k \geq 1 : x_k = 1\}$$

then the distribution of μ is given by

$$\mu(\{k\}) = \mathbb{P}[X = k] = \mathbb{P}[]\{\omega \in \Omega : x_1 = \ldots = x_{k-1} = 0, x_k = 1\} = (1-p)^{k-1}p$$

for singletons, and for arbitrary $A \in \mathcal{B}$, it is

$$\mu(A) = \sum_{k \in A} (1-p)^{k-1}$$

which gives rise to the **geometric** distribution on $\mathbb{N}$, as for $p = \frac{1}{2}$ we obtain the geometric series.

Definition 2.9. Let X be a random variable on $(\Omega, \mathcal{A}, \mathbb{P})$. The function

$$F : \mathbb{R} \to [0, 1], \quad F(b) := \mathbb{P}[X \leq b] = \mu((-\infty, b])$$

is called the **distribution function** of X (or μ).

Remark 2.10. First note the following. For $a < b \in \mathbb{R}$ we have

$$\mu((a,b]) = F(b) - F(a)$$

and we can obtain the discontinuity of F at a point a by evaluating μ at a:

$$\begin{aligned}\mu(\{a\}) &= \lim_{n\to\infty}\left(\left(a - \frac{1}{n}, a\right]\right) \\ &= F(a) - \lim_{h>0\to 0} F(a-h)\end{aligned}$$

Theorem 2.11. *This distribution function has the following properties*

(a) Monotoneity: $a \leq b \implies F(a) \leq F(b)$

(b) ***cadlag***[2]*:* $F(a) = \lim_{h>0\to 0} F(a+h)$

(c) $\lim_{a\to-\infty} F(a) = 0$ *and* $\lim_{a\to\infty} F(a) = 1$

*On the other hand, every function with these three properties is the distribution function of a random variable X (we sometimes write F^{-1}). We call X the **quantile** of the distribution, or more explicitly, we say $X(t)$ the t-quantile of F.*
*An important example is the 50%-Quantile $X(\frac{1}{2})$, also known as the **median**.*

The properties are easy to check. To prove the existence of such a random variable, we require the following lemma

Lemma 2.12. *Let F be such that it satisfies the properties (a) - (c) and define*

$$X(t) = \inf\{x | F(x) \geq t\}$$

then X is monotonous, left-continuous and

$$X(F(x)) \leq x \quad \forall x \in \mathbb{R} \quad \text{and} \quad t \leq F(X(t)) \quad \forall t \in (0,1)$$

Proof Lemma. □

Proof Theorem. If F satisfies these properties, then for $0 < t < 1$ we define

$$X(t) := \inf\{x | F(x) \geq t\}$$

This function is measurable and by the lemma, we we have

$$X(t) \leq x \iff t \leq F(x)$$

Then we chose the equal distribution $\mathbb{P}$ on $[0,1]$ and so we get

$$\mathbb{P}[X \leq b] = \mathbb{P}\left[\{\omega : X(\omega) \leq b\}\right] = \mathbb{P}\left[\{\omega : \omega \leq F(b)\}\right] = F(b)$$

□

[2] Continue à droit, limite à gauche

2.2 Types of distributions

Definition 2.13. A random variable X is **discrete** if there exists a countable set $A \subseteq \mathbb{R}$ such that $\mathbb{P}[X \in A] = 1$. The distribution function

$$F(b) = \sum_{x \in A, x \leq b} \mathbb{P}[X = x]$$

is a step function with discontinuities at points in A.
X is called **absolutely continuous** if there exists a measurable function $f : (\mathbb{R}, \mathcal{B}) \to (\mathbb{R}, \mathcal{B})$ with $f(x) \geq 0$ and $\int_{-\infty}^{\infty} f(x)dx = 1$. such that

$$F(b) = \int_{-\infty}^{b} f(x)dx \quad \text{for all} \quad b \in \mathbb{R}$$

we call f the **density** of X (and is written f_X).

Note that such a function f is unique up to $\mathcal{L}$-measure zero differences.

Example 2.14. For the uniform distribution of X on $[a, b]$, the density is

$$f(x) = \begin{cases} \frac{1}{b-a} & x \in [a, b] \\ 0 & \text{otherwise} \end{cases}$$

The exponential distribution with parameter $\alpha > 0$ (called $\text{Exp}(\alpha)$ has density

$$f(x) = \begin{cases} \alpha e^{-\alpha x} & x \geq 0 \\ 0 & x < 0 \end{cases}$$

and the distribution function has the form

$$F(b) = \begin{cases} 1 - e^{-\alpha x} & x \geq 0 \\ 0 & x < 0 \end{cases}$$

Another important distribution is the **Normal distribution**. It has parameters μ and σ^2 for the center and the variance and we write $\mathcal{N}(\mu, \sigma^2)$. It's density is given by

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

the distribution function F doesn't have a closed form but has a quite distinct look.

Lemma 2.15. *Let* $X_1, \ldots, X_n \overset{i.i.d.}{\sim} \mathcal{N}(\mu, \sigma^2)$. *For* $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$, *we can obtain the normalisation*

$$\frac{\overline{X} - \mathbb{E}[\overline{X}]}{\sqrt{\text{Var}(\overline{X})}} \sim \mathcal{N}(0, 1)$$

Not all continuous random variables are absolutely continous. Take for example the P-space of $0-1$ experiments with parameter $p \in [0,1]$.
If we define the random variable

$$X : \Omega \to [0,1], \quad X(\omega) := \sum_{k=1}^{\infty} x_k 2^{-k}$$

Writing a number $b \in (0,1)$ in terms of its binary representation $b = \sum_{k=1}^{\infty} b_k 2^{-k}$ with $b_k \in \{0,1\}$ and writing s_n for the partial sums $s_n = \sum_{k=1}^{n} b_k$, we can see that the distribution function is given by

$$F(b) = \sum_{n=1}^{\infty} b_n p^{s_{n-1}} q^{n-s_{n-1}}$$

where $q = 1-p$. In particular, for $p = \frac{1}{2}$ we get that

$$F(b) = \mathbb{P}_{\frac{1}{2}}[X \leq b] = \sum_{n=1}^{\infty} b_n 2^{-1} = b$$

which is just the identity on $[0,1]$.
However, for $p \neq \frac{1}{2}$ we see something interesting emerge -auDn. The distribution function for X is continuous, but not absolutely continuous: If there would exist some density f_p, then we would have

$$\mathbb{P}_p[X \in A] = \int_A f_p(x)dx$$

in particular, if $\mathbb{P}_{\frac{1}{2}}[X \in A] = \int_A dx = 0$, then we automatically have $\mathbb{P}_p[X \in A] = 0$. But by the the law of big numbers there must exist some A such that

$$\mathbb{P}_p[X \in A] = 1, \quad \mathbb{P}_{\frac{1}{2}}[X \in A] = 0$$

What is interesting is that it took very long until mathematicians found pathological continuous functions. (See Weierstrass's nowhere differentiable function). But such functions come up quite naturally in probability theory.

2.3 Expectation value

Definition 2.16. Let $X \geq 0$ be a random variable on a continuous P-space $(\Omega, \mathcal{A}, \mathbb{P})$ with distribution μ. Its **expectation value** is defined as

$$\mathbb{E}[X] := \int_{\Omega} X(\omega) d\mathbb{P}(\omega) = \int_{\mathbb{R}} x \mu_X(dx) \in [0, \infty]$$

For random variables with negative values, we again do the same construction as in the discrete case. The expectation value is again linear, monotonous and continous.
In many optimisation problems, convexity is an important feature that assures the existence of solutions. We say that a function $g : \mathbb{R} \to \mathbb{R}$ is **convex**, if for every $x_0 \in \mathbb{R}$ there exists a linear supporting function $l(x) = ax + b$ such that

$$l(x) = g(x) \quad \forall x \in \mathbb{R}, \quad \text{and} \quad l(x_0) = g(x_0)$$

if $-g$ is convex, we say g is **concave**

Distribution	$\mathbb{E}[X]$	$\text{Var}[X]$
Binomial(n, p)	np	$np(1-p)$
Hypergeometric(n,N,K)	$n\frac{K}{N}$	$n\frac{K}{N}(1-\frac{K}{N})\frac{N-n}{N-1}$
Poisson(λ)	λ	λ
Geometric(p)	$\frac{1}{p}$	$\frac{1-p}{p^2}$
Uniform(a, b)	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$
Exponential(α)	$\frac{1}{\alpha}$	$\frac{1}{\alpha^2}$
Normal(μ, σ^2)	μ	σ^2

Table 1: Expectation value and variance of important distributions

Proposition 2.17 (Jensen inequality)**.** *For any random variable X with finite expectation value and $g : \mathbb{R} \to \mathbb{R}$ convex, we have*

$$\mathbb{E}[g(X)] \geq g(\mathbb{E}[X])$$

Proof. Let l be the supporting function on $x_0 = \mathbb{E}[X]$. By linearity and monotoneity:

$$g(E[X]) = l(\mathbb{E}[X]) = \mathbb{E}[l(X)] \leq \mathbb{E}[g(X)]$$

□

The Jensen inequality gives us the assurance that our definition of standard deviation $\sigma(X) = \sqrt{\text{Var}[X]} := \sqrt{\mathbb{E}[X^2] - \mathbb{E}[X]^2}$ is indeed well-defined

Proposition 2.18 (Markov inequality)**.** *Let g be a non-netagive, monotonously increasing function on $\mathbb{R}$. The for every c with $g(c) > 0$:*

$$\mathbb{P}[X \geq c] \leq \frac{\mathbb{E}[g(X)]}{g(c)}$$

Proof. We just use the characteristic function and the properites of g to get

$$\mathbb{1}_{[X \geq c]} \leq \frac{g(X)}{g(c)}$$

and take the expectation value of the above. □

The most notable use of the Markov inequality is the **Chebychev inequality**

$$\mathbb{P}\left[|X - \mathbb{E}[X]| > c\right] \leq \frac{\text{Var}[X]}{c^2}$$

which follows by application on the random variable $Y = |X - \mathbb{E}[X]|$ and the function $g(x) = (\max(x, 0))^2$

2.4 Multiple random variables

If we have multiple random variables $X_1,\ldots,X_n$ then we can view it as a single random variable $\boldsymbol{X}$ with values in $\mathbb{R}^n$.
With the Borel σ-Algebra on $\mathbb{R}^n$, we can write

$$\boldsymbol{X}^{-1}(A_1\times\ldots\times A_n)=\bigcup_{i=1}^n X_i^{-1}(A_i)$$

which allows us to define the distriubtion $\mu_{\boldsymbol{X}}$ under $\mathbb{P}$ as the **collective distribution** of $X_1,\ldots,X_n$ given by

$$\mu_{\boldsymbol{X}}(A)=\mathbb{P}[\boldsymbol{X}^{-1}(A)]=\mathbb{P}[\{\omega|\boldsymbol{X}(\omega)\in A\}]=\mathbb{P}[\boldsymbol{X}\in A]\quad\text{for}\quad A\in\mathcal{B}^n\subseteq\mathbb{R}^n$$

Just like in the 1-dimensional case: If every X_i is discrete, then the image $\boldsymbol{X}(\Omega)$ is countable and we define

$$\mu_{\boldsymbol{X}}(A)=\sum_{\boldsymbol{x}\in\boldsymbol{X}(\Omega)\cap A}\mathbb{P}[\boldsymbol{X}=\boldsymbol{x}]=\sum_{\substack{(x_1,\ldots,x_n)\in A\\ x_i\in X_i(\Omega)}}\mathbb{P}[X_i=x_i]_{i\in I}$$

If the collective distribution **absolutely continuous**, i.e. if there exists a measurable function $f:\mathbb{R}^n\to\mathbb{R}_{\geq 0}$ such that

$$\mu_{\boldsymbol{X}}(A)=\int_A f(\boldsymbol{x})d\boldsymbol{x}$$

and call this function f the **density function** of $\boldsymbol{X}$.
Given the collective distribution, we can obtain a distribution for a single X_i. The **marginal distribution** is the distribution obtained by

$$\mu_{X_i}(B)=\mathbb{P}[X_i\in B]=\mu_{\boldsymbol{X}}(\mathbb{R}\times\ldots\times B\times\ldots\times\mathbb{R})\quad\text{for }B\in\mathcal{B}$$

Note that the marginal distribution does not uniquely determine the collective distribution. The missing information is the (in-)dependence of the random variables of the components of $\boldsymbol{X}$.

Definition 2.19. The random variables $X_1,\ldots,X_n$ are (stochastically) **indepenent**, if for all $A_1,\ldots,A_n\in\mathcal{B}$

$$\mathbb{P}[X_1\in A_1,\ldots,X_n\in A_n]=\prod_{i=1}^n\mathbb{P}[X_i\in A_i]$$

or equivalently

$$\mu_{\boldsymbol{X}}\left(\prod_{i=1}^n A_i\right)=\prod_{i=1}^n\mu_{X_i}(A_i)$$

Example 2.20. We define the **standard normal distribution** for independent random variables $X_1,\ldots,X_n$ using the $N(0,1)$ distribution. The collective distribution then has density

$$f(\boldsymbol{x})=(2\pi)^{-n/2}\exp\left(-\frac{1}{2}\sum_{i=1}^n x_i^2\right)=(2\pi)^{-n/2}e^{-\frac{1}{2}|\boldsymbol{x}|^2}$$

Analogously to the one-dimensional case we can define the push-forwards for a random variable along a **measurable map** $g : (\mathbb{R}^n, \mathcal{B}^n) \to (\mathbb{R}^m, \mathcal{B}^m)$ as

$$\mu_{\boldsymbol{Y}}(A) = \mu_{g \circ \boldsymbol{X}} := \mu_{\boldsymbol{X}}(g^{-1}(A)) \quad \text{for } A \in \mathcal{B}^m$$

Proposition 2.21. *Let $g : \mathbb{R}^n \to \mathbb{R}^m$ be linear and invertible: $g(\boldsymbol{x}) = \boldsymbol{m} + B\boldsymbol{x}$ with $\det B \neq 0$. If $\mu_{\boldsymbol{x}}$ is absolutely continuous, then so is $\mu_{\boldsymbol{Y}}$ and its density is given by*

$$f_{\boldsymbol{Y}}(\boldsymbol{y}) = \frac{1}{|\det B|} f_{\boldsymbol{X}}\left(B^{-1}(\boldsymbol{y} - \boldsymbol{m})\right)$$

Proof. This follows fromthe subtitution rule with $\boldsymbol{x} = B^{-1}(\boldsymbol{y} - \boldsymbol{m})$ as

$$\mu_{\boldsymbol{Y}}(A) = \mu_{\boldsymbol{X}}(g^{-1}(A)) = \int_{g^{-1}(A)} f_{\boldsymbol{X}}(\boldsymbol{x}) d\boldsymbol{x} = \int_A f_{\boldsymbol{X}}(g^{-1}(\boldsymbol{y})) \frac{1}{|\det B|} d\boldsymbol{y}$$

□

Example 2.22. An important example is the push-forward of the n-dimensional standard normal distribution with the substitution $\boldsymbol{Y} = \boldsymbol{m} + B\boldsymbol{X}$. It has density

$$f_{\boldsymbol{Y}}(\boldsymbol{y}) = (2\pi)^{-n/2} \frac{1}{\sqrt{|\det \Sigma|}} \exp\left(-\frac{1}{2}(\boldsymbol{y} - \boldsymbol{m})^T \Sigma^{-1} (\boldsymbol{y} - \boldsymbol{m})\right)$$

where $\Sigma = BB^T$. This gives us is the generalized n-dimensional Nromal distribution $N_n(\boldsymbol{m}, \Sigma)$.

2.5 Covariance and Correlation

Let $g : (\mathbb{R}^n, \mathcal{B}^n) \to (\mathbb{R}, \mathcal{B})$ be measureable. Instead of computing $\mathbb{E}[g \circ X]$ directly, we just use

$$\mathbb{E}[g \circ \boldsymbol{X}] = \int_{\mathbb{R}^n} g(\boldsymbol{x}) \mu_X(d\boldsymbol{x})$$

which in the discrete case is

$$\mathbb{E}[g \circ \boldsymbol{X}] = \sum_{x_i \in X_i(\omega)} g(x_1, \dots, x_n) \mathbb{P}[X_i = x_i]_{i \in I}$$

or in the absolutely continuous case

$$\mathbb{E}[g \circ \boldsymbol{X}] = \int_{\mathbb{R}^n} g(\boldsymbol{x}) f_{\boldsymbol{X}}(\boldsymbol{x}) d\boldsymbol{x}$$

We can use this to define

Definition 2.23. The **Covariance** of random variables X_1, X_2 is defined as

$$\boxed{\operatorname{Cov}(X_1, X_2) := \mathbb{E}\left[(X_1 - \mathbb{E}[X_1])(X_2 - \mathbb{E}[X_2])\right]}$$

Proposition 2.24. *The covariance fulfills the following relations:*

(a) $\operatorname{Cov}(X, X) = \operatorname{Var}[X]$

(b) $\operatorname{Cov}(X_1, X_2) = \operatorname{Cov}(X_2, X_1)$

(c) $\operatorname{Cov}(X_1, X_2) = \mathbb{E}[X_1 X_2] - \mathbb{E}[X_1]\mathbb{E}[X_2]$

(d) $\operatorname{Cov}(X_1, aX_2 + b) = a \operatorname{Cov}(X_1, X_2)$

(e) $\operatorname{Cov}(X_1, X_2 + X_3) = \operatorname{Cov}(X_1, X_2) + \operatorname{Cov}(X_1, X_3)$

(f) $\operatorname{Var}[X_1 + X_2] = \operatorname{Var}[X_1] + \operatorname{Var}[X_2] + 2\operatorname{Cov}(X_1, X_2)$

(g) $|\operatorname{Cov}(X_1, X_2)| \leq \sigma(X_1)\sigma(X_2)$

(h) *If* X_1, X_2 *are independent, then* $\operatorname{Cov}(X_1, X_2) = 0$*. In particular we then have* $\operatorname{Var}[X_1 + X_2] = \operatorname{Var}[X_1] + \operatorname{Var}[X_2]$.

Note that the converse of (h) is wrong. If we set X_1 corresponding to the normal distribution $N(0, 1)$ and $X_2 = X_1^2$, then

$$\operatorname{Cov}(X_1, X_2) = \mathbb{E}\left[(X_1 - \mathbb{E}[X_1])(X_1^2 - \mathbb{E}[X_2]\right] = \mathbb{E}[X_1 X_2] = \mathbb{E}[X_1^3] = 0$$

Remark 2.25. There is alot of similarity between covariance and the standard scalar product in $\mathbb{R}^n$.

Random Variables	Vectors in $\mathbb{R}^n$
$\operatorname{Var}[X] := \mathbb{E}[(X - \mathbb{E}[X])^2]$	$\|v\|^2 := \sum_i v_i^2$
$\operatorname{Cov}(X, Y) := \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])]$	$\langle v, w\rangle := \sum_i v_i w_i$
$\operatorname{Cov}(X, X) = \operatorname{Var}[X]$	$\langle v, v\rangle = \|v\|^2$
$\operatorname{Cov}(X, Y) = \operatorname{Cov}(Y, X)$	$\langle v, w\rangle = \langle w, v\rangle$
$\operatorname{Cov}(\lambda X, Y) = \lambda \operatorname{Cov}(X, Y)$	$\langle \lambda v, w\rangle = \lambda\langle v, w\rangle$
$\operatorname{Cov}(X + Y, Z) = \operatorname{Cov}(X, Z) + \operatorname{Cov}(Y, Z)$	$\langle u + v, w\rangle = \langle u, w\rangle + \langle v, w\rangle$
$\operatorname{Cov}(X, Y + \mu) = \operatorname{Cov}(X, Y)$	No analogue
$\operatorname{Var}[X + Y] = \operatorname{Var}[X] + \operatorname{Var}[Y] + 2\operatorname{Cov}(X, Y)$	$\|v + w\|^2 = \|v\|^2 + \|w\|^2 + 2\langle v, w\rangle$
$\|\operatorname{Cov}(X, Y)\| \leq \sigma(X)\sigma(Y)$	(Cauchy-Schwarz)

Table 2: Similarities and differences between Covariance and scalar products

Definition 2.26. Given two random Variables X_1, X_2 the **correlation** between them is

$$\boxed{\rho(X_1, X_2) := \frac{\operatorname{Cov}(X_1, X_2)}{\sigma(X_1)\sigma(X_2)}}$$

if $\rho(X_1, X_2) = 0$, we say that X_1, X_2 are uncorrelated.

Correlation measures strength and direction of *linear dependence* between random variables.
It's 2021 and the public media is no stranger to the words *correlation*, *standard deviation*, *statistics*, *median* etc. Unfortunately, the numbers are often misused to support statements that aren't actually supported by the data and it is our responsibility to be precise in our usage of statistics.
We often have to ask ourselves what the underlying P-space is to understand what these words mean. Depending on how the P-space is chosen, one can come to wildly idfferent conclusions with the same raw data.

3 Limit theorems

Let $(X_i)_{i\in I}$ be a sequence of independent random variables. If we set $S_n = \sum_{i=1}^n X_i$, then we expect that for large n, the average value $\frac{S_n}{n}$ approaches the arithmetic mean of the $\mathbb{E}[X_i]$. But how fast would it converge to the mean and what happens if the X_i are not independent

3.1 Weak law of big numbers

Assume all X_i have the same expectationvalue $\mathbb{E}[X_i] = m$. We say that the **weak law of big numbers** holds in some P-space, if for all $\epsilon > 0$

$$\mathbb{P}\left[|\frac{S_n}{n} - m| > \epsilon\right] \to 0 \quad \text{for } n \to \infty$$

Using the Chebyshev Inequality we get with $\mathbb{E}[S_n] = nm$

$$\mathbb{P}\left[|\frac{S_n}{n} - m| > \epsilon\right] \leq \frac{\mathrm{Var}[S_n/n]}{\epsilon^2} = \frac{\mathrm{Var}[S_n]}{n^2\epsilon^2}$$

The reason why we cannot always use the law of weak numbers is that the variance $\mathrm{Var}[S_n/n]$ might not exist. If $\mathbb{E}[X_i^2] < \infty$ then it does exist and the law of weak numbers holds.

Example 3.1. A counter example can be given by the **Cauchy-distribution**

$$f(x) = \frac{1}{\pi}\frac{1}{1+x^2}$$

Then the $\mathbb{E}[|X_i|] = \infty$ and we can show that $\frac{S_n}{n}$ again has a Cauchy-distribution. This means that $\frac{S_n}{n}$ has increasing variance for larger and larger n.

We can use the weak law of big numbers to prove Weierstrass's theorem, which states that the polynomials are dense in the set of continuous functions on a compact interval (with the $\|\|_\infty$ norm)
We start by defining the **Bernstein-Polynomials** of degree n on $[0,1]$ as

$$B_{n,k}(x) = \binom{n}{k} x^k (1-x)^{n-k} \quad \text{for} \quad k = 0, \dots n$$

A function $f \in C([0,1])$ now can be approximated to n-th degree by the linear combination

$$B_n^f(x) = \sum_{k=0}^n f(\tfrac{k}{n}) B_{n,k}(x)(x)$$

This really does approximate f, because we know that

$$B_{n,k}(x) = \mathbb{P}_X[S_n = k] \quad \text{for} \quad S_n = \text{ number of successes after } n \text{ throws with parameter } x$$

From this, we see that $B_n^f(x) = \mathbb{E}_x[f(\frac{S_n}{n})]$. But by the weak law of big numbers $\frac{S_n}{n}$ converges to the success parameter x.
This shows that probabilistic arguments can prove useful results from Analysis.

3.2 Strong law of big numbers

Instead of taking the arithmetic mean of all S_n, we can instead only look at the arithmetic mean of the S_k after some point n. So we would like to prove for all $\epsilon > 0$

$$\lim_{n\to\infty} \mathbb{P}\left[\bigcap_{k\geq n} \left\{ |\frac{S_k}{k} - m| \leq \epsilon \right\}\right] = 1$$

i.e. that after some time n, the aritmetic mean stas close to the expectation value m.

4 Limit theorems

4.1 Weak Law of Large Numbers

For the following, assume that we have a collection of random variables X_i that have the same expectation value $\mathbb{E}[X_i] = c \in \mathbb{R}$.

Definition 4.1. We say that the random varialbes follow the **weak law of large numbers**, if for all $\epsilon > 0$

$$\mathbb{P}\left[|\frac{S_n}{n} - c| > \epsilon\right] \to 0 \quad \text{as} \quad n \to \infty$$

5 Statistics

5.1 Point estimates

Given some data $\boldsymbol{x} = (x_1, \dots, x_n)$, we want to find out what random variables $\boldsymbol{X} = (X_1, \dots, X_n)$ could have generated this data.
To do so, we consider a collection of distributions $(\mu_\theta)_{\theta\in\Theta}$ on P-spaces $(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n), \mathbb{P}_\theta)_{\theta\in\Theta}$.
A **point estimate** is therefore a function

$$T : \mathbb{R}^n \to \Theta$$

which to every possible dataset $\boldsymbol{x}$ attributes some P-space $(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n), \mathbb{P}_\theta)$ with random variables $\boldsymbol{X}$ and their corresponding distributions μ_θ.
Usually, we put some structure on Θ.

6 Appendix

6.1 Distributions

Poisson Distribution with parameter λ

$$\Omega = \{0, 1, 2, \dots\}, \quad \mathbb{P}[X = n] = e^{-\lambda}\frac{\lambda^n}{n!}$$

$$X \sim \text{Poi}(\lambda), Y \sim \text{Poi}(\mu) \text{ independent} \implies X + Y \sim \text{Poi}(\lambda + \mu)$$

Proof. For $S = X + Y$

$$\begin{aligned}
\mathbb{P}[S = s] &= \sum_{k=0}^{s} \mathbb{P}[X = k, Y = s - k] \\
&= \sum_{k=0}^{s} \mathbb{P}[X = k] \cdot \mathbb{P}[Y = s - k] \\
&= \sum_{k=0}^{s} e^{-\lambda} \frac{\lambda^s}{s!} e^{-\mu} \frac{\mu^{k-s}}{(s-k)!} \\
&= \frac{e^{-\lambda+\mu}}{s!} \sum_{k=0}^{s} \binom{s}{k} \lambda^k \mu^{s-k} \\
&= \frac{e^{-\lambda+\mu}}{s!} (\lambda + \mu)^s
\end{aligned}$$

□

6.2 Theorems

Borel Cantelli

Let $A_1, A_2, \ldots \in \mathcal{A}$. Set

$$A_\infty := \limsup_{n=\infty} = \bigcap_{n=1}^{\infty} \bigcup_{k=1}^{n} A_k$$

then

(a) $\sum_{k=1}^{\infty} \mathbb{P}[A_k] < \infty \implies \mathbb{P}[A_\infty] = 0$

(b) If the $(A_k)_{k \in \mathbb{N}}$ are independent, then $\sum_{k=1}^{\infty} \mathbb{P}[A_k] = \infty \implies \mathbb{P}[A_\infty] = 1$.

6.3 Gaussian/Normal Distribution with parameters μ, σ^2

For $\mu, \sigma \in \mathbb{R}, \sigma > 0$ the density for $\mathcal{N}(\mu, \sigma^2$ is given by

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

μ is called the **mean** and for $X \sim \mathcal{N}(\mu, \sigma^2)$ is the expectation value $\mathbb{E}[X] = \mu$. σ is called the **standard deviation** and $\sigma^2 =: \mathrm{Var}(X)$ the **variance**

Lemma 6.1. *For $X \sim \mathcal{N}(\mu_X, \sigma_X^2)$ and $Y \sim \mathcal{N}(\mu_Y, \sigma_Y^2)$ independent, their sum is also normal distributed with parameters*

$$Z = X + Y \sim \mathcal{N}(\mu_X + \mu_Y, \sigma_X^2 + \sigma_Y^2)$$

Proof. Since $\mathbb{P}[Z = z] = \mathbb{P}[X = x] \cdot \mathbb{P}[Y = z - x]$, the density of Z is the convolution of the densities of X and Y.

□