2S(3) Statistics – Part I – Spring 2026

Lecture Notes by Dr Nikolaos Fountoulakis

16 January 2026

Contents

Chapter 1

The concept of a probability space

1.0.1 The space of events

What is a probability space? Intuitively, it is a set whose elements are equipped with weights. Events are subsets of this set. That is, a collection of events is a collection of subsets of this set. However, such a collection must satisfy certain conditions. We need to be able to create *composite* events out of basic set operations. If we talk about/consider an event (subset), then we should be able to consider its complement as well. In other words, if a subset is an event, then its complement is an event too. Also, if two subsets A and B are events, then their *union* $A \cup B$ should be an event too. Furthermore, any *countable union* of events should be an event.

More formally, let Ω be a set and let $\mathcal{F}$ be a family of subsets of Ω which satisfies the following properties:

A1 $\Omega \in \mathcal{F}$

A2 if $A \in \mathcal{F}$, then $A^c \in \mathcal{F}$, where $A^c = \Omega \setminus A$

A3 if $A, B \in \mathcal{F}$, then $A \cup B \in \mathcal{F}$

The family $\mathcal{F}$ is called an *algebra*.

If $\{A_n\}_{n\in\mathbb{N}}$ is a countably infinite collection of subsets, the (infinite) *union* of it is denoted by $\cup_{n=1}^{\infty} A_n$ and is consists of those elements s for which there exists some $n \in \mathbb{N}$ such that $s \in A_n$. (Note that this n depends on s and may not be the same for every s.) Now, if $\mathcal{F}$ satisfies the following *instead of A3*

A3′ if $A_n \in \mathcal{F}$, for each $n \in \mathbb{N}$, then $\cup_{n=1}^{\infty} A_n \in \mathcal{F}$,

then $\mathcal{F}$ is called a σ*-algebra*.

Intuitively, a σ-algebra is a collection of events/subsets of a space Ω. The space Ω is usually called the *sample space*. The definition of $\mathcal{F}$ ensures that one can "talk" about these events under any finite or at most countably infinite set-theoretic "sentence", that is, set-theoretic operation. The collection $\mathcal{F}$ may not contain every possible subset of the space Ω, but at least it is closed under these operations. That is, if one takes a collection of members/sets in $\mathcal{F}$ and performs a set-theoretic operation, then the outcome will be a set/member of $\mathcal{F}$ too.

Thus, a σ-algebra is not necessarily a set system that is *complete* in that it contains all possible events/subsets of Ω, but it is *consistent* in that set-theoretic operations on that set system have their outcomes inside it.

Definition. Let Ω be a set and let $\mathcal{F}$ be a σ-algebra of subsets of Ω. The pair $(\Omega, \mathcal{F})$ is called a *measurable* space and the members of $\mathcal{F}$ are called *measurable* sets.

A measurable space may be rich or poor regarding the variety of its measurable subsets.

Example. The smallest possible σ-algebra of events on a sample space Ω is $\mathcal{F} = \{\emptyset, \Omega\}$. It is easy to see that it satisfies all four axioms (verify!). This is the coarsest set of events on the sample space Ω. If we take any subset $\emptyset \subset A \subset \Omega$, then A is not measurable in this σ-algebra.

Example. Let $A \subset \Omega$ and take $\mathcal{F} = (\emptyset, A, A^c, \Omega)$. This is also a σ-algebra. This is a somewhat richer measure space compared to the previous one. However, any set B with $\emptyset \subset B \subset \Omega$ and $B \neq A$ is not measurable. That is, this space is all about A and nothing else.

Example. The powerset of Ω (that is, the set of all subsets of Ω, including Ω itself) is a σ-algebra too.

Example. Consider $\mathbb{R}$ and let $\mathcal{O}$ be the collection of open intervals (a, b) for $a < b$. The σ-algebra that contains all sets derive by any countable sequence of applications of the rules A1– A4 is called the *Borel* σ-algebra of $\mathbb{R}$. We will see in the examples class, that this collection of sets contains all "reasonable" intervals you have dealt with in your courses: closed, one-way infinite, semi-open...

The four axioms A1 – A4 are enough to guarantee that other set-theoretic operations are compatible with a σ-algebra. For example, the following holds:

Lemma 1.1. *Let $\mathcal{F}$ be an algebra on a set Ω. For any $A, B \in \mathcal{F}$, we have*

1. $A \cap B \in \mathcal{F}$.
2. $A \setminus B \in \mathcal{F}$.
3. $A \triangle B \in \mathcal{F}$ *(This is called the* symmetric difference *of A and B and consists of all those elements that belong to* exactly one *of A and B).*

Proof. For 1. we use de Morgan's law: if $(S_1 \cup S_2)^c = S_1^c \cap S_2^c$, for any subsets S_1, S_2. Taking $S_1 = A^c$ and $S_2 = B^c$, the above implies that $(A^c)^c \cap (B^c)^c = (A^c \cup B^c)^c$. But $(A^c)^c = A$ and $(B^c)^c = B$, whereby $(A^c)^c \cap (B^c)^c = A \cap B$. But if $A \in \mathcal{F}$, then $A^c \in \mathcal{F}$ (by A2) (so does B^c). Thus, by A3, $A^c \cup B^c \in \mathcal{F}$ and by A2 $(A^c \cup B^c)^c \in \mathcal{F}$. Thus, $A \cap B \in \mathcal{F}$.

The proofs of 2. and 3. are derived similarly and are left as exercises. □

Note that above does not use the full strength of a σ-algebra (that is, axiom A4) but only that of an algebra of subsets. If $\mathcal{F}$ is a σ-algebra of events, then it also satisfies the following.

Lemma 1.2. *Let $\{A_n\}_{n\in\mathbb{N}}$ be a collection of members of a σ-algebra $\mathcal{F}$ on a set Ω. Then*

$$\cap_{n\in\mathbb{N}} A_n \in \mathcal{F}.$$

(The set $\cap_{n\in\mathbb{N}} A_n$ consists of those elements $s \in \Omega$ such that $s \in A_n$ for all $n \in \mathbb{N}$.)

Proof. The proof of this statement need an infinite version of de Morgan's law. That is, for any collection $\{S_n\}_{n\in\mathbb{N}}$ one has

$$(\cup_{n\in\mathbb{N}} S_n)^c = \cap_{n\in\mathbb{N}} S_n^c.$$

Taking $S_n = A_n^c \in \mathcal{F}$ and with $S_n^c = (A_n^c)^c = A_n$, the statement follows as in the above lemma. □

1.0.2 Probability measures on a space of events

We would like to think of probabilities as a weighting scheme over a measurable space. In other words, a scheme that assigns to the measurable sets. Of course, such a scheme needs to satisfy certain plausible conditions. In particular, let $(\Omega, \mathcal{F})$ be a measurable space. Let $\mathbb{P}(\cdot)$ denote the weighting scheme. More specifically, $\mathbb{P} : \mathcal{F} \to \mathbb{R}^+$ is a function that assigns to the members of $\mathcal{F}$ a non-negative real number. We can assume that this function is *normalised* and assigns $\mathbb{P}(\Omega) = 1$. Furthermore, if $A, B \in \mathcal{F}$ are disjoint, then weight/mass of their union is equal to the sum of their weights/masses. (By axiom A3, the union is in $\mathcal{F}$ too.) More formally, if $A \cap B = \emptyset$, then $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B)$. It is also plausible that this holds for a countably infinite collection of sets in $\mathcal{F}$. If $\{A_n\}_{n\in\mathbb{N}}$ is a sequence of sets in $\mathcal{F}$ such that $A_i \cap A_j = \emptyset$, for any $i \neq j$, then $\mathbb{P}(\cup_{n\in\mathbb{N}} A_n) = \sum_{n\in\mathbb{N}} \mathbb{P}(A_n)$. We summarise these into the following.

Definition. Let $(\Omega, \mathcal{F})$ be a measurable space. A function $\mathbb{P} : \mathcal{F} \to \mathbb{R}^+$ is called a *probability measure*, if

P1 for any $A \in \mathcal{F}$ we have $0 \leq \mathbb{P}(A) \leq 1$.

P2 $\mathbb{P}(\Omega) = 1$ and $\mathbb{P}(\emptyset) = 0$.

P3 Let $\{A_n\}_{n\in\mathbb{N}}$ be a countable collection of sets in $\mathcal{F}$ that are pairwise disjoint, that is, $A_i \cap A_j = \emptyset$, for any $i \neq j$. Then

$$\mathbb{P}(\cup_{n\in\mathbb{N}} A_n) = \sum_{n\in\mathbb{N}} \mathbb{P}(A_n).$$

This property is called *countable additivity*.

The triple $(\Omega, \mathcal{F}, \mathbb{P})$ is called a *probability space*. The members of $\mathcal{F}$ are called *events*. It is those subsets of Ω that $\mathbb{P}$ has assigned weights.

We will show that P3 implies what is called *finite additivity*. This is summarised in the following lemma.

Lemma 1.3. *Let $A_1, \ldots, A_k$ in $\mathcal{F}$ be a finite collection of pairwise disjoint sets. Then*

$$\mathbb{P}\left(\cup_{n=1}^k A_n\right) = \mathbb{P}(A_1) + \cdots + \mathbb{P}(A_k). \tag{1.1}$$

Proof. This follows from P3. Take the infinite collection $\{B_n\}_{n\in\mathbb{N}}$, where $B_n = A_n$, for $n = 1, \ldots, k$ and $B_n = \emptyset$, for $n > k$. Then for $i \neq j$, we have $B_i \cap B_j = \emptyset$. So $\cup_{n\in\mathbb{N}} B_n = \cup_{n=1}^k A_k$. Applying P3, we obtain

$$\mathbb{P}\left(\cup_{n=1}^k A_k\right) = \mathbb{P}(\cup_{n\in\mathbb{N}} B_n) = \sum_{n=1}^{\infty} \mathbb{P}(B_n) = \sum_{n=1}^{k} \mathbb{P}(B_n) = \sum_{n=1}^{k} \mathbb{P}(A_n),$$

as $\mathbb{P}(B_n) = 0$, for all $n > k$. □

A corollary of (1.1) is the following

Corollary 1.4. *For any subset $A \in \mathcal{F}$ we have*

$$\mathbb{P}(A) + \mathbb{P}(A^c) = 1. \tag{1.2}$$

Proof. As $A \cap A^c = \emptyset$, by (1.1) implies that

$$\mathbb{P}(A \cup A^c) = \mathbb{P}(A) + \mathbb{P}(A^c).$$

But also $A \cup A^c = \Omega$, whereby

$$\mathbb{P}(A \cup A^c) = \mathbb{P}(\Omega) = 1.$$

□

Now, take two sets $A, B \in \mathcal{F}$. We will write $A \cup B$ as the union of three pairwise disjoint sets:

$$A \cup B = (A \setminus B) \cup (B \setminus A) \cup (A \cap B).$$

Therefore,

$$\mathbb{P}(A \cup B) = \mathbb{P}(A \setminus B) + \mathbb{P}(B \setminus A) + \mathbb{P}(A \cap B).$$

But also, we can write $A = (A \setminus B) \cup (A \cap B)$ and these two sets are disjoint. So,

$$\mathbb{P}(A) = \mathbb{P}(A \setminus B) + \mathbb{P}(A \cap B).$$

Similarly, we can write

$$\mathbb{P}(B) = \mathbb{P}(B \setminus A) + \mathbb{P}(A \cap B).$$

Substituting these two into the former relation we deduce:

$$\begin{aligned} \mathbb{P}(A \cup B) &= (\mathbb{P}(A) - \mathbb{P}(A \cap B)) + (\mathbb{P}(B) - \mathbb{P}(A \cap B)) + \mathbb{P}(A \cap B) \\ &= \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B). \end{aligned} \tag{1.3}$$

This is like the inclusion-exclusion principle!!! In fact, this holds in its full generality: let $A_1, \ldots, A_n$, with $n \geq 2$, be sets that are members of $\mathcal{F}$. Then

$$\begin{aligned} \mathbb{P}(A_1 \cup \cdots \cup A_n) = &\sum_{i=1}^{n} \mathbb{P}(A_i) - \sum_{i_1 < i_2} \mathbb{P}(A_{i_1} \cap A_{i_2}) + \sum_{i_1 < i_2 < i_3} \mathbb{P}(A_{i_1} \cap A_{i_2} \cap A_{i_3}) - \cdots \\ &+ (-1)^{n+1} \mathbb{P}(A_1 \cap \cdots \cap A_n). \end{aligned} \tag{1.4}$$

Proof of (1.4)**.** We will prove (1.4) by induction on n. The base case is for $n = 2$. Then this is simply (1.3).

Suppose now that this is true for $n = k$ (induction hypothesis). We will show that this holds for $n = k + 1$. We write

$$\begin{aligned} \mathbb{P}(A_1 \cup \cdots \cup A_{k+1}) &= \mathbb{P}((A_1 \cup \cdots \cup A_k) \cup A_{k+1}) \\ &= \sum_{i=1}^{k} \mathbb{P}(A_i) - \sum_{i_1 < i_2 \leq k} \mathbb{P}(A_{i_1} \cap A_{i_2}) + \sum_{i_1 < i_2 < i_3 \leq k} \mathbb{P}(A_{i_1} \cap A_{i_2} \cap A_{i_3}) - \cdots \\ &+ (-1)^{k+1} \mathbb{P}(A_1 \cap \cdots \cap A_k) + \mathbb{P}(A_{k+1}) \\ &- \mathbb{P}((A_1 \cup \cdots \cup A_k) \cap A_{k+1}) \end{aligned} \tag{1.5}$$

having used the induction hypothesis and (1.3). Now, note that that we can write

$$(A_1 \cup \cdots \cup A_k) \cap A_{k+1} = (A_1 \cap A_{k+1}) \cup \cdots \cup (A_k \cap A_{k+1}).$$

Applying the induction hypothesis to the latter we have

$$\begin{aligned}\mathbb{P}\left((A_1 \cup \cdots \cup A_k) \cap A_{k+1}\right) &= \mathbb{P}\left((A_1 \cap A_{k+1}) \cup \cdots \cup (A_k \cap A_{k+1})\right) \\ &= \sum_{i=1}^{k} \mathbb{P}\left(A_i \cap A_{k+1}\right) - \sum_{i_1<i_2\leq k} \mathbb{P}\left(A_{i_1} \cap A_{i_2} \cap A_{k+1}\right) \\ &+ \sum_{i_1<i_2<i_3\leq k} \mathbb{P}\left(A_{i_1} \cap A_{i_2} \cap A_{i_3} \cap A_{k+1}\right) - \cdots \\ &+ (-1)^{k+1}\mathbb{P}\left(A_1 \cap \cdots \cap A_k \cap A_{k+1}\right).\end{aligned} \tag{1.6}$$

Substituting (1.6) in (1.5) we deduce (1.4). □

Example (Finite spaces - counting probability measure)**.** This is a fundamental example. Let Ω be a finite set of size $n \geq 1$. Take $\mathcal{F}$ to be the powerset of Ω which consists of all subsets of Ω (including Ω). We will define a probability measure on $(\Omega, \mathcal{F})$ as follows. For any $S \subseteq \Omega$ (that is, $S \in \mathcal{F}$) we set $N(S) = |S|/n$. It is a straightforward exercise to see that N satisfies axioms P1 – P3.

1.0.3 Conditional probability measures

Another fundament example of a probability space is that of a conditional space. Consider a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and take a set $B \in \mathcal{F}$ such that $\mathbb{P}(B) > 0$.

We will define a probability space on B as follows. Take $\mathcal{F}_B = \{A \cap B \ : \ A \in \mathcal{F}\}$, that is, the intersection of every set/ member of $\mathcal{F}$ with B. Note that to every member of $\mathcal{F}_B$ corresponds a member of $\mathcal{F}$. By the definition of $\mathcal{F}_B$, for every *non-empty* set $\tilde{A} \in \mathcal{F}_B$ there exists some $A \in \mathcal{F}$ such that $\tilde{A} = A \cap B$. For $\tilde{A} \in \mathcal{F}_B$ we let $A \in \mathcal{F}$ be the corresponding set. This correspondence is not unique if $\tilde{A} = \emptyset$. This is the case since there might be multiple sets $A \in \mathcal{F}$ with $A \cap B = \emptyset$.

We *define* a probability measure on $(B, \mathcal{F}_B)$ as follows.

Definition. For $A \in \mathcal{F}$ the *conditional probability measure* of A given B is defined to be

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}. \text{ (Bayes' formula)}$$

The conditional measure defines a probability measure inside B. We denote this by $\mathbb{P}_B(\cdot)$ In particular, take $\tilde{A} \in \mathcal{F}_B$. We set

$$\mathbb{P}_B\left(\tilde{A}\right) = \mathbb{P}(A|B).$$

We can show that this is a indeed a probability measure on $(B, \mathcal{F}_B)$. Indeed:

1. P1 is satisfied as for any *non-empty* $\tilde{A} \in \mathcal{F}_B$ we have

$$\mathbb{P}_B\left(\tilde{A}\right) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)} \leq 1,$$

since $\mathbb{P}(A \cap B) \leq B$. Also, since both $\mathbb{P}(A \cap B), \mathbb{P}(B) \geq 0$, it follows that $\mathbb{P}_B\left(\tilde{A}\right) \geq 0$.

2. For P2, we have
$$\mathbb{P}_B(B) = \frac{\mathbb{P}(B \cap B)}{\mathbb{P}(B)} = \frac{\mathbb{P}(B)}{\mathbb{P}(B)} = 1$$
and
$$\mathbb{P}_B(\emptyset) = \frac{\mathbb{P}(\emptyset \cap B)}{\mathbb{P}(B)} = \frac{\mathbb{P}(\emptyset)}{\mathbb{P}(B)} = 0,$$
since $\mathbb{P}(\emptyset) = 0$.

3. To show P3, consider a sequence of pairwise disjoint sets $\tilde{A}_1, \tilde{A}_2, \cdots$ in $\mathcal{F}_B$. With A_i being such that $\tilde{A}_i = A_i \cap B$ we have
$$\begin{aligned}\mathbb{P}_B\left(\cup_{i=1}^{\infty}\tilde{A}_i\right) &= \frac{\mathbb{P}\left((\cup_{i=1}^{\infty}A_i)\cap B\right)}{\mathbb{P}(B)} = \frac{\mathbb{P}\left(\cup_{i=1}^{\infty}(A_i \cap B)\right)}{\mathbb{P}(B)} = \sum_{i=1}^{\infty}\frac{\mathbb{P}(A_i \cap B)}{\mathbb{P}(B)} \\ &= \sum_{i=1}^{\infty}\mathbb{P}_B\left(\tilde{A}_i\right).\end{aligned} \tag{1.7}$$

What does it mean for the counting probability measure?

Consider again Ω - a finite set of size $n \geq 1$. Take $\mathcal{F}$ to be the powerset of Ω which consists of all subsets of Ω and let B be such a subset. Consider the counting measure that assigns weight $N(A) = |A|/n$ to any set $A \subseteq \Omega$. In other words, it is the fraction of Ω that A occupies.

The definition of the conditional counting probability measure of a subset A given B yields
$$N(A|B) = \frac{N(A \cap B)}{N(B)} = \frac{|A \cap B|/n}{|B|/n} = \frac{|A \cap B|}{|B|}.$$
The meaning of this is the fraction of B that A (or, more precisely, $A \cap B$) occupies.

1.0.4 Products of probability spaces

A very useful and frequently occurring case of a probability space is that of the product of two probability spaces. Let $(\Omega_1, \mathcal{F}_1, \mathbb{P}_1)$ and $(\Omega_2, \mathcal{F}_2, \mathbb{P}_2)$ be two probability spaces. The product of the two sets Ω_1 and Ω_2 is defined as
$$\Omega_1 \times \Omega_2 := \{(\omega_1, \omega_2) \ : \ \omega_1 \in \Omega_1, \omega_2 \in \Omega_2\}.$$
In other words, it consists of all ordered pairs where the first element of every such pair belongs to Ω_1 and the second one to Ω_2. This is the set of elements of this probability space.

The next step is to define a σ-algebra of subsets of the set $\Omega_1 \times \Omega_2$. There is an obvious (well, at least plausible) candidate which is
$$\mathcal{F}_1 \times \mathcal{F}_2 = \{A_1 \times A_2 \ : \ A_1 \in \mathcal{F}_1, A_2 \in \mathcal{F}_2\}.$$
(Note that $A_1 \times A_2 = \{(\omega_1, \omega_2) \ : \ \omega_1 \in A_1, \omega_2 \in A_2\}$ is a subset of $\Omega_1 \times \Omega_2$.)

The problem is that the above set is NOT a σ-algebra. To see this, take sets $A_1, A_1' \in \mathcal{F}_1$ and $A_2, A_2' \in \mathcal{F}_2$ which are disjoint. Consider the product sets $A_1 \times A_2$ and $A_1' \times A_2'$. The union $(A_1 \times A_2) \cup (A_1' \times A_2')$ is not in $\mathcal{F}_1 \times \mathcal{F}_2$. This is the case, for example, as any element (ω_1, ω_2) with $\omega_1 \in A_1$ and $\omega_2 \in A_2'$ does not belong to the union.

1.0.5 Independence

Consider a countable set $\Omega = \{\omega_1, \omega_2, \ldots\}$. To each element ω_n we assign weight p_n such that

1. for all n, we have $p_n \geq 0$,
2. and $\sum_{n\geq 1} p_n = 1$.

We will use these weights in order to construct a probability measure on Ω. We will consider the σ-algebra $\mathcal{F}$ that consists of all subsets of Ω. For each subset $A \subseteq \Omega$, we set $\mathbb{P}(A) = \sum_{n\in A} p_n$. That is, the probability of A is just the sum of the weights of the elements of A. We will verify that $\mathbb{P}$ satisfies axioms P1-P3:

P1 for any $A \subseteq \Omega$ it is clear that $\mathbb{P}(A) \geq 0$ and $\mathbb{P}(A) = \sum_{n\in A} p_n \leq \sum_{n\geq 1} p_n = 1$.

P2 $\mathbb{P}(\emptyset) = 0$ and $\mathbb{P}(\Omega) = \sum_{n\geq 1} p_n = 1$.

P3 Consider a countable collection of pairwise disjoint subsets of Ω, denoted by $A_1, A_2, \ldots$. Then

$$\mathbb{P}(A_1 \cup A_2 \cup \cdots) = \sum_{n\in A_1\cup A_2\cdots} p_n = \sum_{j=1}^{\infty}\sum_{n\in A_j} p_j = \sum_{j=1}^{\infty}\mathbb{P}(A_j).$$

Example: spaces of sequences of unbiased trials

Consider the set of all possible sequences of outcomes of n coin tosses. We indicate the outcome of a single coin toss as *H*eads or *T*ails. With this convention, a sequence of outcomes of n coin tosses is a sequence of length n that consists of the letters H and T. So we can define the following spaces:

$$\begin{aligned} \text{For} \quad & n = 1 \quad \Omega_1 = \{H, T\} \\ & n = 2 \quad \Omega_2 = \{HH, HT, TH, TT\} \\ & n = 3 \quad \Omega_3 = \{HHH, HHT, HTH, HTT, THH, THT, TTH, TTT\} \\ & \quad \vdots \end{aligned}$$

If we assign the same weight $1/2^n$ to each element of Ω_n, then we say that the coin is *fair*.

Consider the following events on Ω_3. Let A_1 be the event that the first trial is H and let A_2 be the event that the second trial is T. Then

$$\mathbb{P}(A_1) = \mathbb{P}(\{HHH\}) + \mathbb{P}(\{HHT\}) + \mathbb{P}(\{HTH\}) + \mathbb{P}(\{HTT\}) = \frac{4}{2^3} = \frac{1}{2},$$

and

$$\mathbb{P}(A_2) = \mathbb{P}(\{HTH\}) + \mathbb{P}(\{HTT\}) + \mathbb{P}(\{TTH\}) + \mathbb{P}(\{TTT\}) = \frac{4}{2^3} = \frac{1}{2}.$$

Furthermore,

$$\mathbb{P}(A_1 \cap A_2) = \mathbb{P}(\{HTH\}) + \mathbb{P}(\{HTT\}) = \frac{2}{2^3} = \frac{1}{4} = \frac{1}{2}\cdot\frac{1}{2} = \mathbb{P}(A_1)\cdot\mathbb{P}(A_2).$$

Definition. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and let $A_1, A_2 \in \mathcal{F}$. We say that A_1 and A_2 are *independent* if $\mathbb{P}(A_1 \cap A_2) = \mathbb{P}(A_1)\cdot\mathbb{P}(A_2)$.

The above definition can be extended to more than two sets but in a more *complicated* way.

Definition. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and let $A_1, \ldots, A_n \in \mathcal{F}$ be a family of measurable sets. We say that the events $A_1, \ldots, A_n$ are *independent* if $\mathbb{P}(A_{i_1} \cap \cdots \cap A_{i_k}) = \mathbb{P}(A_{i_1}) \cdot \cdots \cdot \mathbb{P}(A_{i_k})$, for *any* subset of indices $\{i_1, \ldots, i_k\}$ of the set $\{1, \ldots, n\}$.

Example. Consider the space Ω_n (corresponding to n successive coin tosses). Let $\mathcal{F}_n$ be the σ-algebra on Ω_n which is the powerset of Ω_n, that is, consists of all subsets of Ω_n. The events $A_1, \ldots, A_n$, where A_i is the event that the i element of a sequence in Ω_n is Heads, are independent.

Indeed, let $i_1, \ldots, i_k$ be a subset of indices in $\{1, \ldots, n\}$ and let us consider the events $A_{i_1}, \ldots, A_{i_k}$. We need to show that

$$\mathbb{P}(A_{i_1} \cap \cdots \cap A_{i_k}) = \mathbb{P}(A_{i_1}) \cdot \cdots \cdot \mathbb{P}(A_{i_k}).$$

A generalisation of the argument we presented above shows that $\mathbb{P}(A_{i_1}) = \cdots = \mathbb{P}(A_{i_k}) = 1/2$.

Now, we calculate $\mathbb{P}(A_{i_1} \cap \cdots \cap A_{i_k})$. This is the event (that is, the subset of Ω_n) which consists of those sequences of H and T which have H is positions $i_1, \ldots, i_k$. But there are 2^{n-k} such sequences each having probability equal to $1/2^n$. Hence

$$\mathbb{P}(A_{i_1} \cap \cdots \cap A_{i_k}) = 2^{n-k} \cdot \frac{1}{2^n} = \frac{1}{2^k}.$$

The space of infinite sequences of (unbiased) trials

Let Ω_∞ be the set of sequences $a_1 a_2 \ldots$, where $a_i \in \{H, T\}$, for all $i \in \mathbb{N}$. How is this related to Ω_n (recall that this is the set of all sequences $a_1 a_2 \ldots a_n$ with $a_i \in \{H, T\}$)?

Essentially, one can view the elements of Ω_n as the set of all length-n prefixes of all words in Ω_∞. Take for example the infinite sequence $HHHH \cdots$. The first n letters of it $\underbrace{H \cdots H}_{n} H \cdots$ form the prefix of the infinite sequence which has length n.

More generally, consider a word $w \in \Omega_n$. This corresponds to the subset of all words Ω_∞ which *start with* w. Let S_w denote this subset.

We take the σ-algebra in Ω_∞ which is the smallest σ-algebra which contains all the set S_w, for any $w \in \Omega_n$ and for any $n \in \mathbb{N}$. We denote this σ-algebra by $\mathcal{F}_\infty$.

What is the associated probability measure $\mathbb{P}_\infty$? Since the trials are meant to be unbiased, if we take two different words $w, w' \in \Omega_n$, we would like the sets S_w and $S_{w'}$ to have the same probability weight. In other words, all 2^n possible such subsets must have the same probability which is $1/2^n$. Thus, for any $n \in \mathbb{N}$ and any $w \in \Omega_n$, we should have $\mathbb{P}_\infty(S_w) = 1/2^n$.

Such a probability measure does exist!! The proof of its existence is non-trivial and relies on a general result which is the Daniel-Kolmogorov theorem; this is beyond the scope of this module.

Example: spaces of sequences of biased trials

In this case, we consider Ω_n but we give different weights on the words in Ω_n. In particular, let p be a real number such that $0 \leq p \leq 1$. Now, we do not assign each probability weight

to each sequence. Instead of this, for a word $a_1 a_2 \dots a_n$ we assign the product $p_1 p_2 \cdots p_n$, where $p_i = p$ if $a_i = H$ and $p_i = 1 - p$ if $a_i = T$. Again, the σ-algebra of events consists of all subsets of Ω_n; that is, it is $\mathcal{F}_n$.

Now, consider again the event A_1, namely that the first letter/trial is H. Note that $A_1 \subset \Omega_n$ consists of all words in Ω_n which start with H (however, they have not equal weight...). Any word $H a_2 \dots a_n \in A_1$ has weight which is the product $p p_2 \cdots p_n$. So we have

$$\mathbb{P}(A_1) = p \cdot \sum_{a_2,\dots,a_n \in \{H,T\}} \prod_{i=2}^{n} p_i = p \cdot \sum_{k=0}^{n-1} \binom{n-1}{k} p^k (1-p)^{n-k} = p.$$

This can be extended to the space Ω_∞ with the σ-algebra $\mathcal{F}_\infty$, as above. However, the probability measure $\mathbb{P}_\infty$ now has changed. For every $w = a_1 \dots a_n \in \Omega_n$, the set S_w has $\mathbb{P}_\infty(S_w) = p_1 \cdots p_n$, where $p_i = p$ if $a_i = H$ and $p_i = 1 - p$, otherwise.

Chapter 2

Random variables

2.1 Random variables and their distribution function

A random variable is a function on a probability space but with some restrictions/conditions. However, these conditions are somewhat subtle so that usually most functions one deals with are indeed random variable. Now, let us be more specific about these.

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. A random variable $X : \Omega \to \mathbb{R}$ is a real-valued function on Ω with the property that is *measurable*. This term is defined as follows: for any open set $S \subset \mathbb{R}$ the set $X^{-1}(S) = \{\omega \in \Omega \ : \ X(\omega) \in S\}$ is in $\mathcal{F}$, that is, it is a measurable set.

Remark: note that we use the term *measurable* both for sets and functions.

Example (A fair die)**.** Let $\Omega = \{1, \ldots, n\}$ with $\mathcal{F}$ being the entire powerset of Ω (all subsets of Ω) and $\mathbb{P}$ be the counting measure. Let $X : \Omega \to \mathbb{R}$ defined as $X(i) = i$, for all $i \in \Omega$. One can think of this particular Ω as the space of all possible outcomes of a die with n faces and X the random variable that is the outcome of the die. But is it a random variable? We need to check that the above definition holds.

Let S be an open subset of $\mathbb{R}$. Then $X^{-1}(S) = \{i \in \{1, \ldots, n\} \ : \ X(i) = i \in S$. But this is just a subset of Ω; it may be empty or even the entire Ω. But in any case it belongs to $\mathcal{F}$.

And now a counterexample...

Example. For $n > 2$, let $\Omega = \{1, \ldots, n\}$ again but now let $\mathcal{F} = \{\emptyset, \{1\}, \{2, \ldots, n\}, \Omega\}$. In this case X as defined above is a function that is not a random variable. This is the case, for example, because if we take $S = (n-2, n)$, then $X^{-1}(S) = \{n-1\}$. But this is not a member of $\mathcal{F}$. In this case, the σ-algebra $\mathcal{F}$ is not rich enough to turn the function X into a random variable.

Why do we need this? The informal answer is that we need $\mathcal{F}$ to be rich enough so that it captures/includes the variation of X over Ω. In particular, we are interested in the following sets: $X^{-1}((-\infty, x])$. Is this set in $\mathcal{F}$? We can say that the set $X^{-1}((x, \infty))$ is in $\mathcal{F}$, if X is a random variable. But by Property A2, its complement is in $\mathcal{F}$ too. Note that its complement is $X^{-1}((-\infty, x])$. So this is in $\mathcal{F}$.

To simplify our notation, let us set $\{X \leq x\} := X^{-1}((-\infty, x])$. Similarly, we set $\{X < x\} := X^{-1}((-\infty, x))$ and $\{X > x\} := X^{-1}((x, \infty])$ and $\{X \geq x\} := X^{-1}([x, \infty))$. Other combinations are defined analogously.

2.1.1 The distribution function of a random variable X

The function $F_X(x) = \mathbb{P}(X^{-1}((-\infty, x]))$, for any $x \in \mathbb{R}$, is called the *cumulative distribution function* of X. This is simply the probability that X is at most x. Sometimes we abbreviate this as cdf.

Let us remark that F_X is non-decreasing, non-negative and at most 1. In general, it may not be continuous, that is, it may contain jumps. However, one can show that it may contain at most countably many jumps/discontinuities (we will not show this in this course).

Now, consider two real numbers a, b with $a < b$. Note that $\{X \leq a\} \subseteq \{X \leq b\}$ and let us set

$$\{a < X \leq b\} = \{X \leq b\} \setminus \{X \leq a\}.$$

By Lemma 1.1 Part 2, we can deduce that $\{a < X \leq b\}$ is a measurable set as well. Then

$$\mathbb{P}(\{a < X \leq b\}) = \mathbb{P}(\{X \leq b\}) - \mathbb{P}(\{X \leq a\}) = F_X(b) - F_X(a).$$

This holds as for any two measurable sets $A, B \in \mathcal{F}$ such that $A \subseteq B$, we have $\mathbb{P}(A \setminus B) = \mathbb{P}(A) - \mathbb{P}(B)$, as $A = B \cup (A \setminus B)$ and the latter are disjoint (so we can apply (1.1)).

2.1.2 Common distribution functions

Here is collection of common distribution functions

1. The binomial distribution: a random variable X is binomially distributed with parameter p, if for any $0 \leq k \leq n$

$$\mathbb{P}(X = k) = \binom{n}{k} p^k (1-p)^{n-k},$$

   and 0 otherwise.

2. The geometric distribution: a random variable X is geometrically distributed with parameter p, if for any $k \in \mathbb{N}$

$$\mathbb{P}(X = k) = (1-p)^{k-1} p,$$

   and 0 otherwise.

3. The Bernoulli distribution: a random variable X follows the Bernoulli distribution with parameter $p \in [0, 1]$, if it takes only two values, namely 1 and 0, with probability p and $1 - p$, respectively.

4. A random variable X follows the Gaussian distribution $N(\mu, \sigma)$, if

$$F_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{x} e^{-(z-\mu)^2/\sigma^2} dz.$$

2.1.3 The expected value of a random variable

One may wish to define the expected value of a random variable X as $\mathbb{E}(X) = \sum_{\omega \in \Omega} X(\omega) \cdot \mathbb{P}(\omega)$. This is not formally correct! The value of $X(\omega)$ is well-defined, but what about $\mathbb{P}(\omega)$? Firstly it may be the case that $\{\omega\}$ is not in $\mathcal{F}$ and in this case $\mathbb{P}(\omega)$ is not defined at all (remember that $\mathbb{P}$ is only defined on measurable sets). It is often the case that $\mathbb{P}(\omega) = 0$ for all $\omega \in \Omega$. So in that case the above sum is equal to 0. So we need another idea.

The density of a distribution

A random variable X is called *continuous* if its cumulative distribution function F_X can be written as:

$$F_X(x) = \int_{-\infty}^{x} f_X(y)dy,$$

for some function $f_X : \mathcal{R} \to [0, \infty)$ which is *integrable*. This is called the *probability density function* of X (or simply the *density function*). Of course, this means that if F_X is differentiable, then we can take $f_X = F_X'$.

Let us consider the following example.

Example (Example 2.3.4 from [2])**.** Let $\Omega = [0, 2\pi)$ and let $\mathcal{F}$ be the Borel σ-algebra on Ω. Recall that this is the smallest σ-algebra, which contains all open intervals (a, b) of $\Omega = [0, 2\pi)$. Take a probability measure that assigns

$$\mathbb{P}((a, b)) = \frac{b - a}{2\pi}.$$

In other words, this is the *uniform* probability measure on Ω. You may think of this as the measure that is induced by the angle a straight rod has from the true north if it is flung down at random.

Consider the following two random random variables: for any $\omega \in \Omega$ we set

$$X(\omega) = \omega \text{ and } Y(\omega) = \omega^2.$$

Let us determine their cdfs.

We start with the cdf of X. For any $x \leq 0$, we have $\mathbb{P}(X \leq 0) = 0$, whereby $F_X(x) = 0$. If $0 < x < 2\pi$, then

$$\mathbb{P}(X \leq x) = \mathbb{P}(X \leq x) - 0 = \mathbb{P}(X \leq x) - \mathbb{P}(X \leq 0) = \mathbb{P}((0, x]) = \frac{x}{2\pi}.$$

Finally, if $x \geq 2\pi$, then $\mathbb{P}(X \leq x) = \mathbb{P}(X \leq 2\pi) = 1$. We thus conclude that

$$F_X(x) = \begin{cases} 0, & \text{if } x \leq 0 \\ \frac{x}{2\pi}, & \text{if } 0 < x < 2\pi \\ 1, & \text{if } x \geq 2\pi \end{cases}.$$

Let us turn to Y. Again, for any $x \leq 0$, we have $\mathbb{P}(Y \leq 0) = 0$, whereby $F_Y(x) = 0$. For $0 < x < (2\pi)^2$, we have

$$\mathbb{P}(Y \leq x) = \mathbb{P}(X \leq x^{1/2}) = \frac{x^{1/2}}{2\pi}.$$

Also, for $x \geq (2\pi)^2$ (which means $x^{1/2} \geq 2\pi$), we have

$$\mathbb{P}(Y \leq x) = \mathbb{P}(X \leq x^{1/2}) = 1.$$

To summarise:

$$F_Y(x) = \begin{cases} 0, & \text{if } x \leq 0 \\ \frac{x^{1/2}}{2\pi}, & \text{if } 0 < x < (2\pi)^2 \\ 1, & \text{if } x \geq (2\pi)^2 \end{cases}.$$

What are the probability density functions of X and Y? For the random variable X we can take

$$f_X(x) = \begin{cases} \frac{1}{2\pi}, & \text{if } 0 < x < 2\pi \\ 0, & \text{otherwise} \end{cases}.$$

For Y we can also take

$$f_Y(x) = \begin{cases} \frac{x^{-1/2}}{4\pi}, & \text{if } 0 < x < (2\pi)^2 \\ 0, & \text{otherwise} \end{cases}.$$

Remark. The above example shows that the probability density function is not always smaller than 1 and it is not always continuous.

Definition of the expected value

Let X be a random variable on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$. We will consider two case

The case of a discrete random variable

Assume that X takes only values in a countable set $S \subset \mathbb{R}$. Then we define

$$\mathbb{E}(X) = \sum_{s \in S} s \cdot \mathbb{P}(X = s).$$

Example. The geometric distribution Assume that X is *geometrically distributed* with parameter p. That is, for every $k \in \mathbb{N}$ we have $\mathbb{P}(X = k) = (1-p)^{k-1}p$, where $p \in (0, 1)$. Thus,

$$\mathbb{E}(X) = \sum_{k=1}^{\infty} k \cdot \mathbb{P}(X = k) = \sum_{k=1}^{\infty} k(1-p)^{k-1}p = p\sum_{k=1}^{\infty} k(1-p)^{k-1}.$$

But $\sum_{k=1}^{\infty} k(1-p)^{k-1} = \left(\sum_{k=0}^{\infty}(1-p)^k\right)' = \left(\frac{1}{p}\right)' = \frac{1}{p^2}$. Thus,

$$\mathbb{E}(X) = p \cdot \frac{1}{p^2} = \frac{1}{p}.$$

The case of a continuous random variable

If X has density f, then we define

$$\mathbb{E}(X) = \int_{-\infty}^{\infty} x \cdot f(x)dx.$$

Example. The normal distribution Let X be a random variable that follows the *normal distribution*. Recall that this has density

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

So we have

$$\mathbb{E}(X) = \frac{1}{\sqrt{2\pi}} \cdot \int_{-\infty}^{\infty} x e^{-x^2/2} dx.$$

But $xe^{-x^2/2} = -(-x)e^{(-x)^2/2}$. Thus the above integral is 0 as

$$\int_0^{\infty} x e^{-x^2/2} dx = -\int_{-\infty}^{0} x e^{-x^2/2} dx.$$

Fact 2.1. *Let X be a random variable and let $a, b \in \mathbb{R}$. We have*

$$\mathbb{E}(aX + b) = a\mathbb{E}(X) + b.$$

Proof. We will prove it only for the continuous case - the discrete case can be shown analogously. If X has density f, then

$$\mathbb{E}(aX + b) = \int_{-\infty}^{\infty} (ax + b)f(x)dx = a\int_{-\infty}^{\infty} xf(x)dx + b\int_{-\infty}^{\infty} f(x)dx = a\mathbb{E}(X) + b.$$

□

2.1.4 Functions of random variables

Let $g : \mathbb{R} \to \mathbb{R}$ be If X is a random variable, then $g(X)$ is a random variable: The event $\{g(X) \leq y\}$ is the same as the event $X \in g^{-1}(-\infty, y]$. Now, is $g^{-1}(-\infty, y]$ a Borel measurable set. We shall consider those functions for which this is the case. This will ensure that $g(X)$ is a random variable too! If X is continuous with probability density function f_X, then

$$\mathbb{P}(g(X) \leq y) = \mathbb{P}\left(X \in g^{-1}(-\infty, y]\right) = \int_{g^{-1}(-\infty,y]} g(x)f_X(x)dx.$$

Furthermore,

$$\mathbb{E}(g(X)) = \int_{-\infty}^{\infty} g(x)f_X(x).$$

2.2 The variance of a random variable

Let X be random variable on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ which has expected value $\mathbb{E}(X) = \mu < \infty$. The variance of $\mathrm{Var}(X)$ is defined as

$$\mathrm{Var}(X) = \mathbb{E}\left((X - \mu)^2\right).$$

If we expand the quadratic, then we obtain

$$\mathrm{Var}(X) = \mathbb{E}\left(X^2 - 2\mu X + \mu^2\right) = \mathbb{E}\left(X^2 - 2\mu X\right) + \mu^2$$

But $\mathbb{E}(X^2 - 2\mu X) = \int_{-\infty}^{\infty}(x^2 - 2\mu x)f(x)dx = \int_{-\infty}^{\infty} x^2 f(x)dx - 2\mu\int_{-\infty}^{\infty} xf(x)dx = \mathbb{E}(X^2) - 2\mu^2$. Hence,

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

Example. Assume that X is Bernoulli distributed with parameter $p \in [0, 1]$, that is, it takes only two values: $X = 1$ with probability p and $X = 0$ with probability $1 - p$. The expected value of X is

$$\mathbb{E}(X) = 1 \cdot \mathbb{P}(X = 1) + 0 \cdot \mathbb{P}(X = 0) = \mathbb{P}(X = 1) = p.$$

Now,

$$\mathrm{Var}(X) = \mathbb{E}\left((X - \mathbb{E}(X))^2\right) = \mathbb{E}\left(X^2\right) - \mathbb{E}(X)^2 = \mathbb{E}\left(X^2\right) - p^2.$$

But $\mathbb{E}(X^2) = 1^2 \cdot \mathbb{P}(X = 1) + 0^2 \cdot \mathbb{P}(X = 0) = \mathbb{P}(X = 1) = p$ too. We thus conclude that:

$$\mathrm{Var}(X) = p - p^2 = p(1 - p).$$

The variance of a random variable is not exactly linear: it squares multiplicative constants and absorbs additive constants.

Fact 2.2. *Let X be a random variable and let $a, b \in \mathbb{R}$. We have*

$$\mathrm{Var}\left(aX + b\right) = a^2 \mathrm{Var}\left(X\right).$$

Proof. We will prove it only for the continuous case - the discrete case can be shown analogously. If X has density f, then $\mathbb{E}\left(aX + b\right) = a\mathbb{E}\left(X\right) + b$.

$$\mathrm{Var}\left(aX + b\right) = \mathbb{E}\left((aX + b)^2\right) - \mathbb{E}\left(aX + b\right)^2 = \mathbb{E}\left((aX + b)^2\right) - (a\mathbb{E}\left(X\right) + b)^2.$$

Now, $\mathbb{E}\left((aX + b)^2\right) = \int_{-\infty}^{\infty}(a^2x^2+2abx+b^2)f(x)dx = a^2\int_{-\infty}^{\infty} x^2 f(x)dx+2ab\int_{-\infty}^{\infty} xf(x)dx+ b^2\int_{-\infty}^{\infty} f(x)dx = a^2\mathbb{E}\left(X^2\right)+2ab\mathbb{E}\left(X\right)+b^2$. Also, $(a\mathbb{E}\left(X\right)+b)^2 = a^2\mathbb{E}\left(X\right)^2+2ab\mathbb{E}\left(X\right)+b^2$. Thus,

$$\mathrm{Var}\left(aX + b\right) = a^2(\mathbb{E}\left(X^2\right) - \mathbb{E}\left(X\right)^2) = a^2\mathrm{Var}\left(X\right).$$

□

2.3 Markov and Chebyschev inequalities

Inuitively, one uses the expect value as a way to estimate the magnitude of a random variable. But how likely is it that a random variable is close to its expected value? The simplest way to quantify this is Markov's inequality.

Theorem 2.3 (Markov's inequality)**.** *Let X be random variable on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ which is non-negative and have bounded expected value. For any $t > 0$, we have*

$$\mathbb{P}\left(X \geq t\right) \leq \frac{\mathbb{E}\left(X\right)}{t}.$$

Proof. Assume that X is continuous with density function f_X and expected value $\mathbb{E}\left(X\right) < \infty$. Then we write

$$\begin{aligned}\mathbb{E}\left(X\right) &= \int_{-\infty}^{\infty} x f_X(x)dx \overset{X\geq 0}{=} \int_{0}^{\infty} x f_X(x)dx \geq \int_{t}^{\infty} x f_X(x)dx \\ &= t\int_{t}^{\infty} f_X(x)dx = t\cdot\mathbb{P}\left(X \geq t\right).\end{aligned}$$

The proof for discrete X is analogous - you just replace the integral by a sum. □

Note that Markov's inequality gives a bound on the upper tail of X, that is, the probability that X is t times its expected value. What about the other tail? We would also like a bound on the probability that $|X - \mathbb{E}\left(X\right)| \geq t$. This can be obtained through Markov's inequality.

Note first that $|X - \mathbb{E}\left(X\right)| \geq t$ if and only if $|X - \mathbb{E}\left(X\right)|^2 \geq t^2$. So we have

$$\mathbb{P}\left(|X - \mathbb{E}\left(X\right)| \geq t\right) = \mathbb{P}\left(|X - \mathbb{E}\left(X\right)|^2 \geq t^2\right) \leq \frac{\mathbb{E}\left(|X - \mathbb{E}\left(X\right)|^2\right)}{t^2}.$$

But recall that $\mathbb{E}\left(|X - \mathbb{E}\left(X\right)|^2\right) = \mathrm{Var}\left(X\right)$. Thus, we conclude that for any $t > 0$

$$\mathbb{P}\left(|X - \mathbb{E}\left(X\right)| \geq t\right) \leq \frac{\mathrm{Var}\left(X\right)}{t^2}. \tag{2.1}$$

This is known as *Chebyschev's inequality*.

Of course, one may repeat this argument using higher power than 2. In this case, one would deduce that for any $k \geq 2$

$$\mathbb{P}\left(|X - \mathbb{E}\left(X\right)| \geq t\right) \leq \frac{\mathbb{E}\left(|X - \mathbb{E}\left(X\right)|^k\right)}{t^k}. \tag{2.2}$$

Note that for this to make sense, we must have $\mathbb{E}\left(|X - \mathbb{E}\left(X\right)|^k\right) < \infty$.

2.4 Joint distributions - independence of random variables

Let X, Y be two random variables defined on the same probability space $(\Omega, \mathcal{F}, \mathbb{P})$. Their *joint cumulative distribution function* $F_{X,Y} : \mathbb{R}^2 \to [0,1]$ is such that

$$F_{X,Y}(x,y) = \mathbb{P}\left(X \leq x, Y \leq y\right).$$

We say that X, Y have a joint density $f_{X,Y} : \mathbb{R}^2 \to [0,\infty]$ if for any $x, y \in \mathbb{R}$

$$F_{X,Y}(x,y) = \int_{-\infty}^{x}\int_{-\infty}^{y} f_{X,Y}(z_1,z_2)dz_2dz_1.$$

Of course, this implies that

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_{X,Y}(z_1,z_2)dz_2dz_1 = 1. \tag{2.3}$$

Can we work out the density function of X? Note that

$$\mathbb{P}\left(X \leq x\right) = \mathbb{P}\left(X \leq x, Y < \infty\right) = \int_{-\infty}^{x}\int_{-\infty}^{\infty} f_{X,Y}(z_1,z_2)dz_2dz_1.$$

Thus,

$$F_X(x) = \int_{-\infty}^{x}\left(\int_{-\infty}^{\infty} f_{X,Y}(z_1,z_2)dz_2\right)dz_1,$$

which implies that $f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x,z_2)dz_2$ is the probability density function of X. Similarly, the probability density function of Y is $f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(z_1,y)dz_1$.

More generally, one can extend these definitions to larger collections of random variables. If we consider the random variables $X_1, \ldots, X_k$ all defined on $(\Omega, \mathcal{F}, \mathbb{P})$, their joint cumulative distribution function $F_{X_1,\ldots,X_k} : \mathbb{R}^k \to [0,1]$ is such that for any $x_1, \ldots, x_k \in \mathbb{R}$

$$F_{X_1,\ldots,X_k}(x_1,\ldots,x_k) = \mathbb{P}\left(X_1 \leq x_1, \ldots, X_k \leq x_k\right).$$

Similarly, their joint density function $f_{X_1,\ldots,X_k} : \mathbb{R}^k \to [0,\infty)$ is such that

$$F_{X_1,\ldots,X_k}(x_1,\ldots,x_k) = \int_{-\infty}^{x_1}\cdots\int_{-\infty}^{x_k} f_{X_1,\ldots,X_k}(z_1,\ldots,z_k)dz_k\ldots dz_1.$$

A very useful fact about collections of random variables is the *linearity of the expected value*.

Fact 2.4. *Let X, Y be two random variables on $(\Omega, \mathcal{F}, \mathbb{P})$. Then*

$$\mathbb{E}(X+Y) = \mathbb{E}(X) + \mathbb{E}(Y).$$

Proof. We have

$$\begin{aligned}\mathbb{E}(X+Y) &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(x+y)f_{X,Y}(x,y)dydx \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}xf_{X,Y}(x,y)dydx + \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}yf_{X,Y}(x,y)dydx.\end{aligned}$$

Now, we write

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}xf_{X,Y}(x,y)dydx = \int_{-\infty}^{\infty}x\left(\int_{-\infty}^{\infty}f_{X,Y}(x,y)dy\right)dx.$$

But the *inner* integral is the density function of X. Thus,

$$\int_{-\infty}^{\infty}x\left(\int_{-\infty}^{\infty}f_{X,Y}(x,y)dy\right)dx = \int_{-\infty}^{\infty}xf_X(x)dx = \mathbb{E}(X).$$

Similarly,

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}yf_{X,Y}(x,y)dydx = \int_{-\infty}^{\infty}y\left(\int_{-\infty}^{\infty}f_{X,Y}(x,y)dx\right)dy = \mathbb{E}(Y).$$

Our claim thus follows. □

An inductive argument can extend this to a collection of more than two random variables.

Fact 2.5. *Let $X_1, \ldots, X_k$ be a collection of random variables on $(\Omega, \mathcal{F}, \mathbb{P})$. We have*

$$\mathbb{E}(X_1 + \cdots + X_k) = \mathbb{E}(X_1) + \cdots + \mathbb{E}(X_k).$$

Remark. Note that the above two identities hold no matter how the random variables are correlated.

The linearity of the expected value is a very useful in that it greatly simplifies its calculation in several cases.

Example. Consider the binomial distribution - a random variable X is binomially distributed with parameters n and p, if for any $0 \leq k \leq n$ we have $\mathbb{P}(X=k) = \binom{n}{k}p^k(1-p)^{n-k}$. Let us calculate its expected value.
1st method We calculate explicitly the expected value of X using the definition:

$$\begin{aligned}\mathbb{E}(X) &= \sum_{k=0}^{n} k\cdot\binom{n}{k}p^k(1-p)^{n-k} = \sum_{k=1}^{n}\binom{n}{k}k\cdot p\cdot p^{k-1}(1-p)^{n-k} \\ &= \left(\sum_{k=0}^{n}\binom{n}{k}(xp)^k(1-p)^{n-k}\right)'_{x=1}.\end{aligned}$$

But by the binomial theorem, we have

$$\sum_{k=0}^{n} \binom{n}{k} (xp)^k (1-p)^{n-k} = (xp + 1 - p)^n.$$

The derivative of the latter with respect to x is $np(xp + 1 - p)^{n-1}$ - at $x = 1$, this evaluates to np. Hence,

$$\mathbb{E}(X) = np.$$

This the standard algebraic way.

2nd method

Recall that X actually has meaning. Recall that we defined the space Ω_n which consists of all words of length n on letters $\{H, T\}$ and to each word $a_1 \cdots a_n \in \Omega_n$ we assigned probability equal to $\prod_{i=1}^{n} f(a_i)$, with $f(a_i) = p$, if $a_i = H$ but $f(a_i) = 1 - p$, if $a_i = T$. In the lectures, we proved that the probability weight of the subset of words with exactly k Hs is $\binom{n}{k} p^k (1-p)^{n-k}$. In other words, the number of Hs in a word from this probability space is a random variable that is binomially distributed with parameters n and p. Let us call this random variable X.

We can express X as a sum of *indicator random variables*, that is, random variables that take the value 0 or 1 - in other words, Bernoulli distributed random variables. More specifically, let X_i be the indicator random variable on Ω_n which is 1 if and only if the ith letter of the word is H. Hence, for any $w \in \Omega_n$ the sum $X_1(w) + \cdots + X_n(w)$ is equal to the number of Hs in w.

Therefore, we can write:

$$X = X_1 + \cdots + X_n,$$

and this yields:

$$\mathbb{E}(X) = \mathbb{E}(X_1) + \cdots + \mathbb{E}(X_n). \tag{2.4}$$

So it suffices to find $\mathbb{E}(X_1), \ldots, \mathbb{E}(X_n)$.

We have seen in the example of Section 2.2 that for a Bernoulli distributed random variable with parameter p, its expect value is equal to p. Recall that p is the probability that the random variable is equal to 1. Thus, we need to compute the probability that $X_i = 1$. Let us focus on X_1 for a moment. What is the probability that $X_1 = 1$? In other words: what is the probability that the first letter of a word in Ω_n is H? Using the probability weights we assigned to each word this is

$$\begin{aligned}\mathbb{P}(X_1 = 1) &= \sum_{a_1 \cdots a_n \in \Omega_n, a_1 = H} \mathbb{P}(a_1 \cdots a_n) = p \cdot \sum_{a_2 \cdots a_n \in \Omega_{n-1}} \mathbb{P}(a_2 \cdots a_n) \\ &= p \cdot \sum_{k=0}^{n-1} \binom{n-1}{k} p^k (1-p)^{n-1-k} = p.\end{aligned}$$

Now, by symmetry we can deduce that $\mathbb{P}(X_i = 1) = p$. Therefore,

$$\mathbb{E}(X_i) = p, \text{ for all } i = 1, \ldots, n.$$

and by (2.4) $\mathbb{E}(X) = np$.

An application: the union bound

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and let $A_1, \ldots, A_n \in \mathcal{F}$ be n events on it. We are going to give a simple upper bound on $\mathbb{P}\left(\cup_{i=1}^n A_i\right)$. For every, $i = 1, \ldots, n$, we define an indicator random variable X_i which is equal to 1 only on A_i. That is, $X_i(\omega) = 1$ precisely when $\omega \in A_i$; otherwise it is 0. Note that the event $\{\cup_{i=1}^n A_i\}$ is realised, if and only if at least one of the events A_i is realised. In other words, the event $\{\cup_{i=1}^n A_i\}$ coincides with the event $X_1 + \cdots + X_n \geq 1$. Thus, using Markov's inequality:

$$\mathbb{P}\left(\cup_{i=1}^n A_i\right) = \mathbb{P}\left(X_1 + \cdots + X_n \geq 1\right) \leq \frac{\mathbb{E}\left(X_1 + \cdots + X_n\right)}{1}.$$

But by the linearity of the expected value:

$$\mathbb{E}\left(X_1 + \cdots + X_n\right) = \mathbb{E}\left(X_1\right) + \cdots + \mathbb{E}\left(X_n\right),$$

and note that $\mathbb{E}\left(X_i\right) = 1 \cdot \mathbb{P}\left(X_i = 1\right) = \mathbb{P}\left(A_i\right)$. Thus, we conclude that

$$\mathbb{P}\left(\cup_{i=1}^n A_i\right) \leq \sum_{i=1}^n \mathbb{P}\left(A_i\right). \tag{2.5}$$

2.5 Covariances - the variance of a sum of random variables

It is often needed in applications to be able to bound the variance of the sum of random variables.

Let X, Y be two random variables on the space $(\Omega, \mathcal{F}, \mathbb{P})$ having expected values μ_X and μ_Y, respectively. By the linearity of the expected value, $\mathbb{E}\left(X + Y\right) = \mu_X + \mu_Y$. Let us calculate the *variance* of $X + Y$. We have

$$\begin{aligned}
\mathrm{Var}\left(X + Y\right) &= \mathbb{E}\left((X + Y - (\mu_X + \mu_Y))^2\right) \\
&= \mathbb{E}\left((X + Y)^2 - 2(X + Y)(\mu_X + \mu_Y) + (\mu_X + \mu_Y)^2\right) \\
&= \mathbb{E}\left((X + Y)^2\right) - 2(\mu_X + \mu_Y)\mathbb{E}\left(X + Y\right) + (\mu_X + \mu_Y)^2 \\
&= \mathbb{E}\left((X + Y)^2\right) - 2(\mu_X + \mu_Y)^2 + (\mu_X + \mu_Y)^2 \\
&= \mathbb{E}\left((X + Y)^2\right) - (\mu_X + \mu_Y)^2.
\end{aligned}$$

Now, expanding the squares, we have

$$\mathbb{E}\left((X + Y)^2\right) = \mathbb{E}\left(X^2 + 2XY + Y^2\right) = \mathbb{E}\left(X^2\right) + 2\mathbb{E}\left(XY\right) + \mathbb{E}\left(Y^2\right),$$

and

$$(\mu_X + \mu_Y)^2 = \mu_X^2 + 2\mu_X\mu_Y + \mu_Y^2.$$

So,

$$\begin{aligned}
\mathbb{E}\left((X + Y)^2\right) - (\mu_X + \mu_Y)^2 &= \left(\mathbb{E}\left(X^2\right) - \mu_X^2\right) + 2(\mathbb{E}\left(XY\right) - \mu_X\mu_Y) + (\mathbb{E}\left(Y^2\right) - \mu_Y^2) \\
&= \mathrm{Var}\left(X\right) + 2(\mathbb{E}\left(XY\right) - \mu_X\mu_Y) + \mathrm{Var}\left(Y\right).
\end{aligned}$$

Now, we define the quantity

$$\mathrm{Cov}\left(X, Y\right) := \mathbb{E}\left(XY\right) - \mu_X\mu_Y,$$

which is called the *covariance* of the random variables X and Y. Thus, we have proved that

$$\mathrm{Var}\left(X+Y\right)=\mathrm{Var}\left(X\right)+\mathrm{Var}\left(Y\right)+2\mathrm{Cov}\left(X,Y\right).$$

Note also, that $\mathrm{Cov}\left(X,X\right)=\mathrm{Var}\left(X\right)$.

Now, if we take three random variables X_1, X_2, Y on the same probability space, then we have

$$\begin{aligned}\mathrm{Cov}\left(X_1+X_2,Y\right)&=\mathbb{E}\left(X_1Y+X_2Y\right)-(\mu_{X_1}+\mu_{X_2})\mu_Y=\mathbb{E}\left(X_1Y+X_2Y\right)-(\mu_{X_1}+\mu_{X_2})\mu_Y\\&=\mathbb{E}\left(X_1Y\right)-\mu_{X_1}\mu_Y+\mathbb{E}\left(X_2Y\right)-\mu_{X_2}\mu_Y\\&=\mathrm{Cov}\left(X_1,Y\right)+\mathrm{Cov}\left(X_2,Y\right).\end{aligned}$$

Also, $\mathrm{Cov}\left(X,Y\right)=\mathrm{Cov}\left(Y,X\right)$. Furthermore,

$$\begin{aligned}\mathrm{Cov}\left(aX_1,X_2\right)&=\mathbb{E}\left(aX_1X_2\right)-\mathbb{E}\left(aX_1\right)\mathbb{E}\left(X_2\right)=a\left(\mathbb{E}\left(X_1X_2\right)-\mathbb{E}\left(X_1\right)\mathbb{E}\left(X_2\right)\right)\\&=a\mathrm{Cov}\left(X_1,X_2\right).\end{aligned}$$

Using these properties, and an inductive argument, one can show that for a collection of random variables $X_1,\ldots,X_n$ on the same probability space, we have

$$\mathrm{Var}\left(X_1+\cdots+X_n\right)=\sum_{i=1}^{n}\mathrm{Var}\left(X_i\right)+2\sum_{i<j}\mathrm{Cov}\left(X_i,X_j\right). \tag{2.6}$$

The covariance of two random variables is an important parameter which indicates how two random variables are correlated. It is not always non-negative. However, the following holds.

Lemma 2.6. *For any random variables X,Y on the same probability space with standard deviations $\sigma_X>0$ and $\sigma_Y>0$, respectively, we have*

$$\left|\frac{\mathrm{Cov}\left(X,Y\right)}{\sigma_X\sigma_Y}\right|\leq 1.$$

To this end we will use the *Cauchy-Schwarz inequality*. This states that if X,Y are two random variables on $(\Omega,\mathcal{F},\mathbb{P})$ such that $\mathbb{E}\left(X^2\right),\mathbb{E}\left(Y^2\right)<\infty$, then

$$\left|\mathbb{E}\left(XY\right)\right|\leq\left(\mathbb{E}\left(X^2\right)\right)^{1/2}\left(\mathbb{E}\left(Y^2\right)\right)^{1/2} \tag{2.7}$$

Proof of the Cauchy-Schwarz inequality. Consider the random variable $Z=aX-bY$, where a,b are some real numbers. Then $Z^2\geq 0$ and therefore by the linearity of expectation

$$0\leq\mathbb{E}\left(Z^2\right)=a^2\mathbb{E}\left(X^2\right)-2ab\mathbb{E}\left(XY\right)+b^2\mathbb{E}\left(Y^2\right)=:p(a).$$

(Here, we view this expression as a second-degree polynomial with respect to a.) Since $p(a)$ is non-negative, it follows that it its minimum with respect to a is non-negative. But this is equal to $a_0=b\mathbb{E}\left(XY\right)/\mathbb{E}\left(X^2\right)$. Substituting a_0 for a, we obtain:

$$\begin{aligned}p(a_0)&=a_0^2\mathbb{E}\left(X^2\right)-2a_0b\mathbb{E}\left(XY\right)+b^2\mathbb{E}\left(Y^2\right)\\&=\frac{b^2\mathbb{E}\left(XY\right)^2}{\mathbb{E}\left(X^2\right)^2}\mathbb{E}\left(X^2\right)-2\frac{b^2\mathbb{E}\left(XY\right)^2}{\mathbb{E}\left(X^2\right)}+b^2\mathbb{E}\left(Y^2\right)\\&=-\frac{b^2\mathbb{E}\left(XY\right)^2}{\mathbb{E}\left(X^2\right)}+b^2\mathbb{E}\left(Y^2\right).\end{aligned}$$

As we discussed above, $p(a_0)$ is non-negative, and since we can select $b \not 0$, we deduce that.

$$\mathbb{E}\left(XY\right)^2 \leq \mathbb{E}\left(X^2\right)\mathbb{E}\left(Y^2\right).$$

Taking square roots on both sides, we deduce (2.7). □

A consequence of the Cauchy-Schwarz inequality is that the variance is always non-negative. Taking $Y = 1$ we deduce $\mathbb{E}\left(X\right)^2 = \mathbb{E}\left(X \cdot 1\right)^2 \leq \mathbb{E}\left(X^2\right)\mathbb{E}\left(1\right)^2 = \mathbb{E}\left(X^2\right)$, whereby

$$\mathrm{Var}\left(X\right) = \mathbb{E}\left(X\right)^2 - \mathbb{E}\left(X^2\right) \geq 0.$$

Now, we turn to the proof of Lemma (2.6).

Proof of Lemma 2.6. Set $Z = X - \frac{\mathrm{Cov}(X,Y)}{\sigma_Y^2}Y$. Then

$$\begin{aligned} 0 \leq \mathrm{Var}\left(Z\right) = \mathrm{Cov}\left(Z,Z\right) &= \mathrm{Cov}\left(X - \frac{\mathrm{Cov}\left(X,Y\right)}{\sigma_Y^2}Y, X - \frac{\mathrm{Cov}\left(X,Y\right)}{\sigma_Y^2}Y\right) \\ &= \mathrm{Cov}\left(X,X\right) + \left(\frac{\mathrm{Cov}\left(X,Y\right)}{\sigma_Y^2}\right)^2 \mathrm{Cov}\left(Y,Y\right) - 2\frac{\mathrm{Cov}\left(X,Y\right)}{\sigma_Y^2}\mathrm{Cov}\left(X,Y\right), \end{aligned}$$

by the linearity of covariances. But as $\mathrm{Cov}\left(X,X\right) = \sigma_X^2$ and $\mathrm{Cov}\left(Y,Y\right) = \sigma_Y^2$, we deduce that

$$\begin{aligned} 0 &\leq \sigma_X^2 + \left(\frac{\mathrm{Cov}\left(X,Y\right)}{\sigma_Y^2}\right)^2 \sigma_Y^2 - 2\frac{\mathrm{Cov}\left(X,Y\right)}{\sigma_Y^2}\mathrm{Cov}\left(X,Y\right) \\ &= \sigma_X^2 + \frac{\mathrm{Cov}\left(X,Y\right)^2}{\sigma_Y^2} - 2\frac{\mathrm{Cov}\left(X,Y\right)^2}{\sigma_Y^2} \\ &= \sigma_X^2 - \frac{\mathrm{Cov}\left(X,Y\right)^2}{\sigma_Y^2}. \end{aligned}$$

Rearranging, we get

$$\mathrm{Cov}\left(X,Y\right)^2 \leq \sigma_X^2\sigma_Y^2.$$

□

The ratio $\frac{\mathrm{Cov}(X,Y)}{\sigma_X\sigma_Y}$ is called the *correlation coefficient* of X and Y.

2.6 Independent random variables

Let $X_1, \ldots, X_n$ be n random variables defined on a common probability space $(\Omega, \mathcal{F}, \mathcal{P})$. We say that they are *independent* if for any real numbers $x_1, \ldots, x_n$, the events $\{X_1 \leq x_1\}, \{X_2 \leq x_2\}, \ldots, \{X_n \leq x_n\}$ are independent. This is implies that for any subset of indices $i_1, \ldots, i_\ell$ we have

$$\mathbb{P}\left(X_1 \leq x_1, \ldots, X_n \leq x_n\right) = \mathbb{P}\left(X_1 \leq x_1\right) \cdots \mathbb{P}\left(X_n \leq x_n\right),$$

or equivalently

$$F_{X_1,\ldots,X_n}(x_1, \ldots, x_n) = F_{X_1}(x_1) \cdots F_{X_n}(x_n).$$

In particular, if $X_1, \ldots, X_n$ are continuous random variables, this implies that the following holds for the probability density functions:

$$f_{X_1,\ldots,X_n}(x_1, \ldots, x_n) = f_{X_1}(x_1) \cdots f_{X_n}(x_n).$$

Independence of two random variables can be witnessed by their covariance. In particular, if X_1, X_2 are independent, then $\mathrm{Cov}\,(X_1, X_2) = 0$. To see this,

$$\begin{aligned}\mathbb{E}\,(X_1X_2) &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x_1x_2 f_{X_1,X_2}(x_1, x_2)d_{x_2}d_{x_1} = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x_1x_2 f_{X_1}(x_1)f_{X_2}(x_2)d_{x_2}d_{x_1} \\ &= \left(\int_{-\infty}^{\infty} x_1 f_{X_1}(x_1)dx_1\right) \cdot \left(\int_{-\infty}^{\infty} x_2 f_{X_2}(x_2)dx_2\right) = \mathbb{E}\,(X_1)\,\mathbb{E}\,(X_2)\,.\end{aligned}$$

Hence, $\mathrm{Cov}\,(X_1, X_2) = \mathbb{E}\,(X_1X_2) - \mathbb{E}\,(X_1)\,\mathbb{E}\,(X_2) = 0$.

Remark. Note: the opposite is NOT TRUE. An example of this is to assume that X_1, X_2 are two independent Bernoulli trials with parameter (or probability of success) equal to 1/2. Then the random variables $X_1 + X_2$ and $|X_1 - X_2|$ have covariance equal to 0 but they are not independent. To see the latter, observe that if $X_1 + X_2 = 2$, then this means that $X_1 = X_2 = 1$ and therefore $|X_1 - X_2| = 0$. To see the former, let us write first:

$$\mathrm{Cov}\,(X_1 + X_2, |X_1 - X_2|) = \mathrm{Cov}\,(X_1, |X_1 - X_2|) + \mathrm{Cov}\,(X_2, |X_1 - X_2|)\,.$$

Now,

$$\mathrm{Cov}\,(X_1, |X_1 - X_2|) = \mathbb{E}\,(X_1|X_1 - X_2|) - \mathbb{E}\,(X_1)\,\mathbb{E}\,(|X_1 - X_2|)\,.$$

The product $X_1|X_1 - X_2|$ is positive (actually it is 1) precisely when $X_1 = 1$ but $X_2 = 0$. This occurs with probability $1/4$. Therefore, $\mathbb{E}\,(X_1|X_1 - X_2|) = 1/2$. Also, $\mathbb{E}\,(X_1) = 1/2$. Now, $|X_1 - X_2| = 1$ if and only if $X_1 = 1$ and $X_2 = 0$ or $X_1 = 0$ and $X_2 = 1$; otherwise it is 0. That event occurs with probability $2 \cdot \frac{1}{4} = \frac{1}{2}$, whereby $\mathbb{E}\,(|X_1 - X_2|) = 1/2$. Hence, $\mathrm{Cov}\,(X_1, |X_1 - X_2|) = \frac{1}{4} - \frac{1}{4} = 0$. One can similarly, show that $\mathrm{Cov}\,(X_1, |X_1 - X_2|) = 0$, as X_2 is distributed as X_1. Hence,

$$\mathrm{Cov}\,(X_1 + X_2, |X_1 - X_2|) = 0.$$

Two random variables X_1, X_2 with $\mathrm{Cov}\,(X_1, X_2) = 0$ are called *uncorrelated*.

The variance of sums of independent random variables

Let $X_1, \ldots, X_n$ be n independent random variables on the same probability space. Independence implies that for any $i < j$, we have $\mathrm{Cov}\,(X_i, X_j) = 0$. Using this in (2.6), we deduce that if $X_1, \ldots, X_n$ are independent, then

$$\mathrm{Var}\,(X_1 + \cdots + X_n) = \mathrm{Var}\,(X_1) + \cdots + \mathrm{Var}\,(X_n)\,. \tag{2.8}$$

2.7 Conditional distribution

Let us begin with a heuristic! Suppose that X, Y are two random variables on the same probability space $(\Omega, \mathcal{F}, \mathbb{P})$ which have joint density function $f : \mathbb{R}^2 \to [0, \infty)$. Let f_X, f_Y :

$\mathbb{R} \to [0, \infty)$ denote probability density functions of X and Y, respectively. Let x be such that $f_X(x) > 0$. We can use Bayes' rule and write

$$\begin{aligned}\mathbb{P}\left(Y \leq y \mid x \leq X \leq x + dx\right) &= \frac{\mathbb{P}\left(Y \leq y, x \leq X \leq x + dx\right)}{\mathbb{P}\left(x \leq X \leq x + dx\right)} \\ &\approx \frac{\int_{-\infty}^{y} f(x, z)dz}{f_X(x)dx} \\ &= \int_{-\infty}^{y} \frac{f(x, z)}{f_X(x)}dz. \end{aligned} \tag{2.9}$$

Let us now become more formal. Let $x \in \mathbb{R}$ be such that $f_X(x) > 0$. The *conditional density function* of the random variable Y given $X = x$ is the function $F_{Y|X}(\cdot, y) : \mathbb{R} \to [0, 1]$ such that

$$F_{Y|X}(y|x) = \mathbb{P}\left(Y \leq y \mid X = x\right) = \int_{-\infty}^{y} \frac{f(x, z)}{f_X(x)}dz.$$

The *conditional density function* $f_{Y|X}(\cdot|x) : \mathbb{R} \to [0, \infty)$ is defined as

$$f_{Y|X}(y|x) = \frac{f(x, y)}{f_X(x)}.$$

Recall however that f_X is related to f: $f_X(x) = \int_{-\infty}^{\infty} f(x, y)dy$, whereby one can write

$$f_{Y|X}(y|x) = \frac{f(x, y)}{\int_{-\infty}^{\infty} f(x, y)dy}$$

Example ([2], p.104)**.** Let X, Y be two random variables whose joint density function is

$$f(x, y) = \frac{1}{x}, \text{ for } 0 \leq y \leq x \leq 1,$$

and equal to 0, otherwise. Then the marginal probability density function of X is for $x > 0$

$$f_X(x) = \int_0^x f(x, y)dy = \frac{1}{x}\int_0^x dy = \frac{x}{x} = 1.$$

So the conditional density function of Y given X is

$$f_{Y|X}(y|x) = \frac{f(x, y)}{f_X(x)} = \frac{1}{x}.$$

In other words, Y is uniformly distirbuted on $[0, x]$, conditional on $X = x$.

2.7.1 The conditional expectation of Y given X

Using the above definitions, we will define the conditional expectation of Y given X. This is denoted by $\mathbb{E}\left(Y \mid X\right)$. This is the formal name we give to the function on X $\psi(X)$, where

$$\psi(x) = \mathbb{E}\left(Y \mid X = x\right) = \int_{-\infty}^{\infty} y f_{Y|X}(y, x)dy.$$

Thus, we *define*

$$\mathbb{E}\left(Y \mid X\right) = \psi(X).$$

NOTE: $\mathbb{E}(Y \mid X)$ is a random variable, NOT a number...

The expected value of this random variable can be calculated as follows:

$$\begin{aligned}\mathbb{E}(\mathbb{E}(Y \mid X)) &= \mathbb{E}(\psi(X)) = \int_{-\infty}^{\infty} \psi(x) f_X(x) dx = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} y f_{Y|X}(y,x) dy f_X(x) dx \\ &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} y \frac{f(x,y)}{f_X(x)} f_X(x) dy dx = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} y f(x,y) dy dx \\ &= \int_{-\infty}^{\infty} y f_Y(y) dy = \mathbb{E}(Y).\end{aligned}$$

Chapter 3

Convergence of random variables

3.0.1 The weak law of large numbers

The weak law of large numbers in its simplest form deals with sums of independent but identically distributed random variables defined on the same probability space. It states that no matter what this common distribution is, the average of large sums of these random variables is close to their expected value, with high probability. The latter means that as the number of variables involved in the sum becomes large, the probability that the average of the sum is close to the expected value tends to 1.

Let us now state the (weak) law of large number with precision.

Theorem 3.1 (The Weak Law of Large Numbers)**.** *Let $\{X_k\}_{k\in\mathbb{N}}$ be a sequence of random variables on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$. Suppose that these are independent but have the same distribution. Let $\mu = \mathbb{E}(X_k) < \infty$ be their common expected value and σ^2 their common variance. For any $\varepsilon > 0$, we have*

$$\mathbb{P}\left(\left|\frac{X_1 + \cdots + X_n}{n} - \mu\right| > \varepsilon\right) \to 0 \text{ as } n \to \infty.$$

Proof. Since the random variables are independent, we have

$$\mathrm{Var}(X_1 + \cdots + X_n) = \mathrm{Var}(X_1) + \cdots + \mathrm{Var}(X_n) = n\sigma^2.$$

Now, we write $S_n = X_1 + \cdots + X_n$ and therefore

$$\mathbb{P}\left(\left|\frac{X_1 + \cdots + X_n}{n} - \mu\right| > \varepsilon\right) = \mathbb{P}(|S_n - \mu n| > n\varepsilon).$$

But by the linearity of the expected value, we have $\mathbb{E}(S_n) = \mathbb{E}(X_1) + \cdots \mathbb{E}(X_n) = n\mu$. So we can bound the latter probability using Chebyschev's inequality:

$$\mathbb{P}(|S_n - \mu n| > n\varepsilon) \leq \frac{\mathrm{Var}(S_n)}{n^2\varepsilon^2} = \frac{n\sigma^2}{n^2\varepsilon^2} = \frac{\sigma^2}{\varepsilon^2} \cdot \frac{1}{n} \to 0,$$

as $n \to \infty$. □

So with $S_n = X_1 + \cdots + X_n$, the above states that S_n/n is within ε from μ with probability tending to 1, as $n \to \infty$.

Definition. Let $(X_k)_{k\in\mathbb{N}}$ be a sequence of random variables defined on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and let X be a random variable on the same probability space. We say that the sequence converges to X *in probability* $X_k \xrightarrow{p} X$ if for any $\varepsilon > 0$ we have

$$\mathbb{P}\left(|X_k - X| > \varepsilon\right) \to 0, \text{ as } k \to \infty.$$

Thus, the weak law of large numbers can be rephrased as:

$$\frac{1}{n} S_n \xrightarrow{p} \mu.$$

Remark. By deploying a more elaborate proof one can drop the assumption that the random variables have finite variance. In fact, the above theorem holds also for when the variance does not exist - one only needs that $\mathbb{E}\left(|X_k|\right) < \infty$. For a proof see [**?**].

3.0.2 The strong law of large numbers

One can strengthen the previous statement and actually show that for most $\omega \in \Omega$ the sequence $\frac{1}{n} S_n(\omega)$ converges to μ.

To make this more formal, consider a sequence of random variables $\{X_k\}_{k\in\mathbb{N}}$, that are identically distributed on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$. We say that the sequence satisfies the *strong law of large numbers*, if for any $\varepsilon > 0$ and any $\delta > 0$ there exists n_0 such that with probability at least $1 - \delta$ we have that for all $n > n_0$

$$|S_n/n - \mu| < \varepsilon.$$

Remark. Note the fundamental difference between the above statement and the weak law of large numbers. The latter amounts to saying that for each $n > n_0$ the event $|S_n/n - \mu| < \varepsilon$ has probability at least $1 - \delta$. However, the strong law of large numbers states that with probability $1 - \delta$ the events $|S_n/n - \mu| < \varepsilon$ hold *simultaneously* for all $n > n_0$.

A different way to view this kind of convergence is to consider for each $\omega \in \Omega$ the sequence $(S_k(\omega))_{k\in\mathbb{N}}$. Consider those ωs for which $(X_k(\omega))_{k\in\mathbb{N}}$ converges to μ. Let $\mathcal{C} \subset \Omega$ denote this set. The strong law of large numbers can be stated as

$$\mathbb{P}\left(\mathcal{C}\right) = 1.$$

We say that $\mathcal{C}$ *occurs almost surely (a.s.)*. In other words, the set of those ωs for which $(S_k(\omega))_{k\in\mathbb{N}}$ does not converge has $\mathbb{P}$- measure equal to 0. Note that this DOES NOT imply that this set is empty!!

Let us consider the event $\mathcal{C}$ more closely. Let $\omega \in \mathcal{C}$ and consider the sequence $(S_k(\omega))_{k\in\mathbb{N}}$. Recall the definition of convergence to μ: For any $\varepsilon > 0$ there exists k_0 such that for all $k > k_0$ we have

$$|\frac{1}{k} S_k - \mu| < \varepsilon.$$

How can we rephrase this? Let $A_k(\varepsilon)$ be the event that $|\frac{1}{k} S_k - \mu| < \varepsilon$. Saying that $\mathcal{C}$ occurs is equivalent to saying that $A_k(\varepsilon)$ *occurs eventually*.

By going to the complement of these events, we would like to say that the probability that the events $A_k(\varepsilon)^c$ occur *infinitely often* is 0. That is, the probability that *infinitely often* the difference $|\frac{1}{k} S_k - \mu|$ "jumps" above ε is 0. Or, more formally, the probability that there exists an infinite sequence $k_1, k_2, \ldots$ such that $|\frac{1}{k_i} S_{k_i} - \mu| \geq \varepsilon$, for all i, is 0.

How to express these more concisely? Let $(E_n)_{n\in\mathbb{N}}$ be a sequence of events on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$. What does it mean to say that the events E_n *occur infinitely often*? Take a $k \in \mathbb{N}$. Then at least one of the E_ns occurs for $n \geq k$; that is, $\cup_{n=k}^{\infty} E_n$ occurs. Moreover, this is the case for *any* $k \in \mathbb{N}$, that is, for any natural number k at least one of the E_ns occurs for $n \geq k$. In other words, $\cap_{k=1}^{\infty} \cup_{n=k}^{\infty} E_n$ occurs. We define the event

$$\{E_n \text{ occurs infinitely often}\} = \cap_{k=1}^{\infty} \cup_{n=k}^{\infty} E_n.$$

This motivates us to state and prove the following lemma which is known as the *first Borel-Cantelli* lemma.

Lemma 3.2 (The first Borel-Cantelli Lemma)**.** *Let $(E_n)_{n\in\mathbb{N}}$ be a collection of events on some probability space $(\Omega, \mathcal{F}, \mathbb{P})$. The following holds:*

$$\textit{if } \sum_{n=1}^{\infty} \mathbb{P}(E_n) < \infty, \textit{ then } \mathbb{P}(\{E_n \textit{ occurs infinitely often}\}) = 0.$$

Proof. Set $B_k = \cup_{n=k}^{\infty} E_n$; thus, $\{E_n \text{ occurs infinitely often}\} = \cap_{k=1}^{\infty} B_k$. The B_ks form a decreasing family and therefore (from Practice Question 1, in Examples Sheet 1)

$$\mathbb{P}(\{E_n \text{ occurs infinitely often}\}) = \lim_{k\to\infty} \mathbb{P}(B_k). \tag{3.1}$$

We can also write $B_k = \bigcup_{k'\geq k} \cup_{n=k}^{k'} E_n$, whereby (see Examples Sheet 1),

$$\mathbb{P}(B_k) = \lim_{k'\to\infty} \mathbb{P}\left(\cup_{n=k}^{k'} E_n\right).$$

But by the union bound,

$$\mathbb{P}\left(\cup_{n=k}^{k'} E_n\right) \leq \mathbb{P}(E_k) + \cdots + \mathbb{P}(E_{k'}),$$

and therefore

$$\mathbb{P}(B_k) \leq \sum_{n=k}^{\infty} \mathbb{P}(E_n).$$

Now, note that the assumption that $\sum_{n=1}^{\infty} \mathbb{P}(E_n) < \infty$ implies that the partial sums

$$\sum_{n=k}^{\infty} \mathbb{P}(E_n) \to 0, \text{ as } k \to \infty.$$

In turn, this implies that,

$$\lim_{k\to\infty} \mathbb{P}(B_k) = 0.$$

The lemma follows by the above and (3.1). □

We are going to prove that an infinite sequence of identically distributed random variables $\{X_k\}_{k\in\mathbb{N}}$, defined on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, having bounded expected value μ satisfies the strong law of large numbers.

Theorem 3.3 (The Strong Law of Large Numbers)**.** *Let $\{X_k\}_{k\in\mathbb{N}}$ be a sequence of independent and identically distributed random variables on a probability space $(\Omega, \mathcal{F}_n, \mathbb{P})$ such that $\mathbb{E}(X_k) = \mu$ and $\mathrm{Var}(X_k) = \sigma^2 < \infty$. Set $S_n = \sum_{k=1}^{n} X_k$. Then as $n \to \infty$*

$$\frac{1}{n} S_n \overset{a.s.}{\to} \mu.$$

Proof. (From [2] pp. 326-327.) Since the random variables $X_1, \ldots, X_n$ are independent, we have that

$$\mathrm{Var}\,(S_n) = \mathrm{Var}\,(X_1 + \cdots + X_n) = \mathrm{Var}\,(X_1) + \cdots + \mathrm{Var}\,(X_n) = n\sigma^2.$$

Consider the sub-sequence $n_\ell = \ell^2$ for $\ell \in \mathbb{N}$. Now, by Chebyschev's inequality we have

$$\mathbb{P}\left(\left|\frac{1}{\ell^2}S_{n_\ell} - \mu\right| > \varepsilon_1\right) \leq \frac{\mathrm{Var}\,(S_{n_\ell})}{\varepsilon_1^2\ell^4} = \frac{\ell^2\sigma^2}{\varepsilon^2\ell^4} = \frac{\sigma^2}{\varepsilon_1^2\ell^2}.$$

Set $A_\ell = \{\left|\frac{1}{\ell^2}S_{n_\ell} - \mu\right| > \varepsilon_1\}$. We will set ε_1 later. The above inequality implies that

$$\sum_{\ell=1}^{\infty}\mathbb{P}\,(A_\ell) = \frac{\sigma^2}{\varepsilon_1^2}\sum_{\ell=1}^{\infty}\frac{1}{\ell^2} < \infty.$$

The first Borel-Cantelli lemma implies that

$$\mathbb{P}\,(\{A_\ell \text{ occurs infinitely often}\}) = 0.$$

This proves the a.s. convergence along the subsequence n_ℓ only. What about the rest? We consider two cases.

Case 1: $X_k \geq 0$

Assume first that $X_k \geq 0$. This means that S_n is non-decreasing in n. Taking n such that $n_\ell < n < n_{\ell+1}$ we have

$$S_{n_\ell} < S_n < S_{n_{\ell+1}},$$

which means that

$$\frac{1}{n_{\ell+1}}S_{n_\ell} < \frac{1}{n}S_n < \frac{1}{n_\ell}S_{n_{\ell+1}}.$$

But as $\ell \to \infty$, we have $n_{\ell+1}/n_\ell \to 1$. This implies that there exists $\ell_0 = \ell_0(\varepsilon)$ such that for all $\ell > \ell_0$ we have

$$1 - \varepsilon < \frac{n_\ell}{n_{\ell+1}} \text{ and } \frac{n_{\ell+1}}{n_\ell} < 1 + \varepsilon.$$

Now, we can write the above double inequality as:

$$\frac{n_\ell}{n_{\ell+1}}\frac{1}{n_\ell}S_{n_\ell} < \frac{1}{n}S_n < \frac{n_{\ell+1}}{n_\ell}\frac{1}{n_{\ell+1}}S_{n_{\ell+1}}.$$

Therefore, when $\ell > \ell_0$ we have

$$(1-\varepsilon)\frac{1}{n_\ell}S_{n_\ell} < \frac{1}{n}S_n < \frac{1}{n_{\ell+1}}S_{n_{\ell+1}}(1+\varepsilon).$$

So if both A_ℓ and $A_{\ell+1}$ have not occurred, then

$$\frac{1}{n_\ell}S_{n_\ell} > \mu - \varepsilon \text{ and } \frac{1}{n_{\ell+1}}S_{n_{\ell+1}} < \mu + \varepsilon$$

So in this case

$$(\mu-\varepsilon)(1-\varepsilon_1) < \frac{1}{n}S_n < (\mu+\varepsilon)(1+\varepsilon_1).$$

But

$$(\mu+\varepsilon_1)(1+\varepsilon_1) = \mu+\mu\varepsilon_1+\varepsilon_1+\varepsilon_1^2 \overset{\varepsilon<1}{<} \mu+\mu\varepsilon_1+\varepsilon_1+\varepsilon_1 = \mu+(\mu+2)\varepsilon_1,$$

and

$$(\mu-\varepsilon_1)(1-\varepsilon_1) = \mu-\mu\varepsilon_1-\varepsilon_1+\varepsilon_1^2 > \mu-(\mu+1)\varepsilon_1 > \mu-(\mu+2)\varepsilon_1.$$

Therefore,

$$\left|\frac{1}{n}S_n-\mu\right| < (\mu+2)\varepsilon_1.$$

Let $B_n((\mu+2)\varepsilon_1)$ be the event $\left|\frac{1}{n}S_n-\mu\right| \geq (\mu+2)\varepsilon_1$. Thus, since $\mathbb{P}\left(\{A_\ell \text{ occurs infinitely often}\}\right) = 0$ it follows that

$$\mathbb{P}\left(\{B_n((\mu+2)\varepsilon_1) \text{ occurs infinitely often}\}\right) = 0.$$

Now, we set $\varepsilon_1 = \frac{\varepsilon}{\mu+2}$. In this case, $B_n((\mu+2)\varepsilon_1)$ coincides with $A_n(\varepsilon)^c$, whereby

$$\mathbb{P}\left(\{A_n(\varepsilon)^c \text{ occurs infinitely often}\}\right) = 0.$$

Thus, as $n\to\infty$

$$\frac{1}{n}S_n \overset{a.s.}{\to} \mu, \text{ under the assumption that } X_k \geq 0. \tag{3.2}$$

The general case

If we do not have $X_k \geq 0$, then we resort to the following trick. We write

$$X_k = X_k^+ - X_k^-,$$

where

$$X_k^+(\omega) = \begin{cases} X_k(\omega) & \text{if } X_k \geq 0, \\ 0 & \text{otherwise} \end{cases} \text{ and } X_k^-(\omega) = \begin{cases} 0 & \text{if } X_k \geq 0, \\ |X_k(\omega)| & \text{otherwise} \end{cases}.$$

Hence,

$$S_n = \sum_{k=1}^n X_k = \sum_{k=1}^n (X_k^+ - X_k^-) = \sum_{k=1}^n X_k^+ - \sum_{k=1}^n X_k^- =: S_n^+ - S_n^-.$$

Now, if $\mathbb{E}\left(X_k^+\right) = \mu^+$ and $\mathbb{E}\left(X_k^-\right) = \mu^-$, the linearity of the expected value implies that $\mu = \mu^+ - \mu^-$. Observe that both $X_k^+, X_k^- \geq 0$. So by (3.2), we have as $n\to\infty$

$$\frac{1}{n}S_n^+ \overset{a.s.}{\to} \mu^+ \text{ and } \frac{1}{n}S_n^- \overset{a.s.}{\to} \mu^-.$$

We thus conclude that as $n\to\infty$

$$\frac{1}{n}S_n \overset{a.s.}{\to} \mu^+ - \mu^- = \mu.$$

□

Remarks

1. One can lift the assumption that the random variables X_k are identically distributed. If X_k has variance $\sigma_k^2 < \infty$, the strong law of large numbers holds, if

$$\sum_{k=1}^{\infty} \frac{\sigma_k^2}{k^2} < \infty. \text{ (the Kolmogorov criterion).}$$

   In other words, the variances should not be growing too fast.

2. If we keep the assumption that the random variables X_k are identically distributed, then the above theorem also holds without the assumption that $\mathrm{Var}\,(X_k) < \infty$. In fact, it suffices that $\mathbb{E}\,(|X_k|) < \infty$.

3.1 Convergence in distribution: the central limit theorem

Another kind of convergence of random variables is that of *convergence in distribution*. Here, the sequence of variables do not need to be defined on the same probability space. Let $\{X_k\}_{k\in\mathbb{N}}$ be a sequence of random variables, where X_k is defined on $(\Omega_k, \mathcal{F}_k, \mathbb{P}_k)$. Let X be a random variable defined on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$. Let F_k be the cumulative distribution function of X_k and F be the cumulative distribution function of X. We say that the sequence X_k *converges in distribution* to X (writing, $X_k \xrightarrow{d} X$), if $F_k(x) \to F(x)$, as $k \to \infty$, whenever x is such that F is continuous on x. In other words, for any such x we have $\mathbb{P}_k(X_k \leq x) \to \mathbb{P}(X \leq x)$, as $k \to \infty$.

Example. Consider the sequence of independent and identically distributed random variables $\{X_k\}_{k\in\mathbb{N}}$ defined on the same probability space $(\Omega, \mathcal{F}, \mathbb{P})$. Assume that the assumptions of Theorem 3.1 hold. So with $S_n = X_1 + \ldots + X_n$ and $\mu = \mathbb{E}\,(X_k)$, we have

$$\mathbb{P}\left(\left|\frac{1}{n}S_n - \mu\right| > \varepsilon\right) \to 0.$$

Let us re-interpret this in terms of convergence in distribution. Let X be the random variable on $(\Omega, \mathcal{F}, \mathbb{P})$ such that $X(\omega) = \mu$ for every $\omega \in \Omega$. The cumulative distribution function F of X is

$$F(x) = \mathbb{P}\,(X \leq x) = \begin{cases} 1, & \text{if } x \geq \mu, \\ 0, & \text{if } x < \mu \end{cases}.$$

Note that F is not continuous at $x = \mu$, but it is continuous everywhere else.

Let $x < \mu$. We have

$$\begin{aligned}\mathbb{P}\left(\frac{1}{n}S_n \leq x\right) &= \mathbb{P}\left(\frac{1}{n}S_n \leq \mu - \mu + x\right) = \mathbb{P}\left(\frac{1}{n}S_n - \mu \leq -(\mu - x)\right) \\ &\leq \mathbb{P}\left(\left|\frac{1}{n}S_n - \mu\right| > \mu - x\right) \to 0 = F(x).\end{aligned}$$

Now, if $x > \mu$, then

$$\begin{aligned}\mathbb{P}\left(\frac{1}{n}S_n \leq x\right) &= \mathbb{P}\left(\frac{1}{n}S_n \leq \mu - \mu + x\right) = \mathbb{P}\left(\frac{1}{n}S_n - \mu \leq -(\mu - x)\right) \\ &\leq \mathbb{P}\left(\left|\frac{1}{n}S_n - \mu\right| < \mu - x\right) \to 1 = F(x).\end{aligned}$$

In other words, $\frac{1}{n}S_n$ converges in distribution to X.

More generally, the following is true.

Proposition 3.4. *Let $\{X_k\}_{k\in\mathbb{N}}$ be a sequence of random variables on $(\Omega, \mathcal{F}, \mathbb{P})$ which converge in probability to a random variable X on the same probability space. Then X_k converges in distribution to X.*

Proof. Let $F_k(x) = \mathbb{P}(X_k \leq x)$ and $F(x) = \mathbb{P}(X \leq x)$. We write

$$\begin{aligned} F_k(x) = \mathbb{P}(X_k \leq x) &= \mathbb{P}(X_k \leq x, X \leq x+\varepsilon) + \mathbb{P}(X_k \leq x, X > x+\varepsilon) \\ &\leq \mathbb{P}(X \leq x+\varepsilon) + \mathbb{P}(X - X_k > \varepsilon) \leq F(x+\varepsilon) + \mathbb{P}(|X - X_k| > \varepsilon). \end{aligned}$$

Similarly,

$$\begin{aligned} F(x-\varepsilon) = \mathbb{P}(X \leq x-\varepsilon) &= \mathbb{P}(X \leq x-\varepsilon, X_k \leq x) + \mathbb{P}(X \leq x-\varepsilon, X_k > x) \\ &\leq \mathbb{P}(X_k \leq x) + \mathbb{P}(X_k - X > \varepsilon) \leq F_k(x) + \mathbb{P}(|X - X_k| > \varepsilon). \end{aligned}$$

Combining the above two, we deduce that

$$F(x-\varepsilon) - \mathbb{P}(|X - X_k| > \varepsilon) \leq F_k(x) \leq F(x+\varepsilon) + \mathbb{P}(|X - X_k| > \varepsilon).$$

If x is a point of continuity for F, then for some $\delta = \delta > 0$,

$$F(x) - \delta < F(x+\varepsilon) \leq F(x) + \delta.$$

In fact continuity means that for any δ we can select ε such that the above holds. Also, convergence in probability implies that

$$\mathbb{P}(|X - X_k| > \varepsilon) \to 0, \text{ as } k \to \infty.$$

So, there exists $k_0 = k_0(\varepsilon)$ such that for any $k > k_0$, we have

$$\mathbb{P}(|X - X_k| > \varepsilon) \leq \delta.$$

We conclude that for any $k > k_0$, we have

$$F(x) - 2\delta \leq F_k(x) \leq F(x) + 2\delta.$$

In other words, $F_k(x) \to F(x)$ as $k \to \infty$. □

3.1.1 The central limit theorem

We just saw that for a collection of independent and identically distributed random variables $\{X_k\}_{k\in\mathbb{N}}$, the sum S_n scaled by n converges in distribution to a random variable which is trivial, in that it is equal to $\mu = \mathbb{E}(X_k)$, with probability 1. In fact, one can show a general result that is non-trivial in this sense.

Consider the above setting. Recall that $\mathbb{E}(S_n) = n\mu$ and $\mathrm{Var}(S_n) = n\sigma^2$, as the X_ks are independent, assuming that each has variance $\sigma^2 < \infty$. Let

$$\hat{S}_n = \frac{S_n - n\mu}{\sqrt{n\sigma^2}}.$$

Then $\mathbb{E}\left(\hat{S}_n\right) = 0$ and $\mathrm{Var}\left(\hat{S}_n\right) = 1$ (check!!). The following theorem is the celebrated *central limit theorem*.

Theorem 3.5 (the Central Limit Theorem)**.** *Let $\{X_k\}_{k\in\mathbb{N}}$ be a sequence of independent and identically distributed random variables on the same probability space. Suppose that $\mathbb{E}(X_k) = \mu < \infty$ and $\mathrm{Var}(X_k) = \sigma^2 < \infty$. If N is a random variable that follows the standard normal distribution, then*

$$\hat{S}_n \xrightarrow{d} N, \text{ as } n \to \infty.$$

In other words, under the assumptions of the above theorem, for any $x \in \mathbb{R}$, we have

$$\mathbb{P}\left(\frac{S_n - n\mu}{\sqrt{n\sigma^2}} \leq x\right) \to \Phi(x) := \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-x^2/2} dx.$$

We will see a proof of this, for the case where X_k are Bernoulli-distributed random variables, that is, S_n follows the binomial distribution.

3.1.2 The case of the binomial distribution: the DeMoivre-Laplace theorem

Now, consider the case of a collection of independent Bernoulli-distributed random variables with parameter $p \in [0,1]$. In other words, we consider the space Ω_n consisting of all words on letters $\{H, T\}$ of length n, where each word $w = a_1 \cdots a_n \in \Omega_n$ has probability equal to $\prod_{i=1}^n f(a_i)$, where $f(a_i) = p$, if $a_i = H$, but $f(a_i) = 1 - p$, if $a_i = T$.

Let X_k be the random variable on Ω which is equal to 1 if the kth letter is H and 0 otherwise. We have seen that $\mathbb{P}(X_k = 1) = p$ and the random variables $\{X_k\}_{k=1,\ldots,n}$ are independent. Their sum $S_n := \sum_{k=1}^n X_k$ follows the binomial distribution. It takes values on the set $\{0, 1, \ldots, n\}$ and $\mathbb{P}(S_n = k) = \binom{n}{k} p^k (1-p)^{n-k}$. Recall also that $\binom{n}{k} = \frac{n!}{k!(n-k)!}$. We have seen that $\mathbb{E}(S_n) = np$. Regarding the variance, we have by the independence of the X_ks

$$\mathrm{Var}(S_n) = \mathrm{Var}(X_1 + \cdots + X_k) = \mathrm{Var}(X_1) + \cdots + \mathrm{Var}(X_n).$$

But for each k, we have $\mathrm{Var}(X_k) = \mathbb{E}(X_k^2) - \mathbb{E}(X_k)^2 = p - p^2 = p(1-p)$ and therefore

$$\mathrm{Var}(S_n) = np(1-p).$$

We shall consider the random variable:

$$\hat{S}_n = \frac{S_n - np}{\sqrt{np(1-p)}}.$$

We are first going to prove the following *local limit theorem*.

Theorem 3.6. *Let $p \in [0,1]$ and $x \in \mathbb{R}$. Then as $n \to \infty$*[1]

$$\mathbb{P}\left(S_n = np + x\sqrt{np(1-p)}\right) \to \frac{1}{\sqrt{2\pi}} e^{-x^2/2}.$$

Proof. To this end, we area going to use without proof what is widely know as *Stirling's approximation* for the factorial.

[1] One should have $\lfloor x\sqrt{np(1-p)} \rfloor$ instead of $x\sqrt{np(1-p)}$, but it does not make a difference in the calculations and we omit it.

Proposition 3.7 (The Stirling approximation)**.** *Let* $t(s) = \sqrt{2\pi s} s^s e^{-s}$*. We have that*

$$\frac{s!}{t(s)} \to 1, \textit{ as } s \to \infty.$$

Let us write $q := 1 - p$; thus, $q + p = 1$. We will use this in order to give an asymptotic approximation to $\mathbb{P}(S_n = k) = \binom{n}{k} p^k q^{n-k}$ where $k = np + x\sqrt{npq}$. Hence, as $n \to \infty$ we also have that $k \to \infty$ and also $n - k = nq - x\sqrt{npq}$. We will use the following equalities:

$$\frac{k}{np} = 1 + x\sqrt{\frac{q}{np}} \text{ and } \frac{n-k}{nq} = 1 - x\sqrt{\frac{p}{nq}}. \tag{3.3}$$

Firstly, we apply Stirling's approximation (cf. Proposition 3.7)

$$\begin{aligned}\frac{n!}{k!(n-k)!} &= \sqrt{\frac{n}{2\pi k(n-k)}} \cdot \frac{n^n}{k^k (n-k)^{n-k}} \cdot \frac{e^{-n}}{e^{-k} \cdot e^{-(n-k)}} \\ &= \sqrt{\frac{n}{2\pi k(n-k)}} \cdot \left(\frac{k}{n}\right)^{-k} \left(\frac{n-k}{n}\right)^{-(n-k)}.\end{aligned}$$

This last expression together with $p^k q^{n-k}$ give

$$\begin{aligned}\left(\frac{k}{n}\right)^{-k} \left(\frac{n-k}{n}\right)^{-(n-k)} \cdot p^k q^{n-k} &= \left(\frac{k}{np}\right)^{-k} \left(\frac{n-k}{nq}\right)^{-(n-k)} \\ &\overset{(3.3)}{=} \left(1 + x\sqrt{\frac{q}{np}}\right)^{-k} \left(1 - x\sqrt{\frac{p}{nq}}\right)^{-(n-k)}.\end{aligned}$$

□

Chapter 4

Markov Chains

4.1 Stochastic processes

A *(discrete time) stochastic process* is an infinite collection of random variables $X_0, X_1, \ldots$ defined on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$. One would usually think of the random variable X_n as describing some quantity at time n. For example, one could think of the random variable X_n as being the value of an asset at the end of the nth day after some specific date. Just before that date the value of the asset was given by X_0.

Thus, for any $\omega \in \Omega$, the sequence $X_0(\omega), X_1(\omega), \ldots$ is a *path* of the evolution of the quantity described by the random variables X_n over time. Let us consider a classic example.

Example (Random walks on the integers)**.** Let us recall the probability space Ω_∞. This consists of all infinite sequences of letters in $\{H, T\}$ with the σ-algebra $\mathcal{F}_\infty$, which is generated by all subsets of Ω_∞ which start with a certain prefix. If the prefix has length n, then such a set has $\mathbb{P}_\infty$-measure equal to $1/2^n$. For every such $\omega \in \Omega_\infty$, let ω_k denote the letter at the kth place. Let I_k be the random variable such that $I_k(\omega) = +1$, if $\omega_k = H$ and $I_k(\omega) = -1$, if $\omega_k = T$. Hence, $\mathbb{P}(I_k = +1) = \mathbb{P}(I_k = -1) = 1/2$. Let $X_n = \sum_{k=1}^{n} I_k$, for $n \in \mathbb{N}$ and $X_0 = 0$.

This stochastic process is known is the *unbiased random walk* on the integers starting at 0. If the random is on number s after step n, at the next step it moves to $s + 1$ with probability $1/2$ or to -1 with probability $1/2$.

In the above example, the stochastic process evolves over the set of integers. This is what is called its *state space*. For a stochastic process, the set S which X_n belongs to is called *the state space* of the process.

This notion may be confusing as the X_ns are random variables and take values over the real numbers. However, the state space may not be a subset of $\mathbb{R}$. To be formally correct, we identify S with a subset of the real numbers.

4.2 Markov chains

An important class of discrete time stochastic processes is that of Markov chains. Informally, a Markov chain is a discrete time stochastic process in which just after step n, the distribution of the state of the process after step $n + 1$ depends only on the state at step n. More, precisely a Markov chain is a discrete time stochastic process $X_0, X_1, \ldots$, whose

state space S is discrete, that is, it can identified with a subset of the natural numbers, and satisfies the following condition:

$$\mathbb{P}\left(X_{n+1} = y \mid X_n = x, \{X_0 = x_0, \ldots, X_{n-1} = x_{n-1}\}\right) = \mathbb{P}\left(X_{n+1} = y \mid X_n = x\right) = P(x, y),$$

for any $x, y, x_0, \ldots, x_{n-1} \in S$. Hence, where the process will go at step $n+1$, given that it is currently x does not depend on the previous history, but only on the fact that is is at x.

If S is a finite set, then we can think of P as an $|S| \times |S|$-matrix whose entries are numbers in $[0, 1]$. This is called the *transition matrix*. Moreover, the row corresponding to $x \in S$ and consists of all entries $P(x, \cdot)$ is a *probability distribution* over S. Therefore, for any $x \in S$, we have $\sum_{y \in S} P(x, y) = 1$. In other words, the entries of each add up to 1. Such a matrix is called a *stochastic matrix*. From now on, we will be assuming that the state space S is finite.

This is particularly convenient as it enables us to express succinctly the evolution of the Markov chain. To be more specific, suppose that the initial state X_0 is selected from S according to the probability distribution p_0 on S. What is the probability distribution of the state of the chain after the first step? We can write for any $y \in S$

$$\begin{aligned}\mathbb{P}\left(X_1 = y\right) &= \sum_{x \in S} \mathbb{P}\left(X_0 = x, X_1 = y\right) = \sum_{x \in S} \mathbb{P}\left(X_0 = x\right) \mathbb{P}\left(X_1 = y \mid X_0 = x\right) \\ &= \sum_{x \in S} p_0(x) P(x, y).\end{aligned}$$

If we set $p_1(y) = \mathbb{P}\left(X_1 = y\right)$ and view p_0, p_1 as vectors of dimension $|S|$ with entries in $[0, 1]$, the above can be written as:

$$p_1 = p_0 P.$$

More generally, if p_n is the distribution of the state of the chain just after the nth step, then

$$p_{n+1} = p_n P.$$

Repeating this, we have

$$p_{n+1} = p_{n-1} P^2 = p_{n-2} P^3 = \cdots$$

meaning that P^t describes that t-step transition probabilities of the Markov chain. If $P^t(x, y)$ is the x, y entry of P^t, then this is the probability that the chain is at state y after t steps given that it has started at state x.

A probability distribution π on S is called *stationary* or *invariant* if

$$\pi = \pi P.$$

Thus, if the initial state X_0 is selected according to π, then X_n is distributed according to π for any $n \geq 1$.

Example (Random walks on graphs)**.** Let $G = (V, E)$ be a finite graph on $s = |V|$ vertices that are labelled by the numbers $1, \ldots, s$ (and, therefore, V is identified with the set $\{1, \ldots, s\} \subseteq \mathbb{R}$). A *simple random walk* on G is a Markov chain with state space the set of vertices V and transition matrix P such that $P(x, y) = 1/d(x)$, where $d(x)$ is the degree of vertex x, that is, the number of neighbours of x in G (provided that $d(x) > 0$). If $d(x) = 0$, then for any $y \in V$, we set $P(x, y) = 0$, for $y \neq x$ and $P(x, x) = 1$. In simple terms, if $X_n = x$, then at the next step the state is one of the neighbours of x selected uniformly at

random. One may think of this as particle that moves from vertex to vertex, where if the particle is at vertex x at step n, it jumps to one of its neighbours, selecting one of them uniformly. Note that if a vertex x has degree $d(x) = 0$, then if the chain starts at x, then it stays at x for ever.

The vector π whose entries are indexed by the vertices in V such that for $x \in V$ we have

$$\pi(x) = \frac{d(x)}{2|E|},$$

is an invariant distribution of the Markov chain. Indeed, for a vertex $y \in V$

$$\sum_{x \in V} \pi(x) P(x, y) = \sum_{x : xy \in E} \frac{d(x)}{2|E|} \cdot \frac{1}{d(x)} = \sum_{x : xy \in E} \frac{1}{2|E|} = \frac{d(y)}{2|E|}.$$

Example (Branching processes - Galton-Watson processes)**.** This is an example of a Markov chain whose state space is $\mathbb{N} \cup \{0\}$. The process $X_0, X_1, \ldots$ describes the nth generation of a population. Initially, $X_0 = 1$, that is, there is only one individual. This individual gives birth to random number of children which is a random variable Z taking values in $\mathbb{N} \cup \{0\}$ that has expected value equal to $\rho > 0$. Any individual which has had the chance to reproduce ceases to reproduce after this. Its children form the first generation and their number is X_1. Next each one of the X_1 members of the first generation gives birth to a random number of children distributed as Z, independently. These form the second generation. In other words,

$$X_2 = \sum_{i=1}^{X_1} Z_i,$$

where $Z_1, \ldots, Z_{X_1}$ are independent random variables that are distributed as Z. In general, if the nth generation has X_n individuals, then the $n+1$th generation has

$$X_{n+1} = \sum_{i=1}^{X_n} Z_i^{(n)},$$

where $Z_1^{(n)}, \ldots, Z_{X_1}^{(n)}$ are independent random variables distributed as Z. If $X_n = k$, then

$$\mathbb{E}\left(X_{n+1} | X_n = k\right) = \mathbb{E}\left(\sum_{i=1}^{k} Z_i^{(n)}\right) = \sum_{i=1}^{k} \mathbb{E}\left(Z_i^{(n)}\right) = \rho k.$$

Hence,

$$\begin{aligned}\mathbb{E}\left(X_{n+1}\right) &= \sum_{k=0}^{\infty} \mathbb{E}\left(X_{n+1} | X_n = k\right) \cdot \mathbb{P}\left(X_n = k\right) = \sum_{k=0}^{\infty} \rho k \cdot \mathbb{P}\left(X_n = k\right) \\ &= \rho \sum_{k=0}^{\infty} k \cdot \mathbb{P}\left(X_n = k\right) = \rho \mathbb{E}\left(X_n\right).\end{aligned}$$

As, $ExX_0 = 1$, we have $\mathbb{E}\left(X_n\right) = \rho^n$. In particular, if $\rho < 1$, the expected size of the nth generation drops exponentially. The fundamental theorem of Galton-Watson processes states that:

1. If $\rho \leq 1$ and $\mathrm{Var}\left(Z\right) > 0$, then with probability 1 the process dies out.

2. If $\rho > 1$, then the process dies out with some probability $0 < p < 1$. (Hence, with probability $1 - p$ it survives forever.)

Example (The Ehrenfest model)**.** This probabilistic model is motivated from Physics. It was introduced by T. and P. Ehrenfest (*Physikalische Zeitschrift, vol. 8 (1907), pp. 311–314*) in order to describe how two gases mix and arrive in equilibrium, trying to explain the second law of thermodynamics. Suppose that we have two chambers A and B that are isolated from the environment but they can communicate through a small hole. A gas with n particles is distributed among the two chambers - this is the initial configuration. Thereafter, the parts of the gas start to mix through this little hole on the wall that separates the two chambers. What happens is that at each round a particle selected at random passes through the whole and changes chamber.

We can view this as an urn that contain n balls of colours A and B. Balls of colour A represent the particles that are in chamber A. At each round, one ball is selected uniformly at random and removed from the urn and another ball of the "opposite" colour is added, that is, the corresponding particle changes chamber.

This is a Markov chain whose state space is the numbers $\{0, \dots, n\}$ representing the number of particles which are in chamber A. Here, the transition matrix $P(x, y)$ is non-zero only if $x = i$ and $y = i+1$, $i < n$ or $x = i$ and $y = i-1$, $i > 1$. In particular, $P(i, i+1) = \frac{n-i}{n}$ and $P(i, i-1) = \frac{i}{n}$.

The binomial distribution with parameters n, $1/2$ is the stationary distribution. That is, the vector π with $\pi(i) = \binom{n}{i}\left(\frac{1}{2}\right)^n$. Indeed, for $0 < i < n$

$$\begin{aligned}\pi(i) &= \sum_{j=0}^{n} \pi(j)P(j,i) = \pi(i-1)P(i-1,i) + \pi(i+1)P(i+1,i)\\ &= \binom{n}{i-1}\left(\frac{1}{2}\right)^n \frac{n-(i-1)}{n} + \binom{n}{i+1}\left(\frac{1}{2}\right)^n \frac{i+1}{n}.\end{aligned}$$

But

$$\binom{n}{i-1}\frac{n-(i-1)}{n} = \frac{n!}{(i-1)!(n-i+1)!}\frac{n-(i-1)}{n} = \frac{(n-1)!}{(i-1)!(n-i)!} = \binom{n-1}{i-1}$$

and

$$\binom{n}{i+1}\frac{i+1}{n} = \frac{n!}{(i+1)!(n-i-1)!}\frac{i+1}{n} = \frac{(n-1)!}{i!(n-i-1)!} = \binom{n-1}{i}.$$

But

$$\binom{n-1}{i-1} + \binom{n-1}{i} = \binom{n}{i}.$$

So

$$\pi(i) = \binom{n}{i}\left(\frac{1}{2}\right)^n.$$

Verify this for $i = 0$ and $i = n$.

4.3 Aperiodicity and irreducibility

Recall that the state space S is the set in which the Markov chain takes values in, that is, for each n $X_n \in S$. For states $x, y \in S$, recall that $P^t(x, y)$ is the probability of moving from

state x to state y in exactly t steps. If $P^t(x,y) > 0$, then this means that it is possible to go from x to y in exactly t steps.

A Markov chain with transition matrix P and state space S is called *irreducible*, if for all states $x, y \in S$, there is $t = t(x,y)$ such that $P^t(x,y) > 0$.

Another important notion is that of the *period* of a state $x \in S$. Let $\mathcal{T}(x) = \{j : P^j(x,x) > 0\}$. This is the set of numbers $j > 0$ for which the Markov chain can go from state x to itself in exactly j steps. The *period* of x is defined to be the greatest common divisor of the elements of $\mathcal{T}(x)$. If the period of x is equal to 1, then the state x is called *aperiodic*. Furthermore, if all states are aperiodic, then the Markov chain is called *aperiodic*.

4.4 The stationary distribution and the convergence theorem

For a state $x \in S$ we define its *hitting time* τ_x as

$$\tau_x = \min\{t \geq 0 \ : \ X_t = x\},$$

that is, this is the first time state x is visited. Note that according to this definition, if the Markov chain actually starts at x, then $\tau_x = 0$. We would also be interested in

$$\tau_x^+ = \min\{t > 0 \ : \ X_t = x\},$$

which is the first positive time at which state x is visited. In the case where $X_0 = x$, this quantity is called the *first return time*.

The following theorem establishes the existence of a stationary distribution of an irreducible Markov chain and gives its value explicitly as a function of the first return times.

Theorem 4.1 (Prop 1.14 [**?**])**.** *Let P be the transition matrix of an irreducible Markov chain with state space S. Then*

i. there exists a probability distribution π on S such that $\pi = \pi P$ and $\pi(x) > 0$ for all $x \in S$.

ii. For all $x \in S$

$$\pi(x) = \frac{1}{\mathbb{E}\left(\tau_x^+ \mid X_0 = x\right)}.$$

In other words, the probability weight the stationary distribution assigns to a state x is the inverse of the expected first return time to x. Note that the above holds under the assumption of irreducibility.

The next theorem is central in the theory of Markov chains on finite state spaces. It states that the distribution of X_n converges to the stationary distribution π as $n \to \infty$. To make sense of this, we need to define some kind of a distance between probability distributions on S. If p_1, p_2 are two probability distributions on S, we define the *total variation distance* between p_1 and p_2 as

$$d_{TV}(p_1, p_2) = \frac{1}{2} \sum_{x \in S} |p_1(x) - p_2(x)|.$$

Now, we can speak about convergence more precisely. Recall that if P is the transition matrix of the Markov chain, then $P^n(x,y)$ is the probability that the Markov chain goes from x to y in exactly n steps. Thus, for a given x, the probability distribution of X_n given that $X_0 = x$ is given by $P^n(x, \cdot)$, that is, the row corresponding to x.

Theorem 4.2. *Consider an irreducible and aperiodic Markov chain on a finite state space S with transition matrix P. Then for any $x \in S$ as $n \to \infty$ we have*

$$d_{TV}(P^n(x, \cdot), \pi) \to 0.$$

Bibliography

[1] K-L. Chung and F. AitSahlia, *Elementary Probability Theory: with stochastic processes and an introduction to mathematical finance*, Springer, 4th edition, 2004.

[2] G. Grimmett and D. Stirzaker, *Probability and Random Processes*, Oxford University Press, 3rd edition, 2001.