INTRODUCTION TO PROBABILITY

ARIEL YADIN

Course: 201.1.8001  Fall 2013-2014

Lecture notes updated: March 2, 2014

Contents

Lecture 1. 3
Lecture 2. 9
Lecture 3. 18
Lecture 4. 24
Lecture 5. 33
Lecture 6. 38
Lecture 7. 44
Lecture 8. 51
Lecture 9. 57
Lecture 10. 66
Lecture 11. 73
Lecture 12. 76
Lecture 13. 81
Lecture 14. 87
Lecture 15. 92
Lecture 16. 101
1.1. Example: Bertrand’s Paradox

We begin with an example [this is known as Bertrand’s paradox].

Question 1.1. Consider a circle of radius 1, and an equilateral triangle bounded in the circle, say ABC. (The side length of such a triangle is \( \sqrt{3} \).) Let \( M \) be a randomly chosen chord in the circle.

What is the probability that the length of \( M \) is greater than the length of a side of the triangle (i.e. \( \sqrt{3} \))? 

Solution 1. How to chose a random chord \( M \)?

One way is to choose a random angle, let \( r \) be the radius at that angle, and let \( M \) be the unique chord perpendicular to \( r \). Let \( x \) be the intersection point of \( M \) with \( r \). (See Figure [1] left.)

Because of symmetry, we can rotate the triangle so that the chord \( AB \) is perpendicular to \( r \). Since the sides of the triangle intersect the perpendicular radii at distance \( 1/2 \) from 0, \( M \) is longer than \( AB \) if and only if \( x \) is at distance at most \( 1/2 \) from 0.

\( r \) has length 1, so the probability that \( x \) is at distance at most \( 1/2 \) is \( 1/2 \). \( \square \)

Solution 2. Another way: Choose two random points \( x, y \) on the circle, and let \( M \) be the chord connecting them. (See Figure [1] right.)

Because of symmetry, we can rotate the triangle so that \( x \) coincides with the vertex \( A \) of the triangle. So we can see that \( y \) falls in the arc \( \overarc{BC} \) on the circle if and only if \( M \) is longer than a side of the triangle.
The probability of this is the length of the arc $\widehat{BC}$ over $2\pi$. That is, $1/3$ (the arc $\widehat{BC}$ is one-third of the circle).

Solution 3. A different way to choose a random chord $M$: Choose a random point $x$ in the circle, and let $r$ be the radius through $x$. Then choose $M$ to be the chord through $x$ perpendicular to $r$. (See Figure 1 on the left.)

Again we can rotate the triangle so that $r$ is perpendicular to the chord $AB$.

Then, $M$ will be longer than $AB$ if and only if $x$ lands inside the triangle; that is if and only if the distance of $x$ to 0 is at most $1/2$.

Since the area of a disc of radius $1/2$ is $1/4$ of the disc of radius 1, this happens with probability $1/4$.

How did we reach this paradox?

The original question was ill posed. We did not define precisely what a random chord is.

The different solutions come from different ways of choosing a random chord - these are not the same.

We now turn to precisely defining what we mean by “random”, “probability”, etc.
1.2. Sample spaces

When we do some experiment, we first want to collect all possible outcomes of the experiment. In mathematics, a collection of objects is just a set.

The set of all possible outcomes of an experiment is called a sample space. Usually we denote the sample space by $\Omega$, and its elements by $\omega$.

**Example 1.2.**

- A coin toss. $\Omega = \{H, T\}$. Actually, it is the set of heads and tails, maybe $\{\circ, \oplus\}$, but $H, T$ are easier to write.
- Tossing a coin twice. $\Omega = \{H, T\}^2$. What about tossing a coin $k$ times?
- Throwing two dice. $\Omega = \{\Box, \bigcirc, \heartsuit, \clubsuit, \spadesuit, \diamondsuit\}^2$. It is probably easier to use $\{1, 2, 3, 4, 5, 6\}^2$.
- The lifetime of a person. $\Omega = \mathbb{R}^+$. What if we only count the years? What if we want years and months?
- Bows and arrows: Shooting an arrow into a target of radius $1/2$ meters. $\Omega = \{(x, y) : x^2 + y^2 \leq 1/4\}$. What about $\Omega = \{(r, \theta) : r \in [0, 1/2], \theta \in [0, 2\pi)\}$. What if the arrow tip has thickness of $1$cm? So we don’t care about the point up to radius $1$cm? What about missing the target altogether?
- A random real valued continuously differentiable function on $[0, 1]$. $\Omega = C^1([0, 1])$.
- Noga tosses a coin. If it comes out heads she goes to the probability lecture, and either takes notes or sleeps throughout. If the coin comes out tails, she goes running, and she records the distance she ran in meters and the time it took. $\Omega = \{H\} \times \{\text{notes, sleep}\} \cup \{T\} \times \mathbb{N} \times \mathbb{R}^+$.

1.3. Events

Suppose we toss a die. So the sample space is, say, $\Omega = \{1, 2, 3, 4, 5, 6\}$. We want to encode the outcome “the number on the die is even”.
This can be done by collecting together all the relevant possible outcomes to this event. That is, a sub-collection of outcomes, or a subset of $\Omega$. The event “the number on the die is even” corresponds to the subset $\{2, 4, 6\} \subset \Omega$.

What do we want to require from events? We want the following properties:

- "Everything" is an event; that is, $\Omega$ is an event.
- If we can ask if an event occurred, then we can also ask if it did not occur; that is, if $A$ is an event, then so is $A^c$.
- If we have many events of which we can ask if they have occurred, then we can also ask if one of them has occurred; that is, if $(A_n)_{n\in\mathbb{N}}$ is a sequence of events, then so is $\bigcup_n A_n$.

✓ A word on notation: $A^c = \Omega \setminus A$. One needs to be careful in which space we are taking the complement. Some books use $\overline{A}$.

Thus, if we want to say what events are, we have: If $\mathcal{F}$ is the collection of events on a sample space $\Omega$, then $\mathcal{F}$ has the following properties:

- The elements of $\mathcal{F}$ are subsets of $\Omega$ (i.e. $\mathcal{F} \subset \mathcal{P}(\Omega)$).
- $\Omega \in \mathcal{F}$.
- If $A \in \mathcal{F}$ then $A^c \in \mathcal{F}$.
- If $(A_n)_{n\in\mathbb{N}}$ is a sequence of elements of $\mathcal{F}$, then $\bigcup_n A_n \in \mathcal{F}$.

**Definition 1.3.** $\mathcal{F}$ with the properties above is called a $\sigma$-algebra (or $\sigma$-field).

✓ Explain the name: algebra, $\sigma$-algebra.

**Example 1.4.** Let $\Omega$ be any set (sample space). Then,

- $\mathcal{F} = \{\emptyset, \Omega\}$
- $\mathcal{G} = 2^\Omega = \mathcal{P}(\Omega)$

are both $\sigma$-algebras on $\Omega$.

When $\Omega$ is a countable sample space, then we will always take the full $\sigma$-algebra $2^\Omega$. (We will worry about the uncountable case in the future.)
To sum up: For a countable sample space $\Omega$, any subset $A \subset \Omega$ is an event.

**Example 1.5.**

- The experiment is tossing three coins. The event $A$ is “second toss was heads”. $\Omega = \{H, T\}^3$. $A = \{(x, H, y) : x, y \in \{H, T\}\}$.
- The experiment is “how many minutes passed until a goal is scored in the Manchester derby”. The event $A$ is “the first goal was scored after minute 23 and before minute 45”. $\Omega = \{0, 1, \ldots, 90\}$ (maybe extra-time?). $A = \{23, 24, \ldots, 44\}$.

**Example 1.6.** We are given a computer code of 4 letters. $A_i$ is the event that the $i$-th letter is ‘L’.

What are the following events?

- $A_1 \cap A_2 \cap A_3 \cap A_4$
- $A_1 \cup A_2 \cup A_3 \cup A_4$
- $A_1^c \cap A_2^c \cap A_3^c \cap A_4^c$
- $A_1^c \cup A_2^c \cup A_3^c \cup A_4^c$
- $A_3 \cap A_4$
- $A_1 \cap A_2^c$

What are the events in terms of $A_i$’s?

- There are at least 3 L’s
- There is exactly one L
- There are no two L’s in a row

**Example 1.7.** Every day products on a production line are chosen and labeled ‘-’ if damaged or ‘+’ if good. What are the complements of the following events?

- $A = \text{at least two products are damaged}$
- $B = \text{at most 3 products are damaged}$
\begin{itemize}
\item $C =$ there are at least 3 more good products than damaged products
\item $D =$ there are no damaged products
\item $E =$ most of the products are damaged
\end{itemize}

To which of the above events does the string $---+++---$ belong to?

What are the events $A \cap B, A \cup B, C \cap E, B \cap D, B \cup D$. \hfill \triangle \triangledown \triangle
2.1. Probability

Remark 2.1. Two sets \(A, B\) are called disjoint if \(A \cap B = \emptyset\). For a sequence \((A_n)_{n \in \mathbb{N}}\), we say that \((A_n)\) are mutually disjoint if any two are disjoint; i.e. for any \(n \neq m\), \(A_n \cap A_m = \emptyset\).

What is probability? It is the assignment of a value to each event, between 0 and 1, saying how likely this event is. That is, if \(\mathcal{F}\) is the collection of all events on some sample space, then probability is a function from the events to \([0, 1]\); i.e. \(P : \mathcal{F} \to [0, 1]\).

Of course, not every such assignment is good, and we would like some consistency.

What would we like of \(P\)? For one thing, we would like that the likelyhood of everything is 100%. That is, \(P(\Omega) = 1\). Also, we would like that if \(A, B \in \mathcal{F}\) are two events that are disjoint, i.e. \(A \cap B = \emptyset\), then the likelyhood that one of \(A, B\) occurs is the sum of their likelyhoods; i.e. \(P(A \cup B) = P(A) + P(B)\).

This leads to the following definition:

**Definition 2.2** (Probability measure). Let \(\Omega\) be a sample space, and \(\mathcal{F}\) a \(\sigma\)-algebra on \(\Omega\). A probability measure on \((\Omega, \mathcal{F})\) is a function \(P : \mathcal{F} \to [0, 1]\) such that

- \(P(\Omega) = 1\).
- If \((A_n)\) is a sequence of mutually disjoint events in \(\mathcal{F}\), then
  \[P\left(\bigcup_{n} A_n\right) = \sum_{n} P(A_n).\]

\((\Omega, \mathcal{F}, P)\) is called a probability space.

**Example 2.3.**
• A fair coin toss: \( \Omega = \{H, T\} \), \( \mathcal{F} = \mathcal{P}(\Omega) = \{\emptyset, \{H\}, \{T\}, \Omega\} \). Then \( P(\emptyset) = 0, P(\{H\}) = \frac{1}{2}, P(\{T\}) = \frac{1}{2}, P(\Omega) = 1 \).

• Unfair coin toss: \( P(\emptyset) = 0, P(\{H\}) = \frac{3}{4}, P(\{T\}) = \frac{1}{4}, P(\Omega) = 1 \).

• Fair die: \( \Omega = \{1, 2, \ldots, 6\} \), \( \mathcal{F} = \mathcal{P}(\Omega) \). For \( A \subset \Omega \) let \( P(A) = \frac{|A|}{6} \).

• Let \( \Omega \) be any countable set. Let \( \mathcal{F} = \mathcal{P}(\Omega) \). Let \( p : \Omega \to [0, 1] \) be a function such that \( \sum_{\omega \in \Omega} p(\omega) = 1 \). Then, for any \( A \subset \Omega \) define
  \[
P(A) = \sum_{\omega \in A} p(\omega).
  \]
  \( P \) is a probability measure on \((\Omega, \mathcal{F})\).

• Let \( \gamma = \sum_{n \geq 1} n^{-2} \). For \( A \subset \{1, 2, \ldots, \} \) define
  \[
P(A) = \frac{1}{\gamma} \sum_{a \in A} a^{-2}.
  \]
  \( P \) is a probability measure.

Some properties of probability measures:

**Proposition 2.4.** Let \((\Omega, \mathcal{F}, P)\) be a probability space. Then,

1. \( \emptyset \in \mathcal{F} \) and \( P(\emptyset) = 0 \).
2. If \( A_1, \ldots, A_n \) are a finite number of mutually disjoint events, then
   \[
   A := \bigcup_{j=1}^{n} A_j \in \mathcal{F} \quad \text{and} \quad P(A) = \sum_{j=1}^{n} P(A_j).
   \]
3. For any event \( A \in \mathcal{F} \), \( P(A^c) = 1 - P(A) \).
4. If \( A \subset B \) are events, then \( P(B \setminus A) = P(B) - P(A) \).
5. If \( A \subset B \) are events, then \( P(A) \leq P(B) \).
6. (Inclusion-Exclusion Principle.) For events \( A, B \in \mathcal{F} \),
   \[
P(A \cup B) = P(A) + P(B) - P(A \cap B).
   \]

**Proof.** Let \( A, B, (A_n)_{n \in \mathbb{N}} \) be events in \( \mathcal{F} \).
(1) $\emptyset = \Omega^c \in \mathcal{F}$. Set $A_j = \emptyset$ for all $j \in \mathbb{N}$. This is a sequence of mutually disjoint events. Since $\emptyset = \bigcup_j A_j$, we have that $\mathbb{P}(\emptyset) = \sum_j \mathbb{P}(\emptyset)$ implying that $\mathbb{P}(\emptyset) = 0$.

(2) For $j > n$ let $A_j = \emptyset$. Then $(A_j)_{j \in \mathbb{N}}$ is a sequence of mutually disjoint events, so

$$
\mathbb{P}\left(\bigcup_{j=1}^n A_j\right) = \mathbb{P}\left(\bigcup_{j} A_j\right) = \sum_{j} \mathbb{P}(A_j) = \sum_{j=1}^n \mathbb{P}(A_j).
$$

(3) $A \cup A^c = \Omega$ and $A \cap A^c = \emptyset$. So the additivity of $\mathbb{P}$ on disjoint sets gives that

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

(4) If $A \subset B$, then $B = (B \setminus A) \cup A$, and $(B \setminus A) \cap A = \emptyset$. Thus, $\mathbb{P}(B) = \mathbb{P}(B \setminus A) + \mathbb{P}(A)$.

(5) Since $A \subset B$, we have that $\mathbb{P}(B) - \mathbb{P}(A) = \mathbb{P}(B \setminus A) \geq 0$.

(6) Let $C = A \cap B$. Note that $A \cup B = (A \setminus C) \cup (B \setminus C) \cup C$, where all sets in the union are mutually disjoint. Since $C \subset A, C \subset B$ we get

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

Proposition 2.5 (Boole’s inequality). Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. If $(A_n)_{n \in \mathbb{N}}$ are events, then

$$
\mathbb{P}\left(\bigcup_{n \in \mathbb{N}} A_n\right) \leq \sum_{n \in \mathbb{N}} \mathbb{P}(A_n).
$$

Proof. For every $n$ let $B_0 = \emptyset$, and

$$
B_n = \bigcup_{j=1}^n A_j \quad \text{and} \quad C_n = A_n \setminus B_{n-1}.
$$

Claim 1. For any $n > m \geq 1$, we have that $C_n \cap C_m = \emptyset$.

Proof. Since $m \leq n - 1$, we have that $A_m \subset B_{n-1}$. Since $C_m \subset A_m$,

$$
C_n \cap C_m \subset (A_n \setminus B_{n-1}) \cap B_{n-1} = \emptyset.
$$

Claim 2. $\bigcup_n A_n \subset \bigcup_n C_n$.

Proof. Let $\omega \in \bigcup_n A_n$. So there exists $n$ such that $\omega \in A_n$. Let $m$ be the smallest integer such that $\omega \in A_m$ and $\omega \notin A_{m-1}$ (if $\omega \in A_1$ let $m = 1$). Then, $\omega \in A_m$, and
\( \omega \notin A_k \) for all \( 1 \leq k < m \). In other words: \( \omega \in A_m \) and \( \omega \notin \bigcup_{k=1}^{m-1} A_k = B_{m-1} \). Or: \( \omega \in A_m \setminus B_{m-1} = C_m \). Thus, there exists some \( m \geq 1 \) such that \( \omega \in C_m \); i.e. \( \omega \in \bigcup_n C_n \).

We conclude that \( (C_n)_{n \geq 1} \) is a collection of events that are pairwise disjoint, and \( \bigcup_n A_n \subset \bigcup_n C_n \). Using \# 5 from Proposition 2.4

\[ P\left( \bigcup_n A_n \right) \leq P\left( \bigcup_n C_n \right) = \sum_n P(C_n) \leq \sum_n P(A_n). \]

\( \square \)

**Example 2.6** (Monty Hall paradox). In a famous gameshow, the contestant is given three boxes, two of them contain nothing, but one contains the keys to a new car. All possibilities of key placements are possible. The contestant chooses a box. Then the host reveals one of the other two boxes that does not contain anything, and the contestant is given a choice whether to switch her choice or not.

What should she do?

\( A = \{ \text{an empty box is chosen in the first choice} \} \). \( B = \{ \text{the keys are in the other box (she should switch)} \} \).

One checks that \( B \cap A^c = \emptyset \) and \( A \cap B^c = \emptyset \) so \( A = B \). Since all key placements are equally likely, the probability the key is in the chosen box is \( 1/3 \). That is, \( P(A^c) = 1/3 \), and so \( P(B) = P(A) = 2/3 \). \( \triangle \nabla \triangle \)

### 2.2. Discrete Probability Spaces

Discrete probability spaces are those for which the sample space is countable. We have already seen that in this case we can take all subsets to be events, so \( F = 2^\Omega \). We have also implicitly seen that due to additivity on disjoint sets, the probability measure \( P \) is completely determined by its value on singletons.

That is, let \( (\Omega, 2^\Omega, P) \) be a probability space where \( \Omega \) is countable. If we denote \( p(\omega) = P(\{\omega\}) \) for all \( \omega \in \Omega \), then for any event \( A \subset \Omega \) we have that \( A = \bigcup_{\omega \in A} \{\omega\} \), and this is a countable disjoint union. Thus,

\[ P(A) = \sum_{\omega \in A} P(\{\omega\}) = \sum_{\omega \in A} p(\omega). \]
**Exercise 2.7.** Let $\Omega = \{1, 2, \ldots, \}$ and define a probability measure on $(\Omega, 2^\Omega)$ by

1. $P(\{\omega\}) = \frac{1}{(e-1)\omega!}.$
2. $P'(\{\omega\}) = \frac{1}{3} \cdot (3/4)^\omega.$

(Extend using additivity on disjoint unions.) Show that both uniquely define probability measures.

**Solution.** We need to show that $P(\Omega) = 1$, and that $P$ is additive on disjoint unions.

Indeed,

$$P(\Omega) = \sum_{\omega=1}^{\infty} P(\{\omega\}) = \sum_{\omega=1}^{\infty} \frac{1}{(e-1)\omega!} = 1.$$ 

$$P'(\Omega) = \frac{1}{3} \cdot \sum_{\omega=1}^{\infty} (3/4)^\omega = 1.$$ 

Now let $(A_n)$ be a sequence of mutually disjoint events and let $A = \bigcup_n A_n$. Using the fact that any subset is the disjoint union of the singlet composing it, we have that $\omega \in A$ if and only if there exists a unique $n(\omega)$ such that $\omega \in A_n$. Thus,

$$P(A) = \sum_{\omega \in A} P(\{\omega\}) = \sum_n \sum_{\omega \in A_n} P(\{\omega\}) = \sum_n P(A_n).$$

The same for $P'$.

The next proposition generalizes the above examples, and characterizes all discrete probability spaces.

**Proposition 2.8.** Let $\Omega$ be a countable set. Let $p : \Omega \to [0, 1]$, such that $\sum_{\omega \in \Omega} p(\omega) = 1$. Then, there exists a unique probability measure $P$ on $(\Omega, 2^\Omega)$ such that $P(\{\omega\}) = p(\omega)$ for all $\omega \in \Omega$. ($p$ as above is sometimes called the density of $P$.)

**Proof.** Let $A \subset \Omega$. Define

$$P(A) = \sum_{\omega \in A} p(\omega).$$

We have by the assumption on $p$ that $P(\Omega) = \sum_{\omega \in \Omega} p(\omega) = 1$. Also, if $(A_n)$ is a sequence of events that are mutually disjoint, and $A = \bigcup_n A_n$ is their union, then, for any $\omega \in A$
there exists $n$ such that $\omega \in A_n$. Moreover, this $n$ is unique, since $A_n \cap A_m = \emptyset$ for all $m \neq n$, so $\omega \not\in A_m$ for all $m \neq n$. So

$$P(A) = \sum_{\omega \in A} p(\omega) = \sum_{n} \sum_{\omega \in A_n} p(\omega) = \sum_{n} P(A_n).$$

This shows that $P$ is a probability measure on $(\Omega, 2^\Omega)$.

For uniqueness, let $P : 2^\Omega \to [0, 1]$ be a probability measure such that $P(\{\omega\}) = p(\omega)$ for all $\omega \in \Omega$. Then, for any $A \subset \Omega$,

$$P(A) = \sum_{\omega \in A} P(\{\omega\}) = \sum_{\omega \in A} p(\omega) = P(A).$$

\[\square\]

**Example 2.9.**

(1) An simple but important example is for finite sample spaces. Let $\Omega$ be a finite set. Suppose we assume that all outcomes are equally likely. That is $P(\{\omega\}) = 1/|\Omega|$ for all $\omega \in \Omega$, and so, for any event $A \subset \Omega$ we have that $P(A) = |A|/|\Omega|$.

It is simple to verify that this is a probability measure. This is known as the *uniform measure* on $\Omega$.

(2) We throw two fair dice. What is the probability the sum is 7? What is the probability the sum is 6?

**Solution.** The sample space here is $\Omega = \{1, 2, \ldots, 6\}^2 = \{(i, j) : 1 \leq i, j \leq 6\}$.

Since we assume the dice are fair, all outcomes are equally likely, and so the probability measure is the uniform measure on $\Omega$.

Now, the event that the sum is 7 is

$$A = \{(i, j) : 1 \leq i, j \leq 6, i + j = 7\} = \{(1, 6), (2, 5), (3, 4), (4, 3), (5, 2), (6, 1)\}.$$

So $P(A) = |A|/|\Omega| = 6/36 = 1/6$.

The event that the sum is 6 is

$$B = \{(i, j) : 1 \leq i, j \leq 6, i + j = 6\} = \{(1, 5), (2, 4), (3, 3), (4, 2), (5, 1)\}.$$

So $P(B) = 5/36$. 
(3) There are 15 balls in a jar, 7 black balls and 8 white balls. When removing a ball from the jar, any ball is equally likely to be removed. Shir puts her hand in the jar and takes out two balls, one after the other. What is the probability Shir removes one black ball and one white ball.

Solution. First, we can think of the different balls being represented by the numbers 1, 2, ..., 15. So the sample space is \( \Omega = \{(i, j) : 1 \leq i \neq j \leq 15\} \). Note that the size of \( \Omega \) is \( |\Omega| = 15 \cdot 14 \). Since it is equally likely to remove any ball, we have that the probability measure is the uniform measure on \( \Omega \).

Let \( A \) be the event that one ball is black and one is white. How can we compute \( P(A) = \frac{|A|}{|\Omega|} \)?

We can split \( A \) into two disjoint events, and use additivity of probability measures:

Let \( B \) be the event that the first ball is white and the second ball is black. Let \( B' \) be the event that the first ball is black and the second ball is white. It is immediate that \( B \) and \( B' \) are disjoint. Also, \( A = B \cup B' \). Thus, \( P(A) = P(B) + P(B') \), and we only need to compute the sizes of \( B, B' \).

Note that the number of pairs \((i, j)\) such that ball \( i \) is black and ball \( j \) is white is \( 7 \cdot 8 \). So \( |B| = 7 \cdot 8 \). Similarly, \( |B'| = 8 \cdot 7 \). All in all,

\[
P(A) = P(B) + P(B') = \frac{7 \cdot 8}{15 \cdot 14} + \frac{8 \cdot 7}{15 \cdot 14} = \frac{8}{15}.
\]

(4) A deck of 52 cards is shuffled, so that any order is equally likely. The top 10 cards are distributed among 2 players, 5 to each one. What is the probability that at least one player has a royal flush (10-J-Q-K-A of the same suit)?

Solution. Each player receives a subset of the cards of size 5, so the sample space is

\[
\Omega = \{(S_1, S_2) : S_1, S_2 \subseteq C, |S_1| = |S_2| = 5, S_1 \cap S_2 = \emptyset\},
\]

where \( C \) is the set of all cards \( \{A\spadesuit, 2\spadesuit, \ldots, A\clubsuit, 2\clubsuit, \ldots\} \). There are \( \binom{52}{5} \) ways to choose \( S_1 \) and \( \binom{47}{5} \) ways to choose \( S_2 \), so \( |\Omega| = \binom{52}{5} \cdot \binom{47}{5} \). Also, every choice is equally likely, so \( P \) is the uniform measure.
Let $A_i$ be the event that player $i$ has a royal flush ($i = 1, 2$). For $s \in \{\spadesuit, \diamondsuit, \heartsuit, \clubsuit\}$, let $B(i, s)$ be the event that player $i$ has a royal flush of the suit $s$. So $A_i = \bigcup_s B(i, s)$. $B(i, s)$ is the event that $S_i$ is a specific set of 5 cards, so $|B(i, s)| = \binom{47}{5}$ for any choice of $i, s$. Thus,

$$P(A_i) = \sum_s \frac{|B(i, s)|}{|\Omega|} = \frac{4}{\binom{52}{5}}.$$

Now we use the inclusion-exclusion principle:

$$P(A) = P(A_1 \cup A_2) = P(A_1) + P(A_2) - P(A_1 \cap A_2).$$

The event $A_1 \cap A_2$ is the event that both players have a royal flush, so $|A_1 \cap A_2| = 4 \cdot 3$ since there are 4 options for the first player’s suit, and then 3 for the second. Altogether,

$$P(A) = \frac{4}{\binom{52}{5}} + \frac{4}{\binom{52}{5}} - \frac{4 \cdot 3}{\binom{52}{5} \cdot \binom{47}{5}} = \frac{8}{\binom{52}{5}} \left(1 - \frac{3}{2 \cdot \binom{47}{5}}\right).$$

△▽△

**Exercise 2.10.** Prove that there is no uniform probability measure on $\mathbb{N}$; that is, there is no probability measure on $\mathbb{N}$ such that $P(\{i\}) = P(\{j\})$ for all $i, j$.

**Solution.** Assume such a probability measure exists. By the defining properties of probability measures,

$$1 = P(\mathbb{N}) = \sum_{j=0}^{\infty} P(\{j\}) = \sum_{j=1}^{\infty} p = \left\{\begin{array}{ll} \infty & p > 0 \\ 0 & p = 0 \end{array}\right..$$

A contradiction. □

**Exercise 2.11.** A deck of 52 card is ordered randomly, all orders equally likely. what is the probability that the 14th card is an ace? What is the probability that the first ace is at the 14th card?

**Solution.** The sample space is all the possible orderings of the set of cards $C = \{A\spadesuit, 2\spadesuit, \ldots, A\clubsuit, 2\clubsuit, \ldots\}$. So $\Omega$ is the set of all 1-1 functions $f : \{1, \ldots, 52\} \to C$, where $f(1)$ is the first card, $f(2)$ the second, and so on. So $|\Omega| = 52!$. The measure is the uniform one.
Let $A$ be the event that the 14th card is an ace. Let $B$ be the event that the first ace is at the 14th card.

$A$ is the event that $f(14)$ is an ace, and there are 4 choices for this ace, so if $\mathcal{A}$ is the set of aces, $A = \{f \in \Omega : f(14) \in \mathcal{A}\}$, so $|A| = 4 \cdot 51!$. Thus, $\mathbb{P}(A) = \frac{4}{52} = \frac{1}{13}$.

$B$ is the event that $f(14)$ is an ace, but also $f(j)$ is not an ace for all $j < 14$. Thus, $B = \{f \in \Omega : f(14) \in \mathcal{A}, \forall j < 14 \ f(j) \not\in \mathcal{A}\}$. So $|B| = 4 \cdot 48 \cdot 47 \cdots 36 \cdot 38! = 4 \cdot 48! \cdot 38 \cdot 37 \cdot 36$, and $\mathbb{P}(B) = \frac{4 \cdot 38 \cdot 37 \cdot 36}{52 \cdot 51 \cdot 50 \cdot 49} = 0.031160772$. \qed
3.1. SOME SET THEORY

Recall the inclusion-exclusion principle:

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

This can be demonstrated by the Venn diagram in Figure 2.

This diagram also illustrates the intuitive meaning of \( A \cup B \); namely, \( A \cup B \) is the event that one of \( A \) or \( B \) occurs. Similarly, \( A \cap B \) is the event that both \( A \) and \( B \) occur.

Let \((A_n)_{n=1}^{\infty}\) be a sequence of events in some probability space. We have

\[ \bigcup_{n \geq k} A_n = \{ \omega \in \Omega : \omega \in A_n \text{ for at least one } n \geq k \} \]

\[ = \{ \omega \in \Omega : \text{there exists } n \geq k \text{ such that } \omega \in A_n \}. \]

So \( \bigcap_{k=1}^{\infty} \bigcup_{n \geq k} A_n \) is the set of \( \omega \in \Omega \) such that for any \( k \), there exists \( n \geq k \) such that \( \omega \in A_n \); i.e.

\[ \bigcup_{k=1}^{\infty} \bigcap_{n \geq k} A_n = \{ \omega \in \Omega : \omega \in A_n \text{ for infinitely many } n \}. \]
**Definition 3.1.** Define the set $\limsup A_n$ as

$$
\limsup A_n := \bigcap_{k=1}^{\infty} \bigcup_{n \geq k} A_n.
$$

The intuitive meaning of $\limsup A_n$ is that infinitely many of the $A_n$’s occur.

Similarly, we have that

$$
\bigcup_{k=1}^{\infty} \bigcap_{n \geq k} A_n = \{\omega \in \Omega : \exists n_0 \text{ such that } \forall n > n_0, \omega \in A_n\}
= \{\omega \in \Omega : \omega \in A_n \text{ for all large enough } n\}.
$$

**Definition 3.2.** Define

$$
\liminf A_n := \bigcup_{k=1}^{\infty} \bigcap_{n \geq k} A_n.
$$

$\liminf A_n$ means that $A_n$ will occur from some large enough $n$ onward; that is, eventually all $A_n$ occur.

It is easy to see that

$$
\liminf A_n \subseteq \limsup A_n.
$$

This also fits the intuition, as if all $A_n$ occur eventually, then they occur infinitely many times.

**Definition 3.3.** If $\left( A_n \right)_{n=1}^{\infty}$ is a sequence of events such that $\liminf A_n = \limsup A_n$ (as sets) then we define

$$
\lim A_n := \liminf A_n = \limsup A_n.
$$

**Example 3.4.** Consider $\Omega = \mathbb{N}$, and the sequence

$$
A_n = \{n^j : j = 0, 1, 2, \ldots\}.
$$

If $m < n$ and $m \in A_n$, then it must be that $m = n^0 = 1$. Thus, $\bigcap_{n \geq k} A_n = \{1\}$ and

$$
\liminf_n A_n = \bigcup_{k=1}^{\infty} \bigcap_{n \geq k} A_n = \{1\}.
$$
Also, if \( m < k \leq n \), and \( m \in A_n \), then again \( m = 1 \), so \( \bigcup_{n \geq k} A_n \) does not contain any \( 1 < m < k \); i.e. \( \bigcup_{n \geq k} A_n \subset \{1, k, k + 1, \ldots\} \). Hence,
\[
\limsup A_n = \bigcap_{k=1}^{\infty} \bigcup_{n \geq k} A_n = \{1\}.
\]
Thus, the limit exists and \( \lim_n A_n = \{1\} \).

**Definition 3.5.** A sequence of events is called increasing if \( A_n \subset A_{n+1} \) for all \( n \), and decreasing if \( A_{n+1} \subset A_n \) for all \( n \).

**Proposition 3.6.** Let \( (A_n) \) be a sequence of events. Then,
\[
(\liminf A_n)^c = \limsup A_n^c.
\]
Moreover, if \( (A_n) \) is increasing (resp. decreasing) then \( \lim A_n = \bigcup_n A_n \) (resp. \( \lim A_n = \bigcap_n A_n \)).

**Proof.** The first assertion is de-Morgan.

For the second assertion, note that if \( (A_n) \) is increasing, then
\[
\bigcup_{n \geq k} A_n = \bigcup_{n \geq 1} A_n \quad \text{and} \quad \bigcap_{n \geq k} A_n = A_k.
\]
So
\[
\limsup A_n = \bigcap_k \bigcup_{n \geq k} A_n = \bigcap_k \bigcup_n A_n = \bigcup_n \bigcap_n A_n = A_k.
\]

If \( (A_n) \) is decreasing then \( (A_n^c) \) is increasing. So
\[
\limsup A_n = (\liminf A_n^c)^c = (\bigcup_n A_n^c)^c = (\limsup A_n^c)^c = \liminf A_n,
\]
and \( \lim A_n = (\bigcup_n A_n^c)^c = \bigcap_n A_n \).

**Exercise 3.7.** Show that if \( \mathcal{F} \) is a \( \sigma \)-algebra, and \( (A_n) \) is a sequence of events in \( \mathcal{F} \), then \( \liminf A_n \in \mathcal{F} \) and \( \limsup A_n \in \mathcal{F} \).
3.2. CONTINUITY OF PROBABILITY MEASURES

The goal of this section is to prove

**Theorem 3.8** (Continuity of Probability). Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. Let $(A_n)$ be a sequence of events such that $\lim A_n$ exists. Then,

$$\mathbb{P}(\lim A_n) = \lim_{n \to \infty} \mathbb{P}(A_n).$$

We start with a restricted version:

**Proposition 3.9.** Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. Let $(A_n)$ be a sequence of increasing (resp. decreasing) events. Then,

$$\mathbb{P}(\lim A_n) = \lim_{n \to \infty} \mathbb{P}(A_n).$$

*Proof.* We start with the increasing case. Let $A = \bigcup_n A_n = \lim A_n$. Define $A_0 = \emptyset$, and $B_k = A_k \setminus A_{k-1}$ for all $k \geq 1$. So $(B_k)$ are mutually disjoint and

$$\bigcup_{k=1}^n B_k = \bigcup_{k=1}^n A_k = A_n$$

so $\bigcup_n B_n = A$. Thus,

$$\mathbb{P}(\lim A_n) = \mathbb{P}\left(\bigcup_n B_n\right) = \sum_n \mathbb{P}(B_n) = \lim_{n} \sum_{k=1}^n \mathbb{P}(B_k)$$

$$= \lim_{n} \mathbb{P}\left(\bigcup_{k=1}^n B_k\right) = \lim_{n} \mathbb{P}(A_n).$$

The decreasing case follows from noting that if $(A_n)$ is decreasing, then $(A_n^c)$ is increasing, so

$$\mathbb{P}(\lim A_n) = \mathbb{P}\left(\bigcap_n A_n\right) = 1 - \mathbb{P}\left(\bigcup_n A_n^c\right) = 1 - \lim_{n} \mathbb{P}(A_n^c) = \lim_{n} \mathbb{P}(A_n).$$

□

**Lemma 3.10** (Fatou’s Lemma). Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. Let $(A_n)$ be a sequence of events (that may not have a limit). Then,

$$\mathbb{P}(\lim \inf A_n) \leq \lim \inf \mathbb{P}(A_n) \quad \text{and} \quad \lim \sup \mathbb{P}(A_n) \leq \mathbb{P}(\lim \sup A_n).$$
Proof. For all $k$ let $B_k = \cap_{n \geq k} A_n$. So $\liminf_n A_n = \bigcup_k B_k$. Note that $(B_k)$ is an increasing sequence of events ($B_k = B_{k+1} \cap A_k$). Also, for any $n \geq k$, $B_k \subset A_n$, so $P(B_k) \leq \inf_{n \geq k} P(A_n)$. Thus,

$$P(\liminf_n A_n) = P\left(\bigcup_{n} B_n\right) = \lim_{n} P(B_n) \leq \lim_{k \geq n} P(A_k) = \liminf_{n} P(A_n).$$

For the lim sup,

$$\limsup_{n} P(A_n) = \limsup_{n} (1 - P(A_n^c)) = 1 - \liminf_{n} P(A_n^c)$$

$$\leq 1 - P(\liminf_n A_n^c) = 1 - P((\limsup_n A_n)^c) = P(\limsup_n A_n).$$

Fatou’s Lemma immediately proves Theorem 3.8.

Proof of Theorem 3.8. Just note that

$$\limsup_{n} P(A_n) \leq P(\limsup_n A_n) = P(\lim A_n) = P(\liminf_n A_n) \leq \liminf_{n} P(A_n),$$

so equality holds, and the limit exists.

\[\square\]

As a consequence we get

**Lemma 3.11** (First Borel-Cantelli Lemma). Let $(\Omega, \mathcal{F}, P)$ be a probability space. Let $(A_n)$ be a sequence of events. If $\sum_n P(A_n) < \infty$, then

$$P( A_n \text{ occurs for infinitely many } n ) = P(\limsup_n A_n) = 0.$$

Proof. Let $B_k = \bigcup_{n \geq k} A_n$. So $(B_k)$ is decreasing, and so the decreasing sequence $P(B_k)$ converges to $P(\limsup A_n)$. Thus, for all $k$,

$$P(\limsup A_n) \leq P(B_k) \leq \sum_{n \geq k} P(A_n).$$

Since the right hand side converges to 0 as $k \to \infty$ by the assumption that the series is convergent, we get $P(\limsup A_n) \leq 0$.

\[\square\]
**Example 3.12.** We have a bunch of bacteria in a petri dish. Every second, the bacteria give off offspring randomly, and then the parents die out. Suppose that for any \( n \), the probability that there are no bacteria left by time \( n \) is \( 1 - \exp(-f(n)) \).

What is the probability that the bacteria eventually die out if:

- \( f(n) = \log n \).
- \( f(n) = \frac{n^2 - 7}{2n^2 + 3n + 5} \).

Let \( A_n \) be the event that the bacteria dies out by time \( n \). So \( \mathbb{P}(A_n) = 1 - e^{-f(n)} \), for \( n \geq 1 \).

Note that the event that the bacteria eventually die out, is the event that there exists \( n \) such that \( A_n \); i.e. the event \( \bigcup_n A_n \). Since \( (A_n)_n \) is an increasing sequence, we have that \( \mathbb{P}(\bigcup_n A_n) = \lim_n \mathbb{P}(A_n) \).

In the first case this is 1. In the second case this is

\[
\lim_{n \to \infty} 1 - \exp \left( -\frac{n^2 - 7}{2n^2 + 3n + 5} \right) = 1 - e^{-1/2}.
\]

\( \triangle \nabla \triangle \)

**Example 3.13.** What if we take the previous example, and the information is that the probability that the bacteria at generation \( n \) do not die out without offspring is at most \( \exp(-2\log n) \).

Then, if \( A_n \) is the event that the \( n \)-th generation has offspring, we have that \( \mathbb{P}(A_n) \leq n^{-2} \). Since \( \sum_n \mathbb{P}(A_n) < \infty \), Borel-Cantelli tells us that

\[
\mathbb{P}(A_n \text{ occurs for infinitely many } n) = \mathbb{P}(\lim \sup A_n) = 0.
\]

That is,

\[
\mathbb{P}\left( \exists k \; : \; \forall n \geq k \; : \; A_n^c \right) = \mathbb{P}(\lim \inf A_n^c) = 1.
\]

So a.s. there exists \( k \) such that all generations \( n \geq k \) do not have offspring – implying that the bacteria die out with probability 1.

\( \triangle \nabla \triangle \)
4.1. Conditional Probabilities

We start with a simple observation.

**Proposition 4.1.** Let \((\Omega, \mathcal{F}, \mathbb{P})\) be a probability space. Let \(F \in \mathcal{F}\) be an event such that \(\mathbb{P}(F) > 0\). Define

\[
\mathcal{F}|_F = \{ A \cap F : A \in \mathcal{F} \},
\]

and define \(P : \mathcal{F}|_F \to [0, 1]\) by \(P(B) = \mathbb{P}(B)/\mathbb{P}(F)\) for all \(B \in \mathcal{F}|_F\), and \(Q : \mathcal{F} \to [0, 1]\) by \(Q(A) = \mathbb{P}(A \cap F)/\mathbb{P}(F)\) for all \(A \in \mathcal{F}\). Then, \(\mathcal{F}|_F\) is a \(\sigma\)-algebra on \(F\), and \(P\) is a probability measure on \((\mathcal{F}, \mathcal{F}|_F)\), and \(Q\) is a probability measure on \(\Omega, \mathcal{F}\).

**Notation.** The probability measure \(Q\) above is usually denoted \(\mathbb{P}(\cdot|F)\).

**Proof.** The \(\sigma\)-algebra part just follows from \(F = F \cap F\) and \(\bigcup_n (F \cap A_n) = F \cap \bigcup_n A_n\).

\(P\) is a probability measure since \(P(F) = \mathbb{P}(F)/\mathbb{P}(F) = 1\), and if \((B_n)\) is a sequence of mutually disjoint events in \(\mathcal{F}|_F\), then

\[
P\left(\bigcup_n B_n\right) = \mathbb{P}\left(\bigcup_n B_n\right)/\mathbb{P}(F) = \frac{1}{\mathbb{P}(F)} \sum_n \mathbb{P}(B_n) = \sum_n P(B_n).
\]

\(Q\) is a probability measure since \(Q(\Omega) = \mathbb{P}(\Omega \cap F)/\mathbb{P}(F) = 1\) and if \((A_n)\) is a sequence of disjoint events in \(\mathcal{F}\), then

\[
Q\left(\bigcup_n A_n\right) = \frac{1}{\mathbb{P}(F)} \sum_n \mathbb{P}(A_n \cap F) = \sum_n Q(A_n).
\]

\(\square\)

What is the intuitive meaning of the above measure \(Q\)? What we did is restrict all events to there intersection with \(F\), and look at them only in \(F\), normalized to have
total probability 1. This can be thought of as the probability of an event, given that we
know that the outcome is in $F$.

**Definition 4.2** (Conditional Probability). Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, and let $F \in \mathcal{F}$ be an event such that $\mathbb{P}(F) > 0$.

For any event $A \in \mathcal{F}$, the quantity $\mathbb{P}(A \cap F)/\mathbb{P}(F)$ is called the **conditional probability of $A$ given $F$**.

The probability measure $\mathbb{P}(\cdot | F) := \mathbb{P}(\cdot \cap F)/\mathbb{P}(F)$ is called the **conditional probability given $F$**.

!!! One must be careful with conditional probabilities - intuition here is many
times misleading.

**Example 4.3.** An urn contains 10 white balls, 5 yellow balls and 10 black balls. Shir
takes a ball out of the urn, all balls equally likely.

- What is the probability that the ball removed is yellow?
- What is the probability the ball removed is yellow given it is not black?

**Solution.**

$A =$ the ball is yellow. $B =$ the ball is not black.

$\mathbb{P}(A) = 5/25 = 1/5$. $\mathbb{P}(B) = 15/25 = 3/5$. $\mathbb{P}(A \cap B) = \mathbb{P}(A) = 1/5$. So $\mathbb{P}(A|B) = (1/5)/(3/5) = 1/3$.

**Example 4.4.** Noga has the heart of an artist and the mind of a mathematician. Given
she takes a course in probability she will pass with probability $1/3$. Given she take a
course in contemporary art she will pass with probability $1/2$. She chooses which course
to take using a fair coin toss - heads for probability and tails for art.

What is the probability that Noga takes probability and passes?

What is the probability she passes no matter what course she takes?

**Solution.**

The sample space here is

$$\Omega = \{H, T\} \times \{\text{pass, fail}\}.$$
The events we are interested in are $A = \text{Noga passes the course} = \{H, T\} \times \{\text{pass}\}$. $B = \text{Noga takes probability} = \{H\} \times \{\text{pass, fail}\}$. $B^c = \text{Noga takes art} = \{H\} \times \{\text{pass, fail}\}$. So the event that Noga takes and passes probability is $A \cap B$.

The information in the question tells us that $P(A|B) = \frac{1}{3}$ and $P(A|B^c) = \frac{1}{2}$. Thus,

$$P(A \cap B) = P(A|B)P(B) = \frac{1}{6},$$

and

$$P(A) = P(A \cap B) + P(A \cap B^c) = \frac{1}{6} + \frac{1}{4} = \frac{5}{12}.$$ 

\[\square\]

**Example 4.5.** An urn contains 8 black balls and 4 white balls. Two balls are removed one after the other, any ball being equally likely.

What is the probability that the second ball is black, given that the first is black? What is the probability that the first is black, given that the second is black? \(\triangle \nabla \triangle\)

**Solution.** $A = \text{first ball is black}$. $B = \text{second ball is black}$. We know that $P(A) = \frac{8}{12}$, and that $P(A \cap B) = \frac{67}{1211}$. So (as is intuitive) $P(B|A) = \frac{7}{11}$. However, maybe somewhat less intuitive is $P(A|B)$. Note that $P(B \cap A^c) = \frac{48}{1211}$. So

$$P(B) = P(B \cap A) + P(B \cap A^c) = \frac{8 \cdot (4 + 7)}{12 \cdot 11} = 8/12.$$ 

Thus, $P(A|B) = P(A \cap B)/P(B) = 7/11$. \[\square\]

**Exercise 4.6.** Prove that for all events $A_1, \ldots, A_n$ such that $P(A_1 \cap \cdots \cap A_n) > 0$

$$P(A_1 \cap \cdots \cap A_n) = P(A_n|A_1 \cap \cdots \cap A_{n-1}) \cdot P(A_{n-1}|A_1 \cap \cdots \cap A_{n-2}) \cdots P(A_2|A_1) \cdot P(A_1)$$

$$= P(A_1) \cdot \prod_{j=1}^{n-1} P(A_{j+1}|A_1, \ldots, A_j).$$

**Solution.** By induction. The base is simple. The induction step follows from

$$P(A_1 \cap \cdots \cap A_{n+1}) = P(A_{n+1}|A_1 \cap \cdots \cap A_n) \cdot P(A_1 \cap \cdots \cap A_n).$$

\[\square\]
4.2. Bayes’ Rule

Let \((\Omega, \mathcal{F}, \mathbb{P})\) be a probability space. Let \((A_n)_{n}\) be a partition of \(\Omega\); that is, \((A_n)\) are mutually disjoint and \(\bigcup_n A_n = \Omega\). Assume further that \(\mathbb{P}(A_n) > 0\) for all \(n\). Let \(A, B\) be events of positive probability. Then, since \((B \cap A_n)\) are mutually disjoint, by additivity we have

\[
\mathbb{P}(B) = \sum_n \mathbb{P}(B \cap A_n) = \sum_n \mathbb{P}(B|A_n) \mathbb{P}(A_n).
\]

This is called the law of total probability.

Figure 3. The law of total probability. In this case \(\Omega = A_1 \cup \cdots \cup A_{12}\), and

\[
\mathbb{P}(B) = \mathbb{P}(B|A_1) \mathbb{P}(A_1) + \mathbb{P}(B|A_2) \mathbb{P}(A_2) + \mathbb{P}(B|A_5) \mathbb{P}(A_5) + \mathbb{P}(B|A_6) \mathbb{P}(A_6) + \mathbb{P}(B|A_7) \mathbb{P}(A_7) + \mathbb{P}(B|A_9) \mathbb{P}(A_9) + \mathbb{P}(B|A_{10}) \mathbb{P}(A_{10}).
\]

Another observation is Bayes’ rule:

\[
\mathbb{P}(B|A) = \mathbb{P}(A|B) \cdot \frac{\mathbb{P}(B)}{\mathbb{P}(A)}.
\]

Combining the two, we have that

\[
\mathbb{P}(A_n|B) = \mathbb{P}(B|A_n) \cdot \frac{\mathbb{P}(A_n)}{\sum_n \mathbb{P}(B|A_n) \mathbb{P}(A_n)}.
\]
Exercise 4.7. A new machine is invented that determines if a student will become a millionaire in the next ten years or not. Given that a student will become a millionaire in ten years, the machine succeeds in predicting this 95% of the time. Given that the student will not become a millionaire in the next ten years, the machine predicts (wrongly) that he will become a millionaire in 1% of the cases (this is known as a false positive). Only 0.5% of the students will become millionaires in the next ten years. Given that Zuckerberg has been predicted to become a millionaire by the machine, what is the probability that Zuckerberg will actually become one?

Solution. \( A = \{ \text{Zuckerberg will become a millionaire in the next ten years} \} \). \( B = \{ \text{Zuckerberg is predicted to become a millionaire by the machine} \} \). The information is:

\[ P(A) = 0.005 \quad P(B|A) = 0.95 \quad P(B|A^c) = 0.01. \]

So,

\[
P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B|A) \cdot P(A) + P(B|A^c) \cdot P(A^c)} = \frac{0.95 \cdot 0.005}{0.95 \cdot 0.005 + 0.01 \cdot 0.995}
\]

\[ = \frac{95}{95 + 199} = 0.319727891 < 1/3. \]

Not such a great machine after all... \( \square \)

Exercise 4.8 (Pólya’s Urn). An urn contains one black ball and one white ball. At every time step, a ball is chosen randomly from the urn, and returned with another new ball of the same color. (So at time \( t \) there are a total of \( t + 2 \) balls.) Calculate

\[ p(t, b) = P( \text{there are} \ b \ \text{black balls at time} \ t). \]

Solution. First let’s see how \( p(t, b) \) develops when we change \( t \) and \( b \) by 1.

If at time \( t \) there are \( b \) black balls out of a total of \( t + 2 \), then at time \( t + 1 \) there are \( b + 1 \) black balls with probability \( b/t + 2 \) and \( b \) black balls with probability \( 1 - b/t + 2 \). Thus,

\[ p(t + 1, b) = \frac{b - 1}{t + 2} \cdot p(t, b - 1) + \left( 1 - \frac{b - 1}{t + 2} \right) \cdot p(t, b). \]
Also, the initial conditions are \( p(0, 1) = 1 \). Let \( q(t, b) = (t+1)! \cdot p(t, b) \). Then, \( q(0, 1) = 1 \) and for \( t > 0 \),

\[
q(t, b) = (b-1)q(t-1, b-1) + (t+1-b)q(t-1, b).
\]

Check that \( q(t, b) = t! \) solves this. So \( p(t, b) = 1/(t+1) \).

\[\square\]

**Example 4.9.** The probability a family has \( n \) children is \( p_n \), where \( \sum_{n=0}^\infty p_n = 1 \). Given that a family has \( n \) children, all possibilities for the sex of these children are equally likely. What is the probability the family has one child, given that the family has no girls?

\[\triangle \nabla \triangle\]

**Solution.** The natural sample space to take here is

\[
\Omega = \{(s_1, s_2, \ldots, s_n) : s_j \in \{\text{boy, girl}\}, n \geq 1\} \cup \{\text{no children}\}.
\]

\( A_n = \{ n \text{ children } \}, \) for \( n \geq 0 \). \( B = \{ \text{no girls } \} \).

The information in the question is:

For all \( n \geq 1 \quad P(\{(s_1, \ldots, s_n)\} | A_n) = 2^{-n}, \)

for any combination of sexes \( s_1, \ldots, s_n \in \{\text{boy, girl}\} \). Thus, \( P(B | A_n) = 2^{-n} \) for all \( n \geq 1 \).

Also, \( P(B | A_0) = 1 = 2^{-0} \). The law of total probability gives,

\[
P(B) = \sum_n P(B | A_n) P(A_n) = \sum_{n=0}^\infty 2^{-n} p_n.
\]

Using Bayes’ rule

\[
P(A_1 | B) = \frac{P(A_1) \cdot P(B | A_1)}{P(B)} = \frac{1}{2} \cdot \frac{p_1}{\sum_{n=0}^\infty 2^{-n} p_n}.
\]

\[\square\]

**Example 4.10.** There are 6 machines \( M_1, \ldots, M_6 \) in the basement of the math department. When pressing the red button on machine \( M_j \), the machine outputs a random number in \( \mathbb{N} \), with the probability that number is \( k \) being \( \frac{j}{j+1} \cdot (j+1)^{-k} \).

A bored mathematician has access to these machines. She tosses a die, and uses machine \( M_j \) if the outcome of the die is \( j \), and writes down the resulting number.

What is the probability the resulting number is greater than 0?
What is the probability the outcome of the die is 1 given that the resulting number is greater than 3?

Solution. \( A_j = \{ \text{the die shows } j \} \), \( j = 1, \ldots, 6 \). \( B_k = \{ \text{the resulting number is greater than } k \} \). \( C_k = \{ \text{the resulting number is } k \} \). So \((C_k)\) are mutually disjoint, and \( B_k = \bigcup_{n>k} C_n \).

The information is:

\[
\Pr(B_k|A_j) = \sum_{n>k} \Pr(C_n|A_j) = \sum_{n>k} \frac{j}{j+1} \cdot (j+1)^{-n} = \frac{j}{j+1} \cdot (j+1)^{-(k+1)} \cdot \sum_{n=0}^{\infty} (j+1)^{-n} = (j+1)^{-(k+1)}.
\]

Also, \( \Pr(A_j) = 1/6 \) for all \( j \).

By the law of total probability,

\[
\Pr(B_k) = \frac{1}{6} \sum_{j=1}^{6} (j+1)^{-(k+1)} = \frac{1}{6} \left( 2^{-(k+1)} + 3^{-(k+1)} + \ldots + 7^{-(k+1)} \right).
\]

So \( \Pr(B_0) = \frac{1}{6} \left( \frac{1}{2} + \frac{1}{3} + \ldots + \frac{1}{7} \right) \).

Now, by Bayes

\[
\Pr(A_1|B_3) = \Pr(B_3|A_1) \cdot \frac{\Pr(A_1)}{\Pr(B_3)} = 2^{-4} \cdot \frac{1}{2^{-4} + 3^{-4} + \ldots + 7^{-4}}.
\]

\[\Box\]

4.3. The Erdös-Ko-Rado Theorem

Consider \( n \) elements, say \( \{0, 1, \ldots, n-1\} \) without loss of generality. We want to “collect” subsets of size \( k \leq n/2 \) so that any two of these subsets intersect one another. We want to collect as many such subsets as possible.

One strategy is as follows: Suppose w.l.o.g. \( A = \{0, 1, \ldots, k\} \) is our first set. Any other set that we collect must intersect it, and we want to give ourselves as much freedom as possible. So lets take all subsets of \( \{0, 1, \ldots, n-1\} \) that contain 0. These all must intersect pairwise.

How many sets did we collect? \( \binom{n-1}{k-1} \) because we are just actually choosing \( k-1 \) elements out of \( \{1, \ldots, n-1\} \) for every such subset.

Is this the best possible? The Erdös-Ko-Rado Theorem tells us that it is.
Theorem 4.11 (Erdős-Ko-Rado). Let \( n \geq 2k \). If \( A \subset 2^{\{0, \ldots, n-1\}} \) is a collection of subsets such that for all \( A, B \in A \) : \( |A| = k \) and \( A \cap B \neq \emptyset \). Then, there are at most \( \binom{n-1}{k-1} \) sets in \( A \); that is, \( |A| \leq \binom{n-1}{k-1} \).

This elementary proof is by Katona.

Proof. Let \( A \) be as in the theorem.

Let us choose a uniform element \( \sigma \) from \( S_n \) the set of all permutations on \( \{0, 1, \ldots, n-1\} \), and also a uniform index \( I \) from \( \{0, 1, \ldots, n-1\} \). Set \( A = \{ \sigma(I), \sigma(I+1), \ldots, \sigma(I+k-1) \} \) where addition is modulo \( n \). That is, \( \mathbb{P}[\sigma = s, I = i] = \frac{1}{n!} \frac{1}{n} \).

For any given set \( B = \{b_0, \ldots, b_{k-1}\} \subset 2^{\{0, \ldots, n-1\}} \) with \( |B| = k \), by the law of total probability,

\[
\mathbb{P}[A = B, I = j] = \sum_{\pi \in S_n} \mathbb{P}[\sigma = \pi, I = j, \{\pi(j), \pi(j+1), \ldots, \pi(j+k-1)\} = B]
\]

\[
= \frac{1}{n!} \frac{1}{n} \cdot \# \{ \pi \in S_n : \{\pi(j), \pi(j+1), \ldots, \pi(j+k-1)\} = B \}.
\]

This last set is of size \( k!(n-k)! \) because there are \( k \) choices for \( \pi(j) \), then \( k-1 \) choices for \( \pi(j+1) \), etc. until only 1 choice for \( \pi(j+k-1) \), and then \( n-k \) choices for \( j+k \), \( n-k-1 \) choices for \( j+k+1 \), and so on for all \( j+k, \ldots, j+n-1 \). Thus,

\[
\mathbb{P}[A = B] = \sum_{j=0}^{n-1} \mathbb{P}[A = B, I = j] = \sum_{j=0}^{n-1} \binom{n-1}{k-1} \cdot \frac{1}{n} = \binom{n}{k}^{-1}.
\]

We have found an alternative description to choosing a subset of \( 2^{\{0, \ldots, n-1\}} \) of size exactly \( k \) uniformly.

Now the second part: For \( \pi \in S_n \) and an integer \( s \) consider the subsets \( X_{\pi,s} := \{\pi(s), \pi(s+1), \ldots, \pi(s+k-1)\} \), where as usual addition is modulo \( n \).

Note that if \( 0 \leq m \leq k-1 \) then \( X_{\pi,m} \cap X_{\pi,m+k} = \emptyset \) (here we use the assumption that \( 2k \leq m \)). Indeed if \( x \in X_{\pi,m} \cap X_{\pi,m+k} \) then there exist \( 0 \leq i, j \leq k-1 \) such that \( \pi(m+j) = x = \pi(m+k+i) \). Since \( \pi \) is a permutation, this implies that \( j = k+i \) modulo \( n \) and this is impossible because \( 0 \leq i, j \leq k-1 \) and \( k \leq n/2 \).

Now, for a given \( \pi, s \), the subsets \( X_{\pi,m} \) that intersect \( X_{\pi,s} \) are \( (X_{\pi,m} : m = s-k+1, \ldots, s+k-1) \). Without the case \( m = 0 \), these can be written in \( k-1 \) pairs:
(X_{\pi,s-m}, X_{\pi,s-m+k})_{m=1}^{k-1}. For each pair, since X_{\pi,s-m} \cap X_{\pi,s-m+k} = \emptyset, only one of the pair can be in the intersecting family \( \mathcal{A} \).

We conclude that for any \( 0 \leq j \leq n-1 \), if \( X_{\pi,j} \in \mathcal{A} \) then \( X_{\pi,s} \cap X_{\pi,j} \neq \emptyset \) so \( s-k+1 \leq j \leq s+k-1 \), and \( X_{\pi,s-m} \in \mathcal{A} \) if and only if \( X_{\pi,s-m+k} \not\in \mathcal{A} \). Thus,

\[
\sum_{j=0}^{n-1} 1_{\{X_{\pi,j} \in \mathcal{A}\}} \leq 1_{\{X_{\pi,s} \in \mathcal{A}\}} + \sum_{m=1}^{k-1} 1_{\{X_{\pi,s-m} \in \mathcal{A}\}} + 1_{\{X_{\pi,s-m+k} \not\in \mathcal{A}\}}
\]

\[
\leq 1 + k - 1 = k.
\]

Since all subsets of \( \mathcal{A} \) must intersect one another, this gives that for any \( \pi \in S_n \),

\[
\# \{ 0 \leq j \leq n-1 : X_{\pi,j} \in \mathcal{A}\} \leq k.
\]

Now, if \( \sigma = \pi \) and \( A \in \mathcal{A} \) then it must be that \( I = j \) for some \( j \) such that \( X_{\pi,j} \in \mathcal{A} \).

So, by Boole's inequality,

\[
P[A \in \mathcal{A}, \sigma = \pi] \leq P[\sigma = \pi, I \in \{0 \leq j \leq n-1 : X_{\pi,j} \in \mathcal{A}\}]
\]

\[
\leq \sum_{j=0}^{n-1} 1_{\{X_{\pi,j} \in \mathcal{A}\}} P[\sigma = \pi, I = j] \leq \frac{k}{n! \cdot n}.
\]

Summing over all \( \pi \in S_n \), by the law of total probability,

\[
P[A \in \mathcal{A}] \leq \frac{k}{n}.
\]

We now combine this bound with the fact that \( A \) is uniform over all \( k \)-subsets of \( \{0, \ldots, n-1\} \). Thus,

\[
\frac{|A|}{\binom{n}{k}} = P[A \in \mathcal{A}] \leq \frac{k}{n}.
\]

This is exactly \( |A| \leq \binom{n}{k} \cdot \frac{k}{n} = \binom{n-1}{k-1} \). \qed
5.1. Independence

Until now, we have only used the measure theoretic properties of probability measures. The main difference between probability and measure theory is the notion of independence.

This is perhaps the most important definition in probability.

**Definition 5.1.** Let $A, B$ be events in a probability space $(\Omega, \mathcal{F}, \mathbb{P})$. We say that $A$ and $B$ are independent if

\[ \mathbb{P}(A \cap B) = \mathbb{P}(A) \mathbb{P}(B). \]

If $(A_n)_n$ is a collection of events, we say that $(A_n)$ are mutually independent if for any finite number of events in the collection, $A_{n_1}, A_{n_2}, \ldots, A_{n_k}$ we have that

\[ \mathbb{P}(A_{n_1} \cap A_{n_2} \cap \cdots \cap A_{n_k}) = \mathbb{P}(A_{n_1}) \cdot \mathbb{P}(A_{n_2}) \cdots \mathbb{P}(A_{n_k}). \]

**Example 5.2.** A card is chosen randomly out of a deck of 52, all cards equally likely. $A$ is the event that the card is an ace. $B$ is the event that the card is a spade. $C$ is the event that the card is a jack of diamonds or a 2 of hearts. $D$ is the event that the card is an even number.

Then, $A, B$ are independent, since

\[ \mathbb{P}(A \cap B) = \frac{1}{52} = \frac{4}{52} \cdot \frac{13}{52} = \mathbb{P}(A) \cdot \mathbb{P}(B). \]

$C, D$ are not independent since

\[ \mathbb{P}(C \cap D) = \frac{1}{52} \neq \mathbb{P}(C) \cdot \mathbb{P}(D) = \frac{2}{52} \cdot \frac{20}{52}. \]
Example 5.3. Two dice are tossed, all outcomes equally likely. \( A = \{ \text{the first die is 4} \} \). \( B = \{ \text{the sum of the dice is 6} \} \). \( C = \{ \text{the sum of the dice is 7} \} \).

\( A, B \) are not independent.

\[
\mathbb{P}(A \cap B) = \frac{1}{36} \neq \mathbb{P}(A) \cdot \mathbb{P}(B) = \frac{1}{6} \cdot \frac{5}{36}.
\]

However, \( A, C \) are independent,

\[
\mathbb{P}(A \cap C) = \frac{1}{36} = \frac{1}{6} \cdot \frac{6}{36} = \mathbb{P}(A) \cdot \mathbb{P}(C).
\]

Proposition 5.4. Let \((\Omega, \mathcal{F}, \mathbb{P})\) be a probability space.

(1) Any event \( A \) is independent of \( \Omega \) and of \( \emptyset \).

(2) If \( A, B \) are independent, then \( A^c, B \) are independent.

(3) If \( A, B \) are independent and \( \mathbb{P}(B) > 0 \) then \( \mathbb{P}(A|B) = \mathbb{P}(A) \).

Proof. If \( A, B \) are independent,

\[
\mathbb{P}(B \cap A^c) = \mathbb{P}(B) - \mathbb{P}(B \cap A) = \mathbb{P}(B)(1 - \mathbb{P}(A)) = \mathbb{P}(B) \mathbb{P}(A^c).
\]

If also \( \mathbb{P}(B) > 0 \), then \( \mathbb{P}(A|B) = \mathbb{P}(A) \mathbb{P}(B)/\mathbb{P}(B) = \mathbb{P}(A) \). \( \square \)

Exercise 5.5. Prove that if \( A, B, C \) are mutually independent, then

(1) \( A \) and \( B \cap C \) are independent.

(2) \( A \) and \( B \setminus C \) are independent.

Solution.

\[
\mathbb{P}(A \cap B \cap C) = \mathbb{P}(A) \mathbb{P}(B) \mathbb{P}(C) = \mathbb{P}(A) \mathbb{P}(B \cap C).
\]

This proves the first assertion. For the second assertion, since \( B \setminus C = B \cap C^c \) and since \( A \cap B \) is independent of \( C \), then also \( A \cap B \) and \( C^c \) are independent. So,

\[
\mathbb{P}(A \cap (B \setminus C)) = \mathbb{P}(A \cap B \cap C^c) = \mathbb{P}(A \cap B) \mathbb{P}(C^c) = \mathbb{P}(A) \mathbb{P}(B)(1 - \mathbb{P}(C)).
\]
Finally,

$$P(B \setminus C) = P(B \setminus (B \cap C)) = P(B) - P(B \cap C) = P(B)(1 - P(C)),$$

since $B, C$ are independent. □

**Exercise 5.6.** Show that if $B$ is an event with $P[B] = 0$ then for any event $A$, $A$ and $B$ are independent.

**Solution.** $0 \leq P[A \cap B] \leq P[B] = 0$ and $P[B] \cdot P[A] = 0$. □

**Example 5.7.** Consider the uniform probability measure on $\Omega = \{0, 1\}^2$. Let

$$A = \{(x, y) \in \Omega : x = 1\} \quad B = \{(x, y) \in \Omega : y = 1\},$$

and

$$C = \{(x, y) \in \Omega : x \neq y\}.$$

Any two of these events are independent; indeed,

$$P(A \cap B) = P(A \cap C) = P(B \cap C) = \frac{1}{4},$$

$$P(A) = P(B) = P(C) = \frac{1}{2},$$

(because $C = \{(0, 1), (1, 0)\}$). However, it is not the case that $(A, B, C)$ are mutually independent, since

$$P(A \cap B \cap C) = 0 \neq \frac{1}{8} = P(A) \cdot P(B) \cdot P(C).$$

△ ▽ △

5.2. An example using independence: Riemann zeta function and Euler’s formula

Some number theory via probability:

Let $1 < s \in \mathbb{R}$ be a parameter. Let

$$\zeta(s) = \sum_{n=1}^{\infty} n^{-s}.$$
Let $X$ be a random variable with range $\{1, 2, \ldots, \}$ and distribution given by

$$\mathbb{P}(X = n) = \frac{1}{\zeta(s)} \cdot n^{-s}.$$ 

Let $D_m = m\mathbb{Z}$ be the set of positive integers divisible by $m$. So

$$\mathbb{P}_X(D_m) = \frac{\sum mn^{-s}}{\sum n^{-s}} = m^{-s}.$$ 

Now, let $p_1, p_2, \ldots, p_k$ be $k$ different prime numbers. Then,

$$D_{p_1} \cap D_{p_2} \cap \cdots \cap D_{p_k} = D_{p_1p_2\cdots p_k}.$$ 

So

$$\mathbb{P}_X(D_{p_1} \cap D_{p_2} \cap \cdots \cap D_{p_k}) = (p_1p_2\cdots p_k)^{-s} = \mathbb{P}_X(D_{p_1}) \cdot \mathbb{P}_X(D_{p_2}) \cdots \mathbb{P}_X(D_{p_k}).$$ 

Since this is true for any finite number of primes, we have that the collection $(D_p)_{p \in P}$ are mutually independent, where $P$ is the set of primes.

Since the only number that is not divisible by any prime is 1, we have that $\{1\} = \bigcap_{p \in P} D_p^c$, and so

$$\mathbb{P}(X = 1) = \mathbb{P}_X(\{1\}) = \prod_{p \in P} (1 - \mathbb{P}_X(D_p)) = \prod_{p \in P} (1 - p^{-s}).$$ 

So we conclude Euler’s formula: for any $s > 1$,

$$\sum_{n=1}^{\infty} n^{-s} = \prod_{p \in P} (1 - p^{-s})^{-1}.$$ 

We can also let $s \to 1$ from above, and we get

$$0 = \lim_{s \to 1} \zeta(s)^{-1} = \prod_{p \in P} (1 - p^{-1}).$$ 

Taking logs

$$\sum_{p \in P} \log(1 - p^{-1}) = -\infty.$$ 

Since $\log(1 - x) \geq -3x/2$ for $0 < x < 1$, we have that

$$-\infty \geq -\frac{3}{2} \sum_{p \in P} p^{-1},$$
and so

\[ \sum_{p \in P} \frac{1}{p} = \infty. \]

This is a non-trivial result.
6.1. Uncountable Sample Spaces

Up to now we have dealt with the case of countable sample spaces, but there is no reason to restrict to those. For example, what if we wish to experiment by throwing a dart at a target? The collection of possible outcomes is the collection of points in the target, which is uncountable.

The basic theory to define such spaces is measure theory.

Consider the sample space $\Omega = [0, 1]$. Why not take a probability measure on all subsets of $[0, 1]$?

Well, for example, we would like to consider the experiment of sampling a number from $[0, 1]$ such that it is uniform over the interval; that is, such that for any $0 \leq a < b \leq 1$ the probability the number is in $(a, b)$ is $b - a$.

How can we do this?

The following proposition shows that this cannot be done using all subsets of $[0, 1]$ (!).

**Proposition 6.1.** Let $\Omega = [0, 1]$. Let $P$ be a probability measure on $(\Omega, F)$, where $F$ is some $\sigma$-algebra on $\Omega$, such that for any $x \in [0, 1]$ and $A \in F$, $P(A) = P(A + x)$, where $A + x = \{a + x \pmod{1} : a \in A\}$. Then, there exists a subset $S \subset \Omega$ such that $S \notin F$.

**Proof.** Define the equivalence relation $x \sim y$ iff $x - y \in \mathbb{Q}$. So we can choose a representative for each different equivalence class (using the axiom of choice!). Let $R$ be the set of these representatives.

Define $R_q = R + q = \{r + q \pmod{1} : r \in R\}$, for all $q \in \mathbb{Q} \cap [0, 1)$. For $q \neq p \in \mathbb{Q} \cap [0, 1)$ we have that $R_q \cap R_p = \emptyset$, since if $x \in R_p \cap R_q$ then $x = r + q = r' + p \pmod{1}$, for
$r, r' \in R$, and this implies that $r \sim r'$, a contradiction. Thus, $(R_q)_{q \in \mathbb{Q} \cap [0, 1)}$ is a countable collection of mutually disjoint events, so $1 = \mathbb{P}([0, 1)) = \mathbb{P}(\bigcup_q R_q) = \sum_q \mathbb{P}(R_q) = \sum_q \mathbb{P}(R) = \infty!$

Thus, we have to develop a theory of sets which we allow to be events, so that everything is well defined etc.

### 6.2. $\sigma$-algebras Revisited

Recall the definition of a $\sigma$-algebra:

**Definition 6.2.** Let $\Omega$ be a sample space. A collection of subsets $\mathcal{F} \subset 2^\Omega$ is called a $\sigma$-algebra if

- $\Omega \in \mathcal{F}$.
- If $A \in \mathcal{F}$ then $A^c \in \mathcal{F}$.
- If $(A_n)$ is a collection of events in $\mathcal{F}$ then $\bigcup_n A_n \in \mathcal{F}$.

These were the most basic properties we wanted of events: Everything can occur, if something can occur, it can also not occur, and if there are many possible events, we can ask if there is one of them that has occurred.

**Exercise 6.3.** Show that $\mathcal{F} \subset 2^\Omega$ is a $\sigma$-algebra if and only if it has the following properties.

- $\emptyset \in \mathcal{F}$.
- If $A \in \mathcal{F}$ then $A^c \in \mathcal{F}$.
- If $(A_n)$ is a collection of events in $\mathcal{F}$ then $\bigcap_n A_n \in \mathcal{F}$.

**Solution.** Use de-Morgan.

Suppose we have a collection of subsets of $\Omega$. We want the smallest $\sigma$-algebra containing these subsets.

**Proposition 6.4.** Let $\Omega$ be a sample space, and let $\mathcal{K} \subset 2^\Omega$ be a collection of subsets. There exists a unique $\sigma-$algebra on $\Omega$, denoted by $\sigma(\mathcal{K})$, such that

- $\mathcal{K} \subset \sigma(\mathcal{K})$. 


• If $\mathcal{F}$ is a $\sigma$-algebra such that $\mathcal{K} \subseteq \mathcal{F}$, then $\sigma(\mathcal{K}) \subseteq \mathcal{F}$

So $\sigma(\mathcal{K})$ is the smallest $\sigma$-algebra containing $\mathcal{K}$. Moreover, $\sigma(\mathcal{K}) = \mathcal{K}$ if and only if $\mathcal{K}$ is itself a $\sigma-$algebra.

Proof. Let

$$\Gamma(\mathcal{K}) = \{ \mathcal{F} : \mathcal{F} \text{ is a } \sigma-\text{algebra such that } \mathcal{K} \subseteq \mathcal{F} \},$$

and define $\sigma(\mathcal{K}) := \bigcap \Gamma(\mathcal{K})$. So of course the second property holds. $\sigma(\mathcal{K})$ is a $\sigma$-algebra since $\Omega \in \mathcal{F}$ for all $\mathcal{F} \in \Gamma(\mathcal{K})$ in the above set. Also, if $(A_n)$ is a collection of events in $\sigma$, then $A_n \in \mathcal{F}$ for all $n$ and all $\mathcal{F} \in \Gamma(\mathcal{K})$. Thus, $\bigcup_n A_n \in \mathcal{F}$ for every $\mathcal{F} \in \Gamma(\mathcal{K})$, and so $\bigcup_n A_n \in \sigma(\mathcal{K})$.

As for uniqueness, if $\mathcal{G}$ is another $\sigma$-algebra with the above properties, then $\mathcal{G} \subseteq \sigma(\mathcal{K})$ because $\mathcal{K} \subseteq \mathcal{G}$ and similarly $\sigma(\mathcal{K}) \subseteq \mathcal{G}$.

$\square$

**Definition 6.5.** $\sigma(\mathcal{K})$ as above is called the $\sigma$-algebra generated by $\mathcal{K}$.

$\sigma(\mathcal{K})$ is intuitively all the information carried by the events in $\mathcal{K}$.

**Example 6.6.** Let $\Omega$ be a sample space and $A \subset \Omega$. Then, $\sigma(\{A\}) = \{\emptyset, A, A^c, \Omega\}$. $\triangle \nabla \triangle$

**Definition 6.7.** Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, and let $\mathcal{G} \subset \mathcal{F}$ be a sub-$\sigma$-algebra of $\mathcal{F}$. We say that an event $A \in \mathcal{F}$ is independent of $\mathcal{G}$ if for any $B \in \mathcal{G}$, we have that $A, B$ are independent.

Let $(B_\alpha)_\alpha$ be a collection of events. We say that $A$ is independent of $(B_\alpha)_\alpha$ if $A$ is independent of the $\sigma$-algebra $\sigma((B_\alpha)_\alpha)$.

**Example 6.8.** If $A, B$ are independent, then we have already seen that $A, B^c$ are also independent. Since $A, \emptyset$ are always independent, and also $A, \Omega$ are always independent, we have that $A$ is independent of $\sigma(B) = \{\emptyset, B, B^c, \Omega\}$. $\triangle \nabla \triangle$

### 6.3. Borel $\sigma$-algebra

Let $E$ be a topological space. We can consider the $\sigma$-algebra generated by all open sets. This is called the **Borel $\sigma$-algebra** on $E$, and an event in this $\sigma$-algebra is call a **Borel set**.
A special case is the case $E = \mathbb{R}^d$, and more specifically $E = \mathbb{R}$. In this case we denote the Borel $\sigma$-algebra by $\mathcal{B} = \mathcal{B}(\mathbb{R})$, and it can be shown that $\mathcal{B} = \sigma((a, b) : a < b \in \mathbb{R})$; that is $\mathcal{B}$ is generated by all intervals.

**Exercise 6.9.** Show that

$$\mathcal{B}(\mathbb{R}) = \sigma((a, b] : a < b \in \mathbb{R}) = \sigma([a, b) : a < b \in \mathbb{R}) = \sigma([a, b) : a < b \in \mathbb{R}).$$

**Solution.** Let $a < b \in \mathbb{R}$. Since $(a, b) \cap [0, 1] = \bigcap_n ((a, b - 1/n] \cap [0, 1])$, we get that $(a, b) \cap [0, 1] \in \sigma((a, b] \cap [0, 1] : a < b \in \mathbb{R})$. Since this holds for all $a < b \in \mathbb{R}$, we have that $\sigma((a, b] \cap [0, 1] : a < b \in \mathbb{R}) \subset \sigma((a, b] \cap [0, 1] : a < b \in \mathbb{R})$.

The other inclusions are similar, so all the mentioned $\sigma$–algebras are equal. \qed

### 6.4. Lebesgue Measure

The following basic theorem, due to Lebesgue, is shown in measure theory.

**Theorem 6.10** (Lebesgue). Let $\Omega = [0, 1]$, and let $\mathcal{B}$ be the Borel $\sigma$-algebra on $\Omega$. Let $F : [0, 1] \to [0, 1]$ be a right-continuous, non-decreasing function such that $F(1) = 1$ and $F(0) = 0$. Then, there exists a unique probability measure, denoted $\mathbb{P} = dF$, on $(\Omega, \mathcal{B})$ such that for any $0 \leq a < b \leq 1$, $\mathbb{P}((a, b]) = F(b) - F(a)$.

We will not go into the proof of this theorem, but it gives us many new probability measures on non-discrete sample spaces that we can define.

**Example 6.11.** Take $F(x) = x$ in the above theorem. So, the resulting probability measure, sometimes denoted $\mathbb{P} = dx$, has the property that for any interval $0 \leq a < b \leq 1$, $\mathbb{P}((a, b]) = b - a$. One can also check that this measure is translation invariant, i.e. for any $x \in [0, 1]$,

$$\mathbb{P}(x + (a, b] \pmod 1) = \mathbb{P}(\{x + y \pmod 1 : a < y \leq b\}) = b - a.$$

This is also sometimes called the uniform measure on $[0, 1]$.

**Exercise 6.12.** Show that for the uniform measure on $[0, 1]$, for any $a \in [0, 1]$, $dx(\{a\}) = 0$. \qed
Deduce that for all \( 0 \leq a \leq b \leq 1 \),

\[ dx((a,b)) = dx([a,b]) = dx([a,b)) = dx((a,b]) = b - a. \]

Solution. Let \( a \in [0,1] \). For all \( n \), let \( A_n = (a - \frac{1}{n}, a) \). So \( \mathbb{P}(A_n) = 1/n \). Since \( (A_n) \) is a decreasing sequence of events, and since \( \{a\} = \bigcap_n A_n \), we have by the continuity of probability measures

\[ \mathbb{P}\{a\} = \lim_n \mathbb{P}(A_n) = 0. \]

We can now note that

\[ \mathbb{P}([a,b]) = \mathbb{P}((a,b) \cup \{a\} \cup \{b\}) = \mathbb{P}((a,b)) + \mathbb{P}\{a\} + \mathbb{P}\{b\} = \mathbb{P}((a,b)). \]

Similarly for the other types of intervals. \( \square \)

Example 6.13. Let \( \alpha < \beta \in \mathbb{R} \). Let \( \Omega = [\alpha, \beta] \). For a set \( A \in \mathcal{B}([0,1]) \) let

\[ (\beta - \alpha)A + \alpha = \{((\beta - \alpha)a + \alpha : a \in A \}. \]

Define \( \mathcal{F} = \{(\beta - \alpha)A + \alpha : A \in \mathcal{B}([0,1])\} \).

We have that \( (\beta - \alpha)[0,1] + \alpha = \Omega \), \((\beta - \alpha)A + \alpha)^c = (\beta - \alpha)A^c + \alpha \), and

\[ \bigcup_n((\beta - \alpha)A_n + \alpha) = (\beta - \alpha)(\bigcup_n A_n) + \alpha. \]

So \( \mathcal{F} \) is a \( \sigma \)-algebra on \( \Omega \). Also, if \( dx \) is the uniform measure on \([0,1]\) then

\[ \mathbb{P}((\beta - \alpha)A + \alpha) := dx(A) \]

define a probability measure on \((\Omega, \mathcal{F})\), because \( A \cap B = \emptyset \) if and only if \((\beta - \alpha)A + \alpha \cap (\beta - \alpha)B + \alpha = \emptyset \).

In this case, note that for any \( a \leq \alpha < b \leq \beta \),

\[ \mathbb{P}([a,b]) = \mathbb{P}((\beta - \alpha)(\frac{1}{\beta-\alpha}(a - \alpha), \frac{1}{\beta-\alpha}b] + \alpha) = dx((\frac{1}{\beta-\alpha}(a - \alpha), \frac{1}{\beta-\alpha}b]) = \frac{b - a}{\beta - \alpha}. \]

\( \square \)

Example 6.14. Let \( \Omega \) be a sample space, and let \( f : \Omega \to [0,1] \) be a 1-1 function. Let \( \mathcal{F} = \{f^{-1}(B) : B \in \mathcal{B}([0,1])\} \). This is a \( \sigma \)-algebra on \( \Omega \) since \( f^{-1}([0,1]) = \Omega \) and

\[ \bigcup_n f^{-1}(B_n) = f^{-1}\left(\bigcup_n B_n\right). \]
For $A \in \mathcal{F}$, since $B = f(f^{-1}(B))$, we have that $f(A) \in \mathcal{B}$, and we can define $\mathbb{P}(A) = dx(f(A))$. This results in a probability measure on $(\Omega, \mathcal{F})$ since $\mathbb{P}(\Omega) = 1$ and if $(A_n)$ are mutually disjoint events then so are $(f(A_n))$,

$$
\mathbb{P}(\bigcup_n A_n) = dx(f(\bigcup_n A_n)) = dx(\bigcup_n f(A_n)) = \sum_n dx(f(A_n)) = \sum_n \mathbb{P}(A_n).
$$

Example 6.15. A dart is thrown to a target of radius 1 meter. Points are awarded according to the distance to the center of the target. The probability of being between distance $a$ and distance $b$ is $b^2 - a^2$.

What is the probability of hitting the inner circle of radius $1/2$? What about radius $1/4$?

Solution. Let $F : [0,1] \to [0,1]$ be $F(x) = x^2$. Then, $dF$ is a probability measure on $([0,1], \mathcal{B})$ such that $dF((a,b)) = F(b) - F(a) = b^2 - a^2$. Thus, we can model the dart throwing experiment by the probability space $([0,1], \mathcal{B}, dF)$, where the outcome $\omega \in [0,1]$ is the distance to the target.

Now, the probability of hitting the inner circle of radius $1/2$ is $dF([0,1/2]) = (1/2)^2 = 1/4$.

Similarly the probability of hitting the circle of radius $1/4$ is $1/16$. □
7.1. Random Variables

Up until now we dealt with abstract spaces (such as dice, cards, etc.). In order to say things about objects, we prefer to map them to numbers (or vectors) so that we can perform calculations.

For example, maybe some biological experiment is being carried out in the lab, but in the end, the biologist may only be interested on the number of bacteria in the dish, or other such measurements.

Measurements of experiments are carried out via random variables.

Suppose we have an experiment described by a probability space \((\Omega, \mathcal{F}, \mathbb{P})\). A measurement is just a function taking outcomes to numbers; i.e. a function \(X : \Omega \to \mathbb{R}\). The natural way to define a probability space on \(\mathbb{R}\), would now be to take the natural \(\sigma\)-algebra on \(\mathbb{R}\), i.e. the Borel \(\sigma\)-algebra \(\mathcal{B}\), and then define the measure of a Borel set \(B\) to be the probability that the outcome \(\omega\) is mapped into \(B\):

\[
\forall B \in \mathcal{B} \quad \mathbb{P}_X(B) := \mathbb{P}\{\omega : X(\omega) \in B\} = \mathbb{P}(X^{-1}(B)).
\]

But wait! What if the set \(X^{-1}(B)\) is not in \(\mathcal{F}\)? Then \(\mathbb{P}\) is not defined on that set. This is a technical detail, but one that needs to be addressed, otherwise we will run into paradoxes as in the case of the unmeasurable sets on \([0,1]\).

**Definition 7.1.** Let \((\Omega, \mathcal{F}), (\Omega', \mathcal{F}')\) be a measurable spaces. A function \(X : \Omega \to \Omega'\) is called a measurable map from \((\Omega, \mathcal{F})\) to \((\Omega', \mathcal{F}')\), if for all \(A' \in \mathcal{F}'\), \(X^{-1}(A') \in \mathcal{F}\).

We usually denote this by \(X : (\Omega, \mathcal{F}) \to (\Omega', \mathcal{F}')\) to stress the dependence on \(\mathcal{F}\) and \(\mathcal{F}'\).
To determine if a function $X : (\Omega, \mathcal{F}) \to (\Omega', \mathcal{F}')$ is measurable, it is enough to check for a family of sets generating $\mathcal{F}'$:

**Proposition 7.2.** Let $(\Omega, \mathcal{F}), (\Omega', \mathcal{F}')$ be two measurable spaces, such that $\mathcal{F}' = \sigma(\Pi')$ for some $\Pi' \subset 2^{\Omega'}$. Then, $X : \Omega \to \Omega'$ is measurable if and only if for any $A' \in \Pi'$, $X^{-1}(A') \in \mathcal{F}$.

**Proof.** The “only if” part is clear. For the “if” part, consider

$$\mathcal{G} = \sigma(\{X^{-1}(A') : A' \in \Pi'\}).$$

By assumption, $\mathcal{G} \subset \mathcal{F}$. On the other hand, if

$$\mathcal{G}' = \{A' \subset \Omega' : X^{-1}(A') \in \mathcal{G}\},$$

then $\mathcal{G}'$ is a $\sigma$-algebra containing $\Pi'$, and thus $\mathcal{F}' = \sigma(\Pi') \subset \mathcal{G}'$. Thus, for any $A' \in \mathcal{F}'$ we have that $X^{-1}(A') \in \mathcal{G} \subset \mathcal{F}$. So $X$ is measurable. \qed

**Corollary 7.3.** If $\Omega, \Omega'$ are topological spaces, and $\mathcal{B}(\Omega), \mathcal{B}(\Omega')$ are the Borel $\sigma$-algebras, then any continuous map $f : \Omega \to \Omega'$ is measurable.

**Proof.** $f$ is continuous implies that for any open set $O' \subset \Omega'$, $f^{-1}(O')$ is an open set in $\Omega$. Thus, since $\mathcal{B}(\Omega') = \sigma(\mathcal{O}')$, where $\mathcal{O}'$ is the set of open sets in $\Omega'$, we have that for any $O' \in \mathcal{O}'$, $f^{-1}(O') \in \mathcal{B}(\Omega)$. So $f$ is measurable. \qed

**Definition 7.4.** A (real valued) random variable on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ is a measurable function $X : (\Omega, \mathcal{F}) \to (\mathbb{R}, \mathcal{B})$.

The function $\mathbb{P}_X(B) = \mathbb{P}(X^{-1}(B))$ is a probability measure on $(\mathbb{R}, \mathcal{B})$, and is called the distribution of $X$.

**Example 7.5.** Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. Let $A \in \mathcal{F}$ be an event. This is not a random variable, but there is a natural random variable connected to $A$: Define a function

$$I_A(\omega) = \begin{cases} 
1 & \omega \in A \\
0 & \omega \notin A.
\end{cases}$$
Note that for any \( B \in \mathcal{B} \), if \( 1 \not\in B \) and \( 0 \not\in B \) then \( I_A^{-1}(B) = \emptyset \). If \( 1 \in B \) and \( 0 \not\in B \) then \( I_A^{-1}(B) = A \). If \( 1 \not\in B \) and \( 0 \in B \) then \( I_A^{-1}(B) = A^c \). If \( 1 \in B \) and \( 0 \in B \) then \( I_A^{-1}(B) = \Omega \). So \( I_A \) is measurable.

\( I_A \) is known as the \textbf{indicator} of the event \( A \). Indicators are always measurable, and they are the simplest kind of random variables, taking only two values 0, 1. We denote the indicator of an event \( A \) by \( 1_A \).

\( \Delta \nabla \triangle \)

**Exercise 7.6.** Let \((\Omega, \mathcal{F}, \mathbb{P})\) be a probability space. Prove that if \( X : (\Omega, \mathcal{F}) \to \mathbb{R} \) is a measurable function then indeed \( \mathbb{P}_X(B) = \mathbb{P}(X^{-1}(B)) \) is a probability measure on \((\mathbb{R}, \mathcal{B})\).

\( \text{Proof.} \mathbb{P}_X(\mathbb{R}) = \mathbb{P}(X^{-1}(\mathbb{R})) = \mathbb{P}(\Omega) = 1. \) If \((B_n)\) is a sequence of mutually disjoint events, then \((X^{-1}(B_n))\) are also mutually disjoint and

\[
\mathbb{P}_X(\bigcup_n B_n) = \mathbb{P}(\bigcup_n X^{-1}(B_n)) = \sum_n \mathbb{P}_X(B_n).
\]

\( \square \)

\( \textbf{Notation.} \) To simplify the notation, we will omit the \( \omega \)'s. \( e.g. \) instead of writing \( \mathbb{P}(\{\omega : X(\omega) \in A\}) \) we write \( \mathbb{P}(X \in A) \).

\( \Delta \nabla \triangle \)

**Example 7.7.** Three balls are removed from an urn containing 20 balls numbered 1 to 20. All possibilities are equally likely. What is the probability the at least one has a number 17 or higher?

The sample space here is \( \Omega = \{S \subset \{1, 2, \ldots, 20\} : |S| = 3\} \), and the measure is the uniform measure on this set. We define the random variable \( X : \Omega \to \mathbb{R} \) by \( X(\{i, j, k\}) = \max \{i, j, k\} \).

Why is \( X \) a measurable function? For any Borel set \( B \in \mathcal{B} \), we have that \( X^{-1}(B) \) is a subset of \( \Omega \), and the \( \sigma \)-algebra here is \( 2^n \).

What is the probability that there exists a ball numbered 17 or higher? This is

\[
\mathbb{P}(\{\omega \in \Omega : X(\omega) \geq 17\}) = \mathbb{P}(X \geq 17) = 1 - \mathbb{P}(X < 17).
\]

Since the number of \( S \in \Omega \) with \( \max S < 17 \) is \( \binom{16}{3} \), we have that \( \mathbb{P}(X < 17) = \frac{\binom{16}{3}}{\binom{20}{3}} = \frac{4 \cdot 7}{4 \cdot 19} \).

\( \Delta \nabla \triangle \)
Example 7.8. Two dice are thrown. So \( \Omega = \{(i, j) : 1 \leq i, j \leq 6\} \), and \( P \) is the uniform measure on this set. Define the random variable \( X(i, j) = i + j \). Note that \( 2 \leq X(i, j) \leq 12 \) for any \( i, j \), so

\[
P(X \in [2, 12] \cap \mathbb{N}) = 1,
\]

and \( X \) is a discrete random variable.

What is the probability that the sum is 7?

\[
P(X = 7) = \frac{6}{36} = \frac{1}{6}.
\]

\(\quad \triangle \diamond \triangle\)

7.2. Distribution Function

Proposition 7.9. Let \((\Omega, \mathcal{F})\) be a measurable space, and let \(X : \Omega \to \mathbb{R}\). Then \(X\) is measurable if and only if

\[
\{X \leq r\} = X^{-1}(-\infty, r] \in \mathcal{F} \quad \forall r \in \mathbb{R}.
\]

Proof. The “only if” part is clear.

For the “if” part: Since \(\{X > r\} = \{X \leq r\}^c \in \mathcal{F}\), we have that for any \(a < b \in \mathbb{R}\)

\[
X^{-1}((a, b]) = X^{-1}((-\infty, b] \setminus (-\infty, a]) = \{a < X \leq b\} = \{X \leq b\} \cap \{X > a\} \in \mathcal{F}.
\]

Let

\[
\mathcal{G} = \{B \in \mathcal{B} : X^{-1}(B) \in \mathcal{F}\}.
\]

Since \(X^{-1}(\mathbb{R}) \in \mathcal{F}\) and \(X^{-1}(B^c) = (X^{-1}(B))^c\) and \(X^{-1}(\bigcup_n B_n) = \bigcup_n X^{-1}(B_n)\), we get that \(\mathcal{G}\) is a \(\sigma\)-algebra.

From the above, \((a, b) \in \mathcal{G}\) for all \(a < b \in \mathbb{R}\). Thus, \(\mathcal{B} = \sigma((a, b] : a < b \in \mathbb{R}) \subset \mathcal{G}\). So for any \(B \in \mathcal{B}\), we have that \(B \in \mathcal{G}\), and so \(X^{-1}(B) \in \mathcal{F}\).

Definition 7.10. Let \((\Omega, \mathcal{F}, P)\) be a probability space, and let \(X\) be a random variable. The (cumulative) distribution function of \(X\) is the function

\[
F_X : \mathbb{R} \to [0, 1] \quad F_X(t) = P(X \leq t).
\]
Example 7.11. $X$ is the random variable that is the sum of two dice. $F_X(t) =$ 

$F_X(t) = 0$ for $t < 2$. $F_X(t) = 1$ for $t \geq 12$. $F_X(t) = 1/36$ for $t \in [2, 3)$. $F_X(t) = 2/36$ for $t \in [3, 4)$. In general, $F_X(t) = (k - 1)/36$ for $t \in [k, k + 1)$ and $k \leq 5$. \(\triangle \nabla \triangle\)

Proposition 7.12. Let $X$ be a random variable, and $F_X$ its distribution function. Then,

- $F_X$ is non-decreasing and right-continuous.
- $F_X(t)$ tends to 0 as $t \to -\infty$.
- $F_X(t)$ tends to 1 as $t \to \infty$.

Proof. Since for $a \leq b$ we have that $(-\infty, a] \subset (-\infty, b]$, we get

$$F_X(a) = \mathbb{P}_X((-\infty, a]) \leq \mathbb{P}_X((-\infty, b]) = F_X(b).$$

For right continuity, let $h_n \to 0$ monotonely from the right, so $((-\infty, a + h_n))_n$ is a decreasing sequence and

$$F_X(a) = \mathbb{P}_X(\bigcap_n (-\infty, a + h_n)) = \lim_n \mathbb{P}_X((-\infty, a + h_n)) = \lim_n F_X(a + h_n).$$

The limits at $\infty$ and $-\infty$ are treated similarly. If $a_n \to \infty$ then $\lim_n (-\infty, a_n] = (-\infty, \infty)$. So by the continuity of probability measures $\mathbb{P}_X((-\infty, a_n]) \to \mathbb{P}_X(\mathbb{R}) = 1$. Similarly for $a_n \to -\infty$. \(\square\)

Exercise 7.13. Give an example of a random variable $X$ whose distribution function $F_X$ is not left-continuous.

Solution. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, and let $A \in \mathcal{F}$. Let $X = 1_{\{A\}}$. What is $F_X$?

$$F_X(t) = \begin{cases} 
0 & t < 0 \\
\mathbb{P}(A^c) = 1 - \mathbb{P}(A) & 0 \leq t < 1 \\
1 & t \geq 1.
\end{cases}$$

So if $0 < \mathbb{P}(A) < 1$ then

$$\lim_{s \to 0^-} F_X(s) = 0 \quad \text{and} \quad F_X(0) = 1 - \mathbb{P}(A) > 0,$$
$$\lim_{s \to 1^-} F_X(s) = 1 - \mathbb{P}(A) \quad \text{and} \quad F_X(1) = 1 > 1 - \mathbb{P}(A).$$

So both 0 and 1 are points where $F_X$ is not left-continuous. \(\square\)
✓ Of course one sees that the right-continuity stems from the fact that we define
\( F_X(t) = \mathbb{P}(X \leq t) \). We could have defined \( \mathbb{P}(X \geq t) \) which would have been left-
continuous. This is not important, however we stick to the classical distribution function.

**Exercise 7.14.** Let \( X \) be a random variable with distribution \( F_X \). Show that for any \( a < b \in \mathbb{R} \),

- \( \mathbb{P}(a < X \leq b) = F_X(b) - F_X(a) \).
- \( \mathbb{P}(a < X < b) = F_X(b^-) - F_X(a) \).
- \( \mathbb{P}(X = x) = F_X(x) - F_X(x^-) \). Deduce that \( F_X \) is continuous at \( x \) if and only if \( \mathbb{P}(X = x) = 0 \).

**Solution.**

- Since \( (a,b] = (-\infty,b) \setminus (-\infty,a] \), and \( (-\infty,a] \subset (-\infty,b] \),
  \[ F_X(b) - F_X(a) = \mathbb{P}((-\infty,b]) - \mathbb{P}((-\infty,a]) = \mathbb{P}((a,b]) = \mathbb{P}(a < X \leq b). \]
- The events \( (a,b - 1/n] \) satisfy \( \lim_n (a,b - 1/n] = (a,b) \), and by continuity of probability
  \[ \mathbb{P}(a < X < b) = \mathbb{P}_X(\lim_n (a,b - 1/n]) = \lim_n \mathbb{P}_X(a,b - 1/n)) = F_X(b^-) - F_X(a). \]
- Similarly, \( \{x\} = \lim_n (x - 1/n,x] \) so
  \[ \mathbb{P}(X = x) = \mathbb{P}_X(\{x\}) = F_X(x) - \lim_n F_X(x - 1/n) = F_X(x) - F_X(x^-). \]
  So, \( F_X \) is left-continuous at \( x \) if and only if \( F_X(x^-) = F_X(x) \), which is if and
  only if \( \mathbb{P}(X = x) = 0 \).

Since segments of the form \((a,b], (a,b), (-\infty,a] \) and \((-\infty,a) \) are enough to generate
all sets in \( \mathcal{B}(\mathbb{R}) \) it can be shown that the distribution function uniquely determines the
probability measure of all Borel sets, and thus it determines the random variable \( X \).

In the same way as Lebesgue measure is shown to exist, it is shown in measure
theory that there is a 1-1 correspondence between distribution functions and probability
measures on \((\mathbb{R}, \mathcal{B}(\mathbb{R}))\). That is:
Theorem 7.15 (Lebesgue-Stieltjes). Every function $F : \mathbb{R} \to [0,1]$ that is right-continuous everywhere, non-decreasing and $\lim_{t \to \infty} F(t) = 1$, $\lim_{t \to -\infty} F(t) = 0$ gives rise to a unique probability measure $\mathbb{P}_F$ on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ satisfying $\mathbb{P}_F((a,b]) = F(b) - F(a)$ for all $a < b \in \mathbb{R}$. (That is, for such $F$ there is always a random variable $X$ such that $F = F_X$.)

Conversely, we have already seen that if $X$ is a random variable, then $\mathbb{P}_X((a,b]) = F_X(b) - F_X(a)$ for all $a < b \in \mathbb{R}$, and $F_X$ has the properties mentioned above.
8.1. Discrete Distributions

**Definition 8.1.** A random variable $X$ is called **discrete** if there exists a countable set $R \subset \mathbb{R}$ such that $\mathbb{P}(X \in R) = 1$.

If $X$ is such a random variable, then if we specify $\mathbb{P}(X = r)$ for all $r \in R$, then the distribution of $X$ can be calculated by

$$F_X(t) = \mathbb{P}(X \leq t) = \sum_{R \ni r \leq t} \mathbb{P}(X = r).$$

The set $\{r \in \mathbb{R} : \mathbb{P}(X = r) > 0\}$ is sometimes called the range of $X$, and the function $f_X(x) = \mathbb{P}(X = x)$ is called the density of $X$ (sometimes: probability density function). Note that

$$F_X(t) = \sum_{R \ni x \leq t} f_X(x).$$

8.1.1. **Bernoulli Distribution.** Let $0 < p < 1$. $X$ has the **Bernoulli-$p$** distribution, denoted $X \sim \text{Ber}(p)$ if:

$$F_X(t) = \begin{cases} 
0 & t < 0 \\
1 - p & 0 \leq t < 1 \\
1 & 1 \leq t.
\end{cases}$$

That is, $\mathbb{P}(X \in \{0, 1\}) = 1$, and $\mathbb{P}(X = 0) = 1 - p$, $\mathbb{P}(X = 1) = p$.

The Bernoulli distribution models a coin toss of a biased coin - probability $p$ to fall head (or 1) and $1 - p$ to fall tails (or 0).
8.1.2. **Binomial Distribution.** Suppose we have a biased coin with probability $p$ to fall heads (or 1). What if we toss that coin $n$ times in a row, all tosses mutually independent?

The sample space would be $\Omega = \{0, 1\}^n$. Because of independence the measure would be

$$P(\omega) = p \text{ number of ones in } \omega \cdot (1 - p) \text{ number of zeros in } \omega.$$  

So if $X : \Omega \to \mathbb{R}$ is the number of ones, then since for $0 \leq k \leq n$ there are $\binom{n}{k}$ different $\omega \in \Omega$ with exactly $k$ ones,

$$0 \leq k \leq n \quad f_X(k) = P(X = k) = \binom{n}{k} p^k (1 - p)^{n-k}.$$  

Such an $X$ is said to have **Binomial**-$(n, p)$ distribution, denoted $X \sim \text{Bin}(n, p)$.

What is $F_X$ in this case?

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

So $F_X(t) = 0$ for $t < 0$ and $F_X(t) = 1$ for $t \geq n$.

**Example 8.2.** A mathematician sits at the bar. For the next hour, every 5 minutes he orders a beer with probability $2/3$ and drinks it. With the remaining probability of $1/3$ he does nothing for those 5 minutes. All orders of beer are mutually independent.

What is the probability that he drinks no more than one beer, so he can still drive home (legally)?

**Solution.** If $X$ is the number of beers drunk, then $X \sim \text{Bin}(12, 2/3)$, since every beer is drunk with probability $2/3$ independently, in $12 = 60/5$ trials. So

$$P(X \leq 1) = P(X = 0) + P(X = 1) = \binom{12}{0} (2/3)^0 (1/3)^{12} + \binom{12}{1} (2/3)^1 (1/3)^{11}$$

$$= (1/3)^{12} + 12 \cdot 2 \cdot (1/3)^{12} = \frac{25}{3^{12}} < \frac{1}{20,000}.$$  

$\square$

Example 8.3. A student is watching television. Every minute she tosses a biased coin, if it come up heads she goes to study for the probability exam, if tails she continues watching TV. All coin tosses are mutually independent.

What is the probability she will continue watching TV forever?

What is the probability she watches for exactly \(k\) minutes and then starts studying?

The sample space here is \(\Omega = \mathbb{N} \cup \{\infty\}\).

Let \(T\) be the number of minutes it takes to go study. The event that \(T = k\) is the event that \(k - 1\) coins are tails, and the \(k\)-th coin is heads, where all coins are independent. So

\[
P(T = k) = (1 - p)^{k - 1}p.
\]

Note that

\[
P(T = \infty) = 1 - P(T < \infty) = 1 - \sum_{k=1}^{\infty} P(T = k) = 1 - p \sum_{k=1}^{\infty} (1 - p)^{k - 1} = 0.
\]

So the range of \(T\) is \(\mathbb{N}\).

The distribution of \(T\) above is called the Geometric-\(p\) distribution; that is, \(X\) has the Geometric-\(p\) distribution, denoted \(X \sim \text{Geo}(p)\), if

\[
f_X(k) = P(X = k) = \begin{cases} (1 - p)^{k - 1}p & k \in \{1, 2, \ldots, \} \\ 0 & \text{otherwise.} \end{cases}
\]

What is \(F_X\)?

\[
F_X(t) = \sum_{1 \leq k \leq t : k \in \mathbb{N}} (1 - p)^{k - 1}p = 1 - (1 - p)^{\lfloor t \rfloor}.
\]

So the Geometric distribution is the number of trials of independent identical experiments until the first success.

Example 8.4. An urn contains \(b\) black balls and \(w\) white balls. We randomly remove a ball, all balls equally likely, and then return that ball.

Let \(T\) be the number of trials until we see a black ball.

What is the distribution of \(T\)?
Well, every time we try there is independently a probability of $p := b/(b + w)$ to remove a black ball. So the distribution of $T$ is Geometric-$p$.

**Proposition 8.5** (Memoryless property for geometric distribution). Let $X$ be a Geometric-$p$ random variable. For any $k > m \geq 1$,

$$\mathbb{P}(X = k | X > m) = \mathbb{P}(X = k - m).$$

**Proof.** By definition,

$$\mathbb{P}(X = k | X > m) = \frac{\mathbb{P}(X = k, X > m)}{\mathbb{P}(X > m)} = \frac{(1-p)^{k-1}p}{(1-p)^m} = (1-p)^{k-m}p = \mathbb{P}(X = k - m).$$

That is, if we are given that we haven’t succeeded until time $m$, the probability of succeeding in another $k$ steps is the same as not succeeding for $k$ steps to begin with.

8.1.4. **Poisson Distribution.** A random variable $X$ has the Poisson-$\lambda$ distribution ($0 < \lambda \in \mathbb{R}$) if the range is $\mathbb{N}$ and

$$f_X(k) = \mathbb{P}(X = k) = e^{-\lambda} \cdot \frac{\lambda^k}{k!}.$$ 

Thus,

$$F_X(t) = \sum_{k=0}^{|t|} e^{-\lambda} \cdot \frac{\lambda^k}{k!}.$$ 

The importance of the Poisson distribution is in that it seems to occur naturally in many practical situations (the number of phone calls arriving at a call center per minute, the number of goals per game in football, the number of mutations in a given stretch of DNA after a certain amount of radiation, the number of requests sent to a printer from a network...).

The reason this may happen is that the Poisson serves is the limit of Binomial distributions when the number of trials goes to infinity: Suppose $X_n \sim \text{Bin}(n, \lambda/n)$. Then,

$$\mathbb{P}(X_n = k) = \binom{n}{k} \left(\frac{\lambda}{n}\right)^k \cdot \left(1 - \frac{\lambda}{n}\right)^{n-k}$$

$$= \frac{\lambda^k}{k!} \cdot (1 - \frac{1}{n}) \cdot (1 - \frac{2}{n}) \cdots (1 - \frac{k-1}{n}) \cdot (1 - \frac{\lambda}{n})^{n-k} \to e^{-\lambda} \frac{\lambda^k}{k!}.$$
So Poisson is like doing infinitely many trials of an experiment that has small probability of success, and asking how many successes are observed.

**Example 8.6.** Suppose the number of people at the checkout counter at the supermarket has Poisson distribution. Checkout A has Poisson-5 and Checkout B has Poisson-2. Both counters are independent.

Noga arrives at the counters and chooses the shorter line (and chooses A if they are tied). What is the probability she chooses B?

What is the probability both checkouts are empty?

**Solution.** Let \( X \) be the number of people at A, and \( Y \) the number at B. We want the probability of \( X > Y \).

The information we have is that

\[
P(Y = k, X = m) = P(Y = k) P(X = m),
\]

for all \( k, m \in \mathbb{N} \).

So

\[
P(Y < X) = \sum_{m=0}^{\infty} P(X = m) \sum_{k=m}^{\infty} P(Y = k) = \sum_{m=1}^{\infty} e^{-5} \frac{5^m}{m!} \sum_{k=0}^{m-1} e^{-2} \frac{2^k}{k!}.
\]

The probability that both checkouts are empty is

\[
P(X = 0, Y = 0) = P(X = 0) P(Y = 0) = e^{-5} e^{-2} = e^{-7}.
\]

\[\square\]

8.1.5. **Hypergeometric.** Suppose we have an urn with \( N \geq 1 \) balls, \( 0 \leq m \leq N \) are black and the rest are white. If we choose \( 1 \leq n \leq N \) balls from the urn randomly and uniformly, what is the number of black balls we get?

The space here is uniform measure on

\[
\Omega = \{ \omega \subset \{1, \ldots, N\} : |\omega| = n \} = \binom{\{1, \ldots, N\}}{n}.
\]

We ask for the random variable \( X(\omega) := |\omega \cap \{1, \ldots, m\}|. \)
A simple calculation reveals that for any \(0 \leq k \leq \min\{n, m\},\)

\[
P[X = k] = \frac{{m \choose k} \cdot \left(\frac{N-m}{n-k}\right)}{\left(\frac{N}{n}\right)}.
\]

(We interpret \(\binom{a}{b}\) for \(b > a\) as 0, so actually, \(\max\{0, m + n - N\} \leq k \leq \min\{m, n\}\).

Such a random variable \(X\) is said to have hypergeometric distribution with parameters \(N, m, n\), and we write \(X \sim H(N, m, n)\).

**Example 8.7.** Chagai chooses 5 cards from a deck, all possibilities equal. How much more likely is it for him to choose 3 aces than it is to choose 4 aces? How much more likely is it to choose 4 diamonds than it is to choose 5 diamonds?

Let \(X\) = the number of aces chosen. The size of the deck is \(N = 52\), the number of cards chosen is \(n = 5\) and the number of “special” cards is \(m = 4\). So

\[
\frac{P[X = 3]}{P[X = 4]} = \frac{{m \choose 3} \cdot \left(\frac{N-m}{n-3}\right)}{{m \choose 4} \cdot \left(\frac{N-m}{n-4}\right)} = \frac{4 \cdot (N-m-n+4)}{(m-3) \cdot (n-3)} = \frac{4 \cdot 47}{1 \cdot 2} = 94.
\]

For \(Y\) = the number of diamonds chosen, we get \(N = 52, n = 5, m = 13\). So

\[
\frac{P[X = 4]}{P[X = 5]} = \frac{{m \choose 4} \cdot \left(\frac{N-m}{n-4}\right)}{{m \choose 5} \cdot \left(\frac{N-m}{n-5}\right)} = \frac{5 \cdot (N-m-n+5)}{(m-4) \cdot (n-4)} = \frac{5 \cdot 39}{9 \cdot 1} = \frac{5 \cdot 13}{3} = 21 + \frac{2}{3}.
\]

\[\triangle\nabla\triangle\]

8.1.6. **Negative Binomial.** We say that \(X \sim NB(m, p)\), or \(X\) has negative binomial distribution (or Pascal distribution) if \(X\) is the number of trials until the \(m\)-th success, with each success with probability \(p\).

The event \(X = k\) is the event that there are \(m - 1\) successes in the first \(k - 1\) trials, and then one success in the \(k\)-th trial. So

\[
P[X = k] = \binom{k-1}{m-1} p^m(1-p)^{(k-1)-(m-1)} = \binom{k-1}{m-1} p^m(1-p)^{k-m},
\]

for \(k \geq m\).

**Example 8.8.** Let \(X \sim NB(1, p)\). Then \(X \sim Geo(p)\). Indeed, for any \(k \geq 1\),

\[
P[X = k] = \binom{k-1}{0} p(1-p)^{k-1} = (1-p)^{k-1}p.
\]

\[\triangle\nabla\triangle\]
9.1. CONTINUOUS RANDOM VARIABLES

**Definition 9.1.** Let \((\Omega, \mathcal{F}, \mathbb{P})\) be a probability space. A random variable \(X : \Omega \to \mathbb{R}\) is called **continuous** if the distribution function \(F_X\) is continuous at all points.

**Remark 9.2.** Recall that we showed that for all \(x \in \mathbb{R}\),

\[
\mathbb{P}(X = x) = \lim_{n \to \infty} \mathbb{P}_X((x - 1/n, x]) = F_X(x) - \lim_{n \to \infty} F_X(x - 1/n) = F_X(x) - F_X(x^-).
\]

Thus, if \(F_X\) is continuous at \(x\), then \(\mathbb{P}(X = x) = 0\).

So, if \(X\) is a continuous random variable then \(\mathbb{P}(X = x) = 0\) for all \(x \in \mathbb{R}\).

This is very different than in the discrete case!

Continuous random variables can be very pathological, and we need measure theory to deal with some aspects regarding these. We restrict ourselves for this course to a “nicer” family of random variables.

**Definition 9.3.** Let \((\Omega, \mathcal{F}, \mathbb{P})\) be a probability space. A random variable \(X : \Omega \to \mathbb{R}\) is called **absolutely continuous** if there exists an integrable non-negative function \(f_X : \mathbb{R} \to \mathbb{R}^+\) such that for all \(x \in \mathbb{R}\),

\[
\mathbb{P}(X \leq x) = F_X(x) = \int_{-\infty}^{x} f_X(t)dt.
\]

Such a function \(f_X\) is called the **probability density function** (or just density function, or PDF) of \(X\).

**Remark 9.4.** Recall that the fundamental theorem of calculus says that if \(f_X\) is continuous, then \(F_X\) is differentiable and \(F'_X = f\).
Also, a result of calculus tells us that the function $x \mapsto \int_{-\infty}^{x} f(t)dt$ is a continuous function of $x$. So any absolutely continuous random variable is also continuous.

✓ Under the carpet: We don’t know how to prove that an integrable function $f_X : \mathbb{R} \to \mathbb{R}^+$ uniquely defines a probability measure $P_X$ on $(\mathbb{R}, \mathcal{B})$, since we would need some measure theory for this (actually we could use the Lebesgue-Steiltjes Theorem). However, given that such a measure is uniquely defined, we have the following properties:

**Proposition 9.5.** Let $X$ be an absolutely continuous random variable with PDF $f_X$.

1. By definition,
   $$P_X((a, b]) = F_X(b) - F_X(a) = \int_{-\infty}^{b} f_X(t)dt - \int_{-\infty}^{a} f_X(t)dt = \int_{a}^{b} f_X(t)dt.$$  

2. We don’t always know which sets we can integrate over, but if the integral exists we have
   $$P_X(A) = \int_{A} f_X(t)dt.$$  
   For example, if $A = \bigcup_{j=1}^{n} (a_j, b_j]$ this holds.

3. Continuity of probability gives us that $P_X(\{a\}) = 0$ for all $a$ (because $X$ is a continuous random variable, so $F_X$ is continuous at all points). Thus,
   $$P_X((a, b]) = P_X([a, b)) = P_X((a, b)) = P([a, b]).$$

4. We always have
   $$\int_{\mathbb{R}} f_X(t)dt = P_X(\mathbb{R}) = 1.$$  

**Proof.** Just calculus of the Riemann integral, and continuity of probability measures.

Example 9.6. Let $X$ be a random variable with PDF

$$f_X(t) = \begin{cases} 
C(2t - t^2) & 0 \leq t \leq 2 \\
0 & \text{otherwise.} 
\end{cases}$$

What is $C$? What is the probability that $X > 1$?
To compute $C$ we use the fact that $\int_{\mathbb{R}} f_X(t)dt = 1$. So

$$1 = \int_{-\infty}^{\infty} f_X(t)dt = C \int_{0}^{2} (2t - t^2)dt = C\left(2 - \frac{2}{3}\right) = C\cdot \frac{4}{3}.$$ 

So $C = 3/4$.

The probability that $X > 1$ is

$$P_X((1, \infty)) = 1 - P_X((-\infty, 1]) = 1 - \int_{0}^{1} C(2t - t^2)dt = 1 - \frac{3}{4} \cdot \frac{2}{3} = \frac{1}{2}. \triangle \nabla \triangle$$

9.1.1. **Uniform Distribution.** An absolutely continuous random variable $X$ has the uniform distribution on $[a, b]$, denoted $X \sim U([a, b])$ if $X$ has PDF

$$f_X(t) = \begin{cases} \frac{1}{b-a} & a \leq t \leq b \\ 0 & \text{otherwise.} \end{cases}$$

Indeed $\int_{\mathbb{R}} f_X(t)dt = 1$.

**Example 9.7.** A bus arrives at the station every 15 minutes, starting at 5:00. A passenger arrives between 7:00 and 7:30, uniformly distributed. What is the probability that he does not wait more than 5 minutes for a bus?

If we take $X =$ the time the passenger arrives, then $X$ has uniform distribution on $[0, 30]$. Also, the event that the passenger waits at most 5 minutes is the disjoint union

$$\{X = 0\} \uplus \{10 \leq X \leq 15\} \uplus \{25 \leq X \leq 30\}.$$

Thus the probability is

$$P(X = 0) + P(10 \leq X \leq 15) + P(25 \leq X \leq 30) = 0 + \frac{5}{30} + \frac{5}{30} = \frac{1}{3}. \triangle \nabla \triangle$$

The uniform distribution is very symmetric in the following sense:

**Exercise 9.8.** Let $X \sim U([a, b])$. Show that $Y = X + c$ has the uniform distribution on $[a + c, b + c]$. 


Solution. We need to show that

$$F_Y(x) = \mathbb{P}(\omega : Y(\omega) \leq x) = \int_{-\infty}^{x} \frac{1}{b+c-(a+c)} \cdot 1_{[a+c,b+c]} dt.$$ 

Indeed, since \(\{\omega : Y(\omega) \leq x\} = \{\omega : X(\omega) \leq x - c\}\), we have that

$$F_Y(x) = F_X(x-c) = \int_{-\infty}^{x-c} \frac{1}{b-a} \cdot 1_{[a,b]} dt = \frac{1}{b+c-(a+c)} \cdot \int_{-\infty}^{x} 1_{[a+c,b+c]} du.$$ 

\(\Box\)

**Exercise 9.9.** What is the distribution function of a \(U([a,b])\) random variable?

Proof. For \(a \leq x \leq b\) we have

$$F_X(x) = \frac{1}{b-a} \int_{a}^{x} dt = \frac{x-a}{b-a}.$$ 

For \(x < a\), \(F_X(x) = 0\) and for \(x > b\), \(F_X(x) = 1\).  \(\Box\)

9.1.2. **Exponential Distribution.** Let \(0 < \lambda \in \mathbb{R}\). An absolutely continuous random variable \(X\) is said to have **exponential distribution** of parameter \(\lambda\), denoted \(X \sim \text{Exp}(\lambda)\), if it has PDF

$$f_X(t) = 1_{\{t \geq 0\}} \cdot \lambda e^{-\lambda t}.$$ 

What is the distribution function in this case?

$$F_X(x) = \lambda \int_{0}^{x} e^{-\lambda t} dt = (-e^{-\lambda t})|_{0}^{x} = 1 - e^{-\lambda x},$$ 

for \(x \geq 0\). For \(x < 0\), \(F_x(x) = 0\).

**Example 9.10.** The duration of the wait in line at Kupat Cholim is exponentially distributed with parameter \(1/5\). What is the probability that the wait is longer than 10 minutes?

Suppose we are given that the wait is longer than 20 minutes. What is the probability given this information that the wait is longer than 30 minutes?

Let \(X = \text{waiting time}\). So \(X \sim \text{Exp}(1/5)\).

$$\mathbb{P}(X > 10) = \int_{10}^{\infty} \frac{1}{5} e^{-t/5} dt = (-e^{-t/5})|_{10}^{\infty} = e^{-2}.$$
Now, using the definition of conditional probability,

\[ P(X > 30 | X > 20) = \frac{P(X > 30)}{P(X > 20)}. \]

Since

\[ P(X > x) = (e^{-t/5})^\infty = e^{-x/5}, \]

we get that

\[ P(X > 30 | X > 20) = e^{-6}/e^{-4} = e^{-2}. \]

△▽△

Is it a coincidence that we got the same answer? No. Recall that the geometric distribution had the memoryless property, that is a geometric random variable \( X \) satisfies

\[ P(X > m + n | X > m) = P(X > n). \]

Similarly, the exponential distribution has the memoryless property:

**Proposition 9.11.** Let \( X \sim \text{Exp}(\lambda) \). Then for any \( x, y \in \mathbb{R}^+ \),

\[ P(X > x + y | X > x) = P(X > y). \]

That is, the knowledge of waiting for some time does not give information about how much more we have to wait.

**Proof.** First,

\[ P(X > t) = 1 - F_X(t) = e^{-\lambda t}. \]

So

\[ P(X > x + y | X > x) = \frac{P(X > x + y)}{P(X > x)} = \frac{e^{-\lambda(x+y)}}{e^{-\lambda x}} = e^{-\lambda y} = P(X > y). \]

\[ \square \]

However, it turns out that the exponential distribution is the *only* continuous random variable with the memoryless property:

**Proposition 9.12.** Let \( X \) be a continuous random variable, such that for all \( x, y \in \mathbb{R}^+ \),

\[ P(X > x + y | X > x) = P(X > y). \]

Then, \( X \sim \text{Exp}(\lambda) \) for some \( \lambda > 0 \).
Proof. \( X \) is continuous, so that means \( F_X \) is a continuous function. Let \( g(t) = 1 - F_X(t) = P(X > t) \). So \( g \) is a continuous function as well.

Now, for any \( x, y \in \mathbb{R}^+ \),

\[
g(x + y) = g(x)g(y).
\]

Let \( \lambda \geq 0 \) be such that \( g(1) = e^{-\lambda} \) (recall that \( g : \mathbb{R} \to [0, 1] \)). For any \( 0 < n \in \mathbb{N} \) we have that

\[
g(n) = g(n - 1)g(1) = \cdots = g(1)^n.
\]

Since \( g(n) \to 0 \) we have that \( g(1) < 1 \) and \( \lambda > 0 \).

For any \( 0 < m \in \mathbb{N} \) we have that

\[
g(1) = g\left(\frac{1}{m}\right)g(1 - \frac{1}{m}) = \cdots = g\left(\frac{1}{m}\right)^m,
\]

so \( g(1/m) = g(1)^{1/m} \). In a similar way, we get that for any rational number

\[
g\left(\frac{n}{m}\right) = g(1)^{n/m} = e^{-\lambda n/m}.
\]

Since \( g \) is continuous and coincides with \( e^{-\lambda t} \) for all \( q \in \mathbb{Q} \), we get that \( g(t) = e^{-\lambda t} \).

9.1.3. Normal Distribution.

**Some calculus.** Let us look at the following integral:

\[
I = \int_{-\infty}^{\infty} e^{-t^2} dt.
\]

How can we compute this value?

We use two-dimensional calculus to compute it: Note that

\[
I^2 = \int_{\mathbb{R}^2} e^{-(x^2+y^2)} dxdy.
\]

Setting \( x = r \cos \theta \) and \( y = r \sin \theta \) (or \( r = \sqrt{x^2 + y^2} \) and \( \theta = \arctan(y/x) \)) so the Jacobian is

\[
J = \begin{bmatrix}
\frac{dx}{dr} & \frac{dx}{d\theta} \\
\frac{dy}{dr} & \frac{dy}{d\theta}
\end{bmatrix} = \begin{bmatrix}
\cos \theta & -r \sin \theta \\
\sin \theta & r \cos \theta
\end{bmatrix}
\]

So \( \det(J) = r \). Thus,

\[
I^2 = \int_0^{2\pi} \int_0^{\infty} e^{-r^2} r dr d\theta.
\]
Substituting \( s = r^2 \) so that \( ds = 2rdr \), we have that

\[
I^2 = \int_0^{2\pi} \int_0^\infty \frac{1}{2} e^{-s} ds d\theta = \pi.
\]

So \( I = \sqrt{\pi} \).

Now, we can use another trick to compute

\[
I = \int_{-\infty}^\infty \exp\left( -\frac{(t - \mu)^2}{2\sigma^2} \right) dt.
\]

Letting \( x = \frac{t - \mu}{\sqrt{2}\sigma} \) we have that \( dt = \sqrt{2}\sigma dx \). Thus,

\[
I = \sqrt{2}\sigma \int_{-\infty}^\infty e^{-x^2} dx = \sqrt{2\pi}\sigma.
\]

We are ready to define the normal distribution:

Let \( \mu \in \mathbb{R} \) and \( \sigma \in \mathbb{R}^+ \). An absolutely continuous random variable has normal-\((\mu, \sigma)\) distribution, denoted \( X \sim N(\mu, \sigma) \), if it has PDF

\[
f_X(t) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left( -\frac{(t - \mu)^2}{2\sigma^2} \right).
\]

A normal-(0, 1) random variable is called a standard normal random variable.

**Example 9.13.** The distribution of the lifetime of people in Oceania is distributed as a normal-(75, 8) random variable. What is the probability a person lives more than 100 years? Show that it is less than \( e^{-\xi^2/2} \cdot \frac{1}{\sqrt{2\pi}} \), where \( \xi = 25/8 \).

If \( X = \text{lifetime} \), then \( X \sim N(75, 8) \). So

\[
P(X > 100) = \int_{100}^{\infty} \frac{1}{\sqrt{2\pi} \cdot 8} \cdot e^{(t-75)^2/(2\cdot8^2)} dt.
\]

Taking \( s = (t - 75)/8 \) we have that \( dt = 8ds \) so

\[
P(X > 100) = \int_{\xi}^{\infty} \frac{1}{\sqrt{2\pi}} e^{s^2/2} ds.
\]

Since on \((\xi, \infty)\) we have that \( 1 < \frac{s}{\xi} \) we get that

\[
P(X > 100) < \int_{\xi}^{\infty} \frac{1}{\sqrt{2\pi}} se^{-s^2/2} ds.
\]

Substituting \( u = s^2/2 \), so \( du = sds \) we have that

\[
P(X > 100) < \int_{\xi^2/2}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-u} du = e^{-\xi^2/2} \cdot \frac{1}{\sqrt{2\pi}}.
\]
A nice property of the normal distribution is the following:

**Exercise 9.14.** Let $X \sim N(\mu, \sigma)$. Define $Y(\omega) = aX(\omega) + b$. Show that $Y \sim N(a\mu + b, a\sigma)$.

**Solution.** We need to prove that for any $y \in \mathbb{R}$,

$$P(Y \leq y) = \int_{-\infty}^{y} \frac{1}{\sqrt{2\pi} \cdot a\sigma} \exp \left( -\frac{(t - a\mu - b)^2}{2a^2\sigma^2} \right) dt.$$

Since

$$\{\omega : Y(\omega) \leq y\} = \{\omega : X(\omega) \leq (y - b)/a\},$$

we have that

$$P(Y \leq y) = P(X \leq (y - b)/a) = \int_{-\infty}^{(y-b)/a} \frac{1}{\sqrt{2\pi} \cdot \sigma} \exp \left( -\frac{(t - \mu)^2}{2\sigma^2} \right) dt.$$

Substituting $t - \mu = \frac{s - a\mu - b}{a}$ (or $t = \frac{s - b}{a}$), we have that $dt = \frac{1}{a} ds$ and

$$P(Y \leq y) = \int_{-\infty}^{y} \frac{1}{\sqrt{2\pi} \cdot a\sigma} \exp \left( -\frac{(s - a\mu - b)^2}{2a^2\sigma^2} \right) ds.$$

\[\Box\]

**Corollary 9.15.** If $X$ is a normal-$(\mu, \sigma)$ random variable, then $\frac{X - \mu}{\sigma}$ is a standard normal random variable.

**Example 9.16.** Let $X \sim N(\mu, \sigma)$. Show that $F_X(t) > 0$ for all $t$.

Indeed,

$$F_X(t) = \int_{-\infty}^{t} f_X(s) ds \geq \int_{t-1}^{t} f_X(s) ds \geq \inf_{s \in [t-1,t]} f_X(s).$$

Since $f_X$ is increasing up to $\mu$ and decreasing after $\mu$, we have that the infimum on $[t-1, t]$ is obtained at the endpoints. So

$$F_X(t) \geq \min \{ f_X(t - 1), f_X(t) \} > 0.$$  

\[\Box\]
Example 9.17. Let $X \sim N(0,1)$. Show that $X^2$ is absolutely continuous. Is $X^2$ normally distributed?

First, to show that $X^2$ is absolutely continuous we need to show that

$$F_{X^2}(t) = \int_{-\infty}^{t} f(s)ds$$

for some function $f : \mathbb{R} \to \mathbb{R}^+$. Indeed, if $t < 0$ then $\{X^2 \leq t\} = \emptyset$ so $F_{X^2}(t) = 0$.

If $t \geq 0$,

$$F_{X^2}(t) = \mathbb{P}[X^2 \leq t] = \mathbb{P}[-\sqrt{t} \leq X \leq \sqrt{t}] = \int_{-\sqrt{t}}^{\sqrt{t}} \frac{1}{\sqrt{2\pi}} e^{-s^2/2}ds.$$

Since this last function in the integrand is even (symmetric around 0), we have that

$$F_{X^2}(t) = 2 \int_{0}^{\sqrt{t}} \frac{1}{\sqrt{2\pi}} e^{-s^2/2}ds = \int_{0}^{t} \frac{1}{\sqrt{2\pi x}} e^{-x/2}dx,$$

where we have used the change of variables $x = s^2$ so $ds = \frac{1}{2\sqrt{x}}dx$.

Now, if we define

$$f(s) = \begin{cases} \frac{1}{\sqrt{2\pi s}} e^{-s/2} & \text{if } s > 0 \\ 0 & \text{if } s \leq 0, \end{cases}$$

then we get from all the above that

$$F_{X^2}(t) = \int_{-\infty}^{t} f(s)ds.$$

So $X^2$ is absolutely continuous with density $f = f_X$.

Is $X^2$ normal? No. This can be seen since $F_{X^2}(0) = 0$ but $F_{N}(0) > 0$ for any $N \sim N(\mu, \sigma)$. △ ▽ △
10.1. Independence Revisited

10.1.1. Some reminders. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. Given a collection of subsets $\mathcal{K} \subset \mathcal{F}$, recall that the $\sigma$-algebra generated by $\mathcal{K}$, is

$$\sigma(\mathcal{K}) = \bigcap \{ \mathcal{G} : \mathcal{G} \text{ is a } \sigma\text{-algebra }, \mathcal{K} \subset \mathcal{G} \},$$

and this $\sigma$-algebra is the smallest $\sigma$-algebra containing $\mathcal{K}$. $\sigma(\mathcal{K})$ can be thought of as all possible information that can be generated by the sets in $\mathcal{K}$.

Recall that a collection of events $(A_n)$ are mutually independent if for any finite number of these events $A_1, \ldots, A_n$ we have that

$$\mathbb{P}(A_1 \cap \cdots \cap A_n) = \mathbb{P}(A_1) \cdots \mathbb{P}(A_n).$$

We can also define the independence of families of events:

**Definition 10.1.** Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. Let $(\mathcal{K}_n)_n$ be a collection of families of events in $\mathcal{F}$. Then, $(\mathcal{K}_n)_n$ are **mutually independent** if for any finite number of families from this collection, $\mathcal{K}_1, \ldots, \mathcal{K}_n$ and any events $A_1 \in \mathcal{K}_1, A_2 \in \mathcal{K}_2, \ldots, A_n \in \mathcal{K}_n$, we have that

$$\mathbb{P}(A_1 \cap \cdots \cap A_n) = \mathbb{P}(A_1) \cdots \mathbb{P}(A_n).$$

For example, we would like to think that $(A_n)$ are mutually independent, if all the information from each event in the sequence is independent from the information from the other events in the sequence. This is the content of the next proposition.
Proposition 10.2. Let \((A_n)\) be a sequence of events on a probability space \((\Omega, \mathcal{F}, P)\). Then, \((A_n)\) are mutually independent if and only if the \(\sigma\)-algebras \((\sigma(A_n))_n\) are mutually independent.

It is not difficult to prove this by induction:

Proof. By induction on \(n\) we show that \((\sigma(A_1), \ldots, \sigma(A_n), \{A_{n+1}\}, \ldots)\) are mutually independent.

The base \(n = 0\) is just the assumption.

So assume that \((\sigma(A_1), \ldots, \sigma(A_{n-1}), \{A_n\}, \{A_{n+1}\}, \ldots)\) are mutually independent. Let \(n_1 < n_2 < \ldots < n_k < n_{k+1} < \ldots < n_m\) be a finite number of indices such that \(n_k < n < n_{k+1}\). Let \(B_j \in \sigma(A_{n_j})\) for \(j = 1, \ldots, k\), and \(B_j = A_{n_j}\) for \(j = k+1, \ldots, m\).

Let \(B \in \sigma(A_n)\). If \(B = \emptyset\) then \(P(B_1 \cap \cdots \cap B_n \cap B) = 0 = P(B_1) \cdots P(B_n) \cdot P(B)\). If \(B = \Omega\) then \(P(B_1 \cap \cdots \cap B_n \cap B) = P(B_1) \cdots P(B_n) = P(B_1) \cdots P(B_n)\), by the induction hypotheses. If \(B = A_n\) then this also holds by the induction hypotheses. So we only have to deal with the case \(B = A_n^c\). In this case, for \(X = B_1 \cap \cdots \cap B_n\),

\[
P(X \cap B) = P(X \setminus (X \cap A_n)) = P(X) - P(X \cap A_n) = P(X)(1 - P(A_n)) = P(X)P(B),
\]

where we have used the induction hypotheses to say that \(X\) and \(A_n\) are independent.

Since this holds for any choice \(a\) of a finite number of events, we have the induction step. \(\square\)

However, we will take a more windy road to get stronger results... (Corollary 10.8)

10.2. \(\pi\)-systems and Independence

Definition 10.3. Let \(\mathcal{K}\) be a family of subsets of \(\Omega\).

- We say that \(\mathcal{K}\) is a \(\pi\)-system if \(\emptyset \in \mathcal{K}\) and \(\mathcal{K}\) is closed under intersections; that is, for all \(A, B \in \mathcal{K}\), \(A \cap B \in \mathcal{K}\).
- We say that \(\mathcal{K}\) is a Dynkin system (or \(\lambda\)-system) if \(\mathcal{K}\) is closed under set complements and countable disjoint unions; that is, for any \(A \in \mathcal{K}\) and any sequence \((A_n)\) in \(\mathcal{K}\) of pairwise disjoint subsets, we have that \(A^c \in \mathcal{K}\) and \(\biguplus_n A_n \in \mathcal{K}\).
The main goal now is to show that probability measures are uniquely defined once they are defined on a \( \pi \)-system. This is the content of Theorem 10.7.

**Proposition 10.4.** If \( \mathcal{F} \) is Dynkin system on \( \Omega \) and \( \mathcal{F} \) is also a \( \pi \)-system on \( \Omega \), then \( \mathcal{F} \) is a \( \sigma \)-algebra.

**Proof.** Since \( \mathcal{F} \) is a Dynkin system, it is closed under complements, and \( \Omega = \emptyset^c \in \mathcal{F} \).

Let \( (A_n) \) be a sequence of subsets in \( \mathcal{F} \). Set \( B_1 = A_1, C_1 = \emptyset \), and for \( n > 1 \),

\[
C_n = \bigcup_{j=1}^{n-1} A_j \quad \text{and} \quad B_n = A_n \setminus C_n.
\]

Since \( \mathcal{F} \) is a \( \pi \)-system and closed under complements, \( C_n = (\bigcap_{j=1}^{n-1} A_j^c)^c \in \mathcal{F} \), and so \( B_n = A_n \cap C_n^c \in \mathcal{F} \). Since \( (B_n) \) is a sequence of pairwise disjoint sets, and since \( \mathcal{F} \) is a Dynkin system,

\[
\bigcup_n A_n = \bigcup_n B_n \in \mathcal{F}.
\]

\( \square \)

**Proposition 10.5.** If \( (\mathcal{D}_\alpha)_\alpha \) is a collection of Dynkin systems (not necessarily countable). Then \( \mathcal{D} = \bigcap_\alpha \mathcal{D}_\alpha \) is a Dynkin system.

**Proof.** Since \( \emptyset \in \mathcal{D}_\alpha \) for all \( \alpha \), we have that \( \emptyset \in \mathcal{D} \).

If \( A \in \mathcal{D} \), then \( A \in \mathcal{D}_\alpha \) for all \( \alpha \). So \( A^c \in \mathcal{D}_\alpha \) for all \( \alpha \), and thus \( A^c \in \mathcal{D} \).

If \( (A_n)_n \) is a countable sequence of pairwise disjoint sets in \( \mathcal{D} \), then \( A_n \in \mathcal{D}_\alpha \) for all \( \alpha \) and all \( n \). Thus, for any \( \alpha \), \( \bigcup_n A_n \in \mathcal{D}_\alpha \). So \( \bigcup_n A_n \in \mathcal{D} \).

\( \square \)

**Lemma 10.6** (Dynkin’s Lemma). If a Dynkin system \( \mathcal{D} \) contains a \( \pi \)-system \( \mathcal{K} \), then \( \sigma(\mathcal{K}) \subset \mathcal{D} \).

**Proof.** Let

\[
\mathcal{F} = \bigcap \{ \mathcal{D}' : \mathcal{D}' \text{ is a Dynkin system containing } \mathcal{K} \}.
\]

So \( \mathcal{F} \) is a Dynkin system and \( \mathcal{K} \subset \mathcal{F} \subset \mathcal{D} \). We will show that \( \mathcal{F} \) is a \( \sigma \)-algebra, so \( \sigma(\mathcal{K}) \subset \mathcal{F} \subset \mathcal{D} \).
Suppose we know that $\mathcal{F}$ is closed under intersections (which is Claim 3 below). Since $\emptyset \in \mathcal{K} \subset \mathcal{F}$, we will then have that $\mathcal{F}$ is a $\pi$-system. Being both a Dynkin system and a $\pi$–system, $\mathcal{F}$ is a $\sigma$–algebra.

Thus, to show that $\mathcal{F}$ is a $\sigma$–algebra, it suffices to show that $\mathcal{F}$ is closed under intersections.

Note that $\mathcal{F}$ is closed under complements (because all Dynkin systems are).

**Claim 1.** If $A \subset B$ are subsets in $\mathcal{F}$, then $B \setminus A \in \mathcal{F}$.

**Proof.** If $A, B \in \mathcal{F}$, then since $\mathcal{F}$ is a Dynkin system, also $B^c \in \mathcal{F}$. Since $A \subset B$, we have that $A, B^c$ are disjoint, so $A \cup B^c \in \mathcal{F}$ and so $B \setminus A = (A \cap B) = (A \cup B^c)^c \in \mathcal{F}$. $\Box$

**Claim 2.** For any $K \in \mathcal{K}$, if $A \in \mathcal{F}$ then $A \cap K \in \mathcal{F}$.

**Proof.** Let $E = \{ A : A \cap K \in \mathcal{F} \}$.

Let $A \in E$ and $(A_n)$ be a sequence of pairwise disjoint subsets in $E$.

Since $K \in \mathcal{F}$ and $A \cap K \in \mathcal{F}$, by Claim 1 we have that $A^c \cap K = K \setminus (A \cap K) \in \mathcal{F}$. So $A^c \in E$.

Since $(A_n \cap K)_n$ is a sequence of pairwise disjoint subsets in $\mathcal{F}$, we get that

$$\bigcup_n A_n \cap K = \bigcup_n (A_n \cap K) \in \mathcal{F}.$$

So we conclude that $E$ is a Dynkin system. Since $\mathcal{K}$ is closed under intersections, $E$ contains $\mathcal{K}$. Thus, by definition $\mathcal{F} \subset E$.

So for any $A \in \mathcal{F}$ we have that $A \in E$, and $A \cap K \in \mathcal{F}$. $\Box$

**Claim 3.** For any $B \in \mathcal{F}$, if $A \in \mathcal{F}$ then $A \cap B \in \mathcal{F}$.

**Proof.** Let $E = \{ A : A \cap B \in \mathcal{F} \}$.

Let $A \in E$ and $(A_n)$ be a sequence of pairwise disjoint subsets in $E$.

Since $B \in \mathcal{F}$ and $A \cap B \in \mathcal{F}$, by Claim 1 we have that $A^c \cap B = B \setminus (A \cap B) \in \mathcal{F}$. So $A^c \in E$.

Since $(A_n \cap B)_n$ is a sequence of pairwise disjoint subsets in $\mathcal{F}$, we get that

$$\bigcup_n A_n \cap B = \bigcup_n (A_n \cap B) \in \mathcal{F}.$$
So we conclude that $E$ is a Dynkin system. By Claim 2, $\mathcal{K}$ is contained in $E$. So by definition, $\mathcal{F} \subset E$. 

Since $\mathcal{F}$ is closed under intersections, this completes the proof. 

The next theorem tells us that a probability measure on $(\Omega, \mathcal{F})$ is determined by it’s values on a $\pi$-system generating $\mathcal{F}$.

**Theorem 10.7** (Uniqueness of Extension). Let $\mathcal{K}$ be a $\pi$-system on $\Omega$, and let $\mathcal{F} = \sigma(\mathcal{K})$ be the $\sigma$-algebra generated by $\mathcal{K}$. Let $P, Q$ be two probability measures on $(\Omega, \mathcal{F})$, such that for all $A \in \mathcal{K}$, $P(A) = Q(A)$. Then, $P(B) = Q(B)$ for any $B \in \mathcal{F}$.

**Proof.** Let $\mathcal{D} = \{A \in \mathcal{F} : P(A) = Q(A)\}$. So $\mathcal{K} \subset \mathcal{D}$. We will show that $\mathcal{D}$ is a Dynkin system, and since it contains $\mathcal{K}$, the by Dynkin’s Lemma it must contain $\mathcal{F} = \sigma(\mathcal{K})$.

If $A \in \mathcal{D}$, then $P(A^c) = 1 - P(A) = 1 - Q(A) = Q(A^c)$, so $A^c \in \mathcal{D}$.

Let $(A_n)$ be a sequence of pairwise disjoint sets in $\mathcal{D}$. Then,

$$P(\bigcup_n A_n) = \sum_n P(A_n) = \sum_n Q(A_n) = Q(\bigcup_n A_n).$$

So $\bigcup_n A_n \in \mathcal{D}$. 

**Corollary 10.8.** Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. Let $(\Pi_n)_n$ be a sequence of $\pi$-systems, and let $\mathcal{F}_n = \sigma(\Pi_n)$. Then, $(\Pi_n)_n$ are mutually independent if and only if $(\mathcal{F}_n)_n$ are mutually independent.

**Proof.** We will prove by induction on $n$ that for any $n \geq 0$, the collection $(\mathcal{F}_1, \mathcal{F}_2, \ldots, \mathcal{F}_n, \Pi_{n+1}, \Pi_{n+2}, \ldots)$ are mutually independent.

For $n = 0$ this is the assumption. For $n > 1$, let $n_1 < n_2 < \ldots < n_k < n_{k+1} < \ldots < n_m$ be a finite number of indices such that $n_k < n < n_{k+1}$. Let $A_j \in \mathcal{F}_{n_j}, j = 1, \ldots, k$ and $A_j \in \Pi_{n_j}, j = k + 1, \ldots, m$.

For any $A \in \mathcal{F}_n$, if $\mathbb{P}(A_1 \cap \cdots \cap A_m) = 0$ then $A$ is independent of $A_1 \cap \cdots \cap A_m$. So assume that $\mathbb{P}(A_1 \cap \cdots \cap A_m) > 0$.

For any $A \in \mathcal{F}_n$ define the probability measure

$$P(A) := \mathbb{P}(A|A_1 \cap \cdots \cap A_m).$$
By induction, the collection \((\mathcal{F}_1, \ldots, \mathcal{F}_{n-1}, \Pi_n, \Pi_{n+1}, \ldots)\) are mutually independent, so \(P(A) = P(A)\) for any \(A \in \Pi_n\). Since \(\Pi_n\) is a \(\pi\)-system generating \(\mathcal{F}_n\), we have by Theorem 10.7 that \(P(A) = P(A)\) for any \(A \in \mathcal{F}_n\). Since this holds for any choice of a finite number of events \(A_1, \ldots, A_m\), we get that the collection \((\mathcal{F}_1, \ldots, \mathcal{F}_{n-1}, \mathcal{F}_n, \Pi_{n+1}, \ldots)\) are mutually independent.

**Corollary 10.9.** Let \((\Omega, \mathcal{F}, P)\) be a probability space, and let \((A_n)_n\) be a sequence of mutually independent events. Then, the \(\sigma\)-algebras \((\sigma(A_n))_n\) are mutually independent.

**10.3. Second Borel-Cantelli Lemma**

We conclude this lecture with

**Lemma 10.10** (Second Borel-Cantelli Lemma). Let \((A_n)\) be a sequence of mutually independent events. If \(\sum_n P(A_n) = \infty\) then \(P(\limsup_n A_n) = P(A_n \text{ i.o.}) = 1\).

**Proof.** It suffices to show that \(P(\liminf_n A_n^c) = 0\).

For fixed \(m > n\) note that by independence

\[
P(\bigcap_{k=n}^m A_k^c) = \prod_{k=n}^m (1 - P(A_k)).
\]

Since \(\lim_n \bigcap_{k=n}^m A_k^c = \bigcap_{k \geq n} A_k^c\), we have that

\[
P(\bigcap_{k \geq n} A_k^c) = \lim_{m \to \infty} \prod_{k=n}^m (1 - P(A_k)) = \prod_{k \geq n} (1 - P(A_k)) \leq \exp\left(- \sum_{k \geq n} P(A_k)\right),
\]

where we have used \(1 - x \leq e^{-x}\).

Since \(\sum_n P(A_n) = \infty\), we have that \(\sum_{k \geq n} P(A_k) = \infty\) for all fixed \(n\). Thus, \(P(\bigcap_{k \geq n} A_k^c) = 0\) for all fixed \(n\), and

\[
P(\liminf_n A_n^c) = P(\bigcup_n \bigcap_{k \geq n} A_k^c) = P(\lim_n \bigcap_{k \geq n} A_k^c) = \lim_{n \to \infty} P(\bigcap_{k \geq n} A_k^c) = 0.
\]

**Example 10.11.** A biased coin is tossed infinitely many times, all tosses mutually independent. What is the probability that the sequence 01010 appears infinitely many times?
Let $p$ be the probability the coin’s outcome is 1. Let $a_n$ be the outcome of the $n$-th toss. Let $A_n$ be the event that $a_n = 0, a_{n+1} = 1, a_{n+2} = 0, a_{n+3} = 1, a_{n+4} = 0$. Since the tosses are mutually independent, we have that the sequence of events $(A_{5n+1})_{n \geq 0}$ are mutually independent. Since $\mathbb{P}(A_{5n+1}) = p^{-2}(1-p)^{-3} > 0$, we have that $\sum_{n=0}^{\infty} \mathbb{P}(A_{5n+1}) = \infty$. So the Borel-Cantelli Lemma tells us that $\mathbb{P}(A_{5n+1} \text{ i.o.}) = 1$. Since $\{A_{5n+1} \text{ i.o.}\}$ implies $\{A_n \text{ i.o.}\}$ we are done.
11.1. Independent Random Variables

Let $X : (\Omega, \mathcal{F}) \to (\mathbb{R}, \mathcal{B})$ be a random variable. Recall that in the definition of a random variable we require that $X$ is a measurable function; i.e. for any Borel set $B \in \mathcal{B}$ we have $X^{-1}(B) \in \mathcal{F}$. We want to define a $\sigma$-algebra that is all the possible information that can be inferred from $X$.

**Definition 11.1.** Let $(\Omega, \mathcal{F}, P)$ be a probability space and let $(X_\alpha)_{\alpha \in I} : \Omega \to \mathbb{R}$ be a collection of random variables. The $\sigma$-algebra generated by $(X_\alpha)_{\alpha \in I}$ is defined as

$$\sigma(X_\alpha : \alpha \in I) := \sigma(\{X_\alpha^{-1}(B) : \alpha \in I, B \in \mathcal{B}\}).$$

Note that for a random variable $X : \Omega \to \mathbb{R}$, the collection $\{X^{-1}(B) : B \in \mathcal{B}\}$ is a $\sigma$-algebra, so $\sigma(X) = \{X^{-1}(B) : B \in \mathcal{B}\}$.

We now define independent random variables.

**Definition 11.2.** Let $(X_n)_n, (Y_n)_n$ be collections of random variables on some probability space $(\Omega, \mathcal{F}, P)$. We say that $(X_n)_n$ are mutually independent if the $\sigma$-algebras $(\sigma(X_n))_n$ are mutually independent.

We say that $(X_n)_n$ are independent of $(Y_n)_n$ if the $\sigma$-algebras $\sigma((X_n)_n)$ and $\sigma((Y_n)_n)$ are independent.

An important property (that is a consequence of the $\pi$-system argument above):

**Proposition 11.3.** Let $(X_n)_n$ be a collection of random variables on $(\Omega, \mathcal{F}, P)$. $(X_n)$ are mutually independent, if for any finite number of random variables from the collection,
\(X_1, \ldots, X_n\), and any real numbers \(a_1, a_2, \ldots, a_n\), we have that

\[
P[X_1 \leq a_1, X_2 \leq a_2, \ldots, X_n \leq a_n] = P[X_1 \leq a_1] \cdot P[X_2 \leq a_2] \cdots P[X_n \leq a_n].
\]

**Proof.** Let \(X_1, \ldots, X_n\) be a finite number of random variables.

Define two probability measure on the Borel sets of \(\mathbb{R}^n\): For any \(B \in \mathcal{B}^n\) let

\[
P_1(B) = P((X_1, \ldots, X_n) \in B) \quad \text{and} \quad P_2(B) = P(X_1 \in \pi_1 B) \cdots P(X_n \in \pi_n B),
\]

where \(\pi_j\) is the projection onto the \(j\)-th coordinate. Since these two measure are identical on the \(\pi\)-system

\[\{(-\infty, a_1] \times \cdots \times (-\infty, a_n] : a_1, a_2, \ldots, a_n \in \mathbb{R}\},\]

and since this \(\pi\)-system generates all Borel sets on \(\mathbb{R}^n\), we get that \(P_1 = P_2\) for all Borel sets on \(\mathbb{R}^n\).

Thus, if \(A_j \in \sigma(X_j), j = 1, \ldots, n\), then for all \(j\), \(A_j = X_j^{-1}(B_j)\) for some Borel set on \(\mathbb{R}\), so

\[
P(A_1 \cap \cdots \cap A_n) = P((X_1, \ldots, X_n) \in B_1 \times \cdots \times B_n) = P(X_1 \in B_1) \cdots P(X_n \in B_n) = P(A_1) \cdots P(A_n).
\]

\(\square\)

**Example 11.4.** We toss two dice. \(Y\) is the number on the first die and \(X\) is the sum of both dice.

Here the probability space is the uniform measure on all pairs of dice results, so \(\Omega = \{1, 2, \ldots, 6\}^2\).

What is \(\sigma(X)\)?

\[
\sigma(X) = \{\{(x, y) \in \Omega : x + y \in A, A \subset \{2, \ldots, 12\}\}\}.
\]

(We have to know that all subsets of \(\{2, \ldots, 12\}\) are Borel sets. This follows from the fact that \(\{r\} = \bigcup_n (r - 1/n, r]\).) On the other hand,

\[
\sigma(Y) = \{\{(x, y) \in \Omega : x \in A, A \subset \{1, \ldots, 6\}\}\}.
\]

Now note

\[
P[X = 7, Y = 3] = \frac{1}{36} = P[X = 7] \cdot P[Y = 3].
\]
This is mainly due to \( \mathbb{P}[X = 7] = 1/6 \). Similarly, for any \( y \in \{1, \ldots, 6\} \), we have that the events \( \{X = 7\} \) and \( \{Y = y\} \) are independent.

Are \( X \) and \( Y \) independent random variables? NO!

For example,

\[
\mathbb{P}[X = 6, Y = 3] = \frac{1}{36} \neq \frac{5}{36} \cdot \frac{1}{6} = \mathbb{P}[X = 6] \cdot \mathbb{P}[Y = 3].
\]

For independence we need all information from \( X \) to be independent of the information from \( Y \)!

\( \triangle \ \nabla \ \triangle \)

**Example 11.5.** Let \((X_n)_n\) be independent random variables such that \(X_n \sim \text{Exp}(1)\) for all \(n\). Then, for any \(\alpha > 0\),

\[
\mathbb{P}[X_n > \alpha \log n] = e^{-\alpha \log n} = n^{-\alpha}.
\]

Note that in this case,

\[
\sum_{n \geq 1} \mathbb{P}[X_n > \alpha \log n] = \begin{cases} 
\infty & \text{if } \alpha \leq 1 \\
< \infty & \text{if } \alpha > 1.
\end{cases}
\]

For every \(n\) the event \( \{X_n > \alpha \log n\} \in \sigma(X_n) \). So the sequence \((\{X_n > \alpha \log n\})_n\) is independent. By the Borel-Cantelli Lemma (both parts)

\[
\mathbb{P}[X_n > \alpha \log n \ i.o.] = \begin{cases} 
1 & \text{if } \alpha \leq 1 \\
0 & \text{if } \alpha > 1.
\end{cases}
\]

\( \triangle \ \nabla \ \triangle \)
12.1. Joint Distributions

In the same way that we constructed the Borel sets on $\mathbb{R}$, we can consider the $\sigma$-algebra of subsets of $\mathbb{R}^d$, generated by sets of the form $(a_1, b_1] \times (a_2, b_2] \times \cdots \times (a_d, b_d]$. This is the Borel $\sigma$-algebra on $\mathbb{R}^d$, usually denoted by $\mathcal{B}^d$ or $\mathcal{B}(\mathbb{R}^d)$.

Suppose now that we have $d$ random variables $X_1, \ldots, X_d$ on some probability space $(\Omega, \mathcal{F}, \mathbb{P})$. Then, we can define a function $X : \Omega \to \mathbb{R}^d$ by $X(\omega) = (X_1(\omega), X_2(\omega), \ldots, X_d(\omega))$. Suppose we are told that this function is measurable from $(\Omega, \mathcal{F})$ to $(\mathbb{R}^d, \mathcal{B}^d)$. In this case $X$ is called a $\mathbb{R}^d$-valued random variable, and we say that $X_1, \ldots, X_d$ have a joint distribution on $(\Omega, \mathcal{F}, \mathbb{P})$.

**Proposition 12.1.** $X_1, \ldots, X_d$ are random variables if and only if $X = (X_1, \ldots, X_d)$ is a $\mathbb{R}^d$-valued random variable.

**Proof.** Note that given a measurable function $X : (\Omega, \mathcal{F}) \to (\mathbb{R}^d, \mathcal{B}^d)$ we can always write $X = (X_1, \ldots, X_d)$, and each $X_j$ is a random variable, since for any $B \in \mathcal{B}$,

$$X_j^{-1}(B) = X^{-1}(\mathbb{R} \times \cdots \times B \times \cdots \times \mathbb{R}) \in \mathcal{F}.$$

Now suppose $X_1, \ldots, X_d$ are random variables on $(\Omega, \mathcal{F}, \mathbb{P})$. We need to show that $X^{-1}(B) \in \mathcal{F}$ for all $B \in \mathcal{B}^d$.

Since $\mathcal{B}^d = \sigma(\{(a_1, b_1] \times (a_2, b_2] \times \cdots \times (a_d, b_d] : a_j < b_j, j = 1, \ldots, d\})$, by Proposition 7.2 it suffices to prove that

$$X^{-1}((a_1, b_1] \times (a_2, b_2] \times \cdots \times (a_d, b_d]) \in \mathcal{F}$$
for all $a_j < b_j$, $j = 1, 2, \ldots, d$. This follows from the fact that

$$X^{-1}((a_1, b_1] \times (a_2, b_2] \times \cdots \times (a_d, b_d]) = \bigcap_{j=1}^d X_j^{-1}((a_j, b_j]) \in \mathcal{F}.$$ 

The probability measure $\mathbb{P}_X$ on $(\mathbb{R}^d, \mathcal{B}^d)$ defined by $\mathbb{P}_X(B) = \mathbb{P}(X \in B) = \mathbb{P}(X^{-1}(B))$ is called the distribution of $X$. We can also define the joint distribution function of $X_1, \ldots, X_d$:

$$F_{(X_1, \ldots, X_d)}(t_1, \ldots, t_d) = F_X(\mathbf{t}) = \mathbb{P}(X_1 \leq t_1, X_2 \leq t_2, \ldots, X_d \leq t_d).$$

A restatement of Proposition 11.3 is:

**Proposition 12.2.** Let $(X_1, \ldots, X_d)$ be $d$ random variables on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ with a joint distribution. Then, $(X_1, \ldots, X_d)$ are mutually independent if and only if

$$\forall \ (t_1, \ldots, t_d) \in \mathbb{R}^d \quad F_X(t_1, \ldots, t_d) = F_{X_1}(t_1) \cdots F_{X_d}(t_d).$$

**Example 12.3.** An urn contains 3 red balls, 4 white balls and 5 blue balls. We remove three balls from the urn, all balls equally likely. Let $X$ be the number of red balls removed and $Y$ the number of white balls removed.

What is the joint distribution of $(X, Y)$?

A natural probability space here is the uniform measure on $\Omega = \{S : |S| = 3, S \subset \{1, \ldots, 12\}\}.$

Suppose $f : \{1, \ldots, 12\} \to \{\text{red, white, blue}\}$ is the function assigning a color to each ball in the urn.

So

$$(X, Y)(S) = (\# \{s \in S : f(s) = \text{red}\}, \# \{s \in S : f(s) = \text{white}\}).$$

So

$$\mathbb{P}_{(X,Y)}(\{(0,0)\}) = \mathbb{P}(\{S \in \Omega : f(s) = \text{blue} \forall s \in S\}) = \binom{5}{3} \binom{12}{3}.$$

For any $S \in \Omega$, we can write $S = S_r \uplus S_w \uplus S_b$ where $S_r = \{s \in S : f(s) = \text{red}\}$ and similarly for white and blue. So $|S| = |S_r| + |S_w| + |S_b|.$
If \((X,Y)(S) = (0,1)\) then \(S_r = 0, S_w = 1\) and \(S_b = 2\). So
\[
\mathbb{P}_{(X,Y)}(\{(0,1)\}) = \frac{\binom{3}{0} \cdot \binom{1}{1} \cdot \binom{3}{2}}{\binom{12}{3}}.
\]
Similarly, if \((X,Y)(S) = (1,1)\) then \(S_r = 1, S_w = 1\) and \(S_b = 1\). So
\[
\mathbb{P}_{(X,Y)}(\{(1,1)\}) = \frac{\binom{3}{1} \cdot \binom{1}{1} \cdot \binom{5}{1}}{\binom{12}{3}}.
\]
Generally, for integers \(x, y \geq 0\), such that \(x + y \leq 3\),
\[
\mathbb{P}_{(X,Y)}(\{(x,y)\}) = \frac{\binom{3}{x} \cdot \binom{4}{y} \cdot \binom{5}{3-x-y}}{\binom{12}{3}}.
\]
Of course \(F_{(X,Y)}\) can be calculated by summing these. \(\triangle \nabla \triangle\)

**Example 12.4.** Two fair coins are tossed, there outcomes independent. \(X \in \{0,1\}\) is the outcome of the first coin. \(Y \in \{0,1\}\) is the outcome of the second coin. \(Z = X + Y \pmod{2}\). Show that \(X, Z\) are independent.

First \(\mathbb{P}(X = 0) = \mathbb{P}(X = 1) = \frac{1}{2}\). Also,
\[
\mathbb{P}(Z = 0) = \mathbb{P}(X = Y = 0) + \mathbb{P}(X = Y = 1) = \frac{1}{4} + \frac{1}{4} = \frac{1}{2}.
\]
Similarly, \(\mathbb{P}(Z = 1) = \frac{1}{2}\).

Now,
\[
\mathbb{P}(X = 0, Z = 0) = \mathbb{P}(X = 0, Y = 0) = \frac{1}{4} = \mathbb{P}(X = 0) \cdot \mathbb{P}(Z = 0).
\]
Similarly for other values. \(\triangle \nabla \triangle\)

As in the case of 1-dimensional random variables, we would like to define the information a random variable carries as some \(\sigma\)-algebra. Thus, for a \(d\)-dimensional random variable \(X\) we define
\[
\sigma(X) := \left\{X^{-1}(B) \mid B \in \mathcal{B}\right\}.
\]
This is a \(\sigma\)-algebra.

Recall that if \(X = (X_1, \ldots, X_d)\) then
\[
\sigma(X_1, \ldots, X_d) = \sigma\left(\left\{X_j^{-1}(B) \mid B \in \mathcal{B}, 1 \leq j \leq d\right\}\right)
= \sigma\left(\left\{X^{-1}(\mathbb{R} \times \cdots \times \mathbb{R} \times B \times \mathbb{R} \times \cdots \times \mathbb{R}) \mid B \in \mathcal{B}\right\}\right).
\]
The following proposition shows that these notations agree.

**Proposition 12.5.** Let \( X = (X_1, \ldots, X_d) \) be a \( d \)-dimensional random variable on some probability space. Then, \( \sigma(X) = \sigma(X_1, \ldots, X_d) \).

**Proof.** Define \( \mathcal{G} = \{ B \in \mathcal{B}^d : X^{-1}(B) \in \sigma(X_1, \ldots, X_d) \} \) we get that this is a \( \sigma \)-algebra, and that the set of rectangles \( \mathcal{K} = \{ (a_1, b_1] \times \cdots \times (a_d, b_d] : a_j < b_j \in \mathbb{R} \} \) is contained in \( \mathcal{G} \). Since \( \mathcal{B}^d \) is generated by \( \mathcal{K} \), we get that \( \mathcal{B}^d \subset \mathcal{G} \), and \( \sigma(X) \subset \sigma(X_1, \ldots, X_n) \). On the other hand, if \( B \in \mathcal{B} \) then \( X^{-1}(B) = X^{-1}(\mathbb{R} \times \cdots \times B \times \cdots \times \mathbb{R}) \in \sigma(X) \). So \( \sigma(X_1, \ldots, X_n) \subset \sigma(X) \) and they are equal. \( \square \)

### 12.2. Marginals

**Lemma 12.6.** Let \( X = (X_1, \ldots, X_d) \) be a \( d \)-dimensional random variable on some probability space \( (\Omega, \mathcal{F}, \mathbb{P}) \). Let \( Y = (X_1, \ldots, X_{d-1}) \) and let \( Z = (X_2, \ldots, X_d) \) (which are \( d-1 \)-dimensional). Then, for any \( \vec{t} \in \mathbb{R}^{d-1} \),

\[
F_Y(\vec{t}) = \lim_{t \to \infty} F_X(\vec{t}, t) \quad \text{and} \quad F_Z(\vec{t}) = \lim_{t \to \infty} F_X(t, \vec{t}).
\]

**Proof.** Fix \( \vec{t} = (t_1, \ldots, t_{d-1}) \in \mathbb{R}^{d-1} \). Let \( A = \{ X_1 \leq t_1, \ldots, X_{d-1} \leq t_{d-1} \}. \)

Let \( s_n \nearrow \infty \), and set \( A_n = \{ X_d \leq s_n \} \). Since \( (s_n) \) is an increasing sequence, we have that \( A_n \subset A_{n+1} \), so \( (A_n)_n \) is an increasing sequence, and thus \( (A \cap A_n)_n \) is also an increasing sequence. Note that

\[
\lim_n A_n = \bigcup_n A_n = \{ X < \infty \} = \Omega,
\]

so \( \lim_n (A \cap A_n) = A \). Thus, using continuity of probability

\[
F_Y(\vec{t}) = \mathbb{P}[A] = \lim_n \mathbb{P}[A \cap A_n] = \lim_{n \to \infty} F_X(\vec{t}, s_n).
\]

Since this holds for any sequence \( s_n \nearrow \infty \), we have the first assertion.

The proof of the second assertion is almost identical, switching the roles of the first and last coordinate. \( \square \)
Corollary 12.7. Let $X = (X_1, \ldots, X_d)$ be a $d$-dimensional random variable on some probability space $(\Omega, \mathcal{F}, \mathbb{P})$. Then, for any $t \in \mathbb{R}$,

$$F_{X_j}(t) = \lim_{t_i \to \infty : i \neq j} F_X(t_1, \ldots, t_{j-1}, t, t_{j+1}, \ldots, t_d).$$

$F_{X_j}$ above is called the **marginal distribution function** of $X_j$. 
13.1. DISCRETE JOINT DISTRIBUTIONS

Recall that a random variable is discrete if there exists a countable set \( R \) such that \( \mathbb{P}(X \in R) = 1 \). \( R \) is the range of \( X \), and the function \( f_X(r) = \mathbb{P}(X = r) \) is the density of \( X \).

Suppose that \( X, Y \) are discrete random variables, with ranges \( R_X, R_Y \) and densities \( f_X, f_Y \). Note that \( R := R_X \cup R_Y \) is also a range for both \( X \) and \( Y \).

Now, we can define the joint density of \( X \) and \( Y \) by

\[
    f_{X,Y}(x,y) = \mathbb{P}(X = x, Y = y) = \mathbb{P}((X, Y) = (x, y)).
\]

Note that

\[
    \mathbb{P}((X, Y) \notin R^2) = \mathbb{P}((X \notin R) \cup (Y \notin R)) \leq \mathbb{P}(X \notin R) + \mathbb{P}(Y \notin R) = 0.
\]

So \( \mathbb{P}((X, Y) \in R^2) = 1 \), and we can think of \((X, Y)\) as a \( \mathbb{R}^2 \)-valued discrete random variable.

This of course can be generalized:

If \( X_1, \ldots, X_n \) are discrete random variables. Then \((X_1, \ldots, X_n)\) is a \( \mathbb{R}^n \) valued random variable, supported on some countable set in \( \mathbb{R}^n \), and the density of \((X_1, \ldots, X_n)\) is defined to be

\[
    f_{(X_1, \ldots, X_n)}(x_1, \ldots, x_n) = \mathbb{P}(X_1 = x_1, \ldots, X_n = x_n).
\]

Exercise 13.1. Let \( X_1, \ldots, X_d \) be random variables. Show that \( X = (X_1, \ldots, X_d) \) is a jointly discrete random variable if and only if \( X_1, \ldots, X_d \) are each discrete.
Assume that $X_1, \ldots, X_d$ are discrete random variables. Show that $X_1, \ldots, X_d$ are mutually independent, if and only if

$$\forall (t_1, \ldots, t_d) \in \mathbb{R}^d \quad f_{(X_1, \ldots, X_d)}(t_1, \ldots, t_d) = f_{X_1}(t_1) \cdot f_{X_2}(t_2) \cdots f_{X_d}(t_d).$$

### 13.2. Discrete Marginal Densities

**Proposition 13.2.** If $X = (X_1, \ldots, X_d)$ is a discrete joint random variable with range $\mathbb{R}^d$ and density $f_X$, then the density of $X_j$ is given by

$$f_{X_j}(x) = \sum_{x_i \in \mathbb{R} : i \neq j} f_X(x_1, \ldots, x_{j-1}, x, x_{j+1}, \ldots, x_d).$$

**Proof.** This follows from

$$F_X(t_1, \ldots, t_d) = \sum_{R \ni r_j \leq t_j : j=1,\ldots,d} f_X(r_1, \ldots, r_d).$$

So

$$\lim_{t_i \to \infty : i \neq j} F_X(t_1, \ldots, t_{j-1}, t, t_{j+1}, \ldots, t_d) = \sum_{R \ni r_j \leq t} \sum_{r_i \in \mathbb{R} : i \neq j} f_X(r_1, \ldots, r_d).$$

Thus,

$$f_{X_j}(r) = F_{X_j}(r) - F_{X_j}(r^-) = \sum_{R \ni r_j \leq r} \sum_{r_i \in \mathbb{R} : i \neq j} f_X(r_1, \ldots, r_d) - \sum_{R \ni r_j < r} \sum_{r_i \in \mathbb{R} : i \neq j} f_X(r_1, \ldots, r_d)$$

$$= \sum_{r_i \in \mathbb{R} : i \neq j} f_X(r_1, \ldots, r_{j-1}, r, r_{j+1}, \ldots, r_d).$$

\[\square\]

The above density is called the **marginal density** of $X_j$.

**Example 13.3.** Noga and Shir come home from school and decide if to do homework or watch TV.

Let $X = 1$ if Noga does homework and $X = 0$ if she watches TV. Let $Y = 1$ if Shir does homework and $Y = 0$ if she watches TV.

We are given that

$$\mathbb{P}((X, Y) = (1, 1)) = \frac{1}{4}, \quad \mathbb{P}((X, Y) = (0, 0)) = \frac{1}{2}, \quad \mathbb{P}((X, Y) = (1, 0)) = \frac{3}{16}.$$ 

What are the marginal densities of $X$ and $Y$?
Who is more likely to watch TV?

Note that since the density must sum to 1, \( P((X, Y) = (0, 1)) = \frac{1}{16} \).

Well, let's calculate the marginal densities:

\[
f_X(1) = P((X, Y) = (1, 0)) + P((X, Y) = (1, 1)) = \frac{7}{16}.
\]

So it must be that \( f_X(0) = \frac{9}{16} \) (why?). Similarly,

\[
f_Y(1) = P((X, Y) = (0, 1)) + P((X, Y) = (1, 1)) = \frac{5}{16}.
\]

So \( f_Y(0) = \frac{11}{16} \).

We see that Shir is more likely to watch TV.

\[\triangle \nabla \triangle\]

13.3. Conditional Distributions (Discrete)

Example 13.4. After mating season sea turtles land on the beach and lay their eggs. After hatching, the baby turtles must make it back to the sea on their own, however, there are many dangers and predators awaiting.

Each female hatches a Poisson-\( \lambda \) number of eggs. After hatching, each baby turtle has probability \( p \) to make it back to the sea, all turtles mutually independent.

What is the probability that exactly \( k \) turtles make it back to sea? \[\triangle \nabla \triangle\]

How do we solve this? We would like to use conditional probabilities, to say that conditioned on their being \( n \) eggs, the number of surviving turtles is Binomial-(\( n, p \)).

So we want to define distributions conditioned on the results of other distributions.

Definition 13.5. Let \( X, Y \) be discrete random variables defined on some probability space \((\Omega, \mathcal{F}, \mathbb{P})\). For \( y \) such that \( f_Y(y) > 0 \) define the random variable \( X|Y = y \) as the random variable whose density is

\[
f_{X|Y}(x|y) = f_{X|Y=y}(x) = P(X = x|Y = y).
\]

✓ Note that the correct way to do this would be to define a different density \( f_{X|y}(\cdot) \) for every \( y \). This is captured in the notation \( f_{X|Y}(\cdot|y) \).
Example 13.6. Let the density of $X, Y$ be given by:

$$
\begin{array}{c|cc}
X \setminus Y & 0 & 1 \\
\hline
0 & 0.4 & 0.1 \\
1 & 0.2 & 0.3 \\
\end{array}
$$

So

$$
P(Y = 1) = 0.4 \quad \text{and} \quad P(Y = 0) = 0.6.
$$

Thus,

$$
f_{X|Y}(1|1) = P(X = 1|Y = 1) = 0.3/0.4 = 3/4,$$

$$
f_{X|Y}(0|1) = P(X = 0|Y = 1) = 0.1/0.4 = 1/4,$$

$$
f_{X|Y}(1|0) = 0.2/0.6 = 1/3,$$

$$
f_{X|Y}(0|0) = 0.4/0.6 = 2/3.
$$

△▽△

Let’s return to solve the example above:

Solution of Example 13.4. Let $X$ be the number of eggs, and let $Y$ be the number of turtles that survive.

What we are given, is that $X \sim \text{Poi}(\lambda)$ and that $Y|X = n \sim \text{Bin}(n, p)$.

That is, for all $k \leq n$,

$$
P(Y = k|X = n) = \binom{n}{k}p^k(1-p)^{n-k}.
$$

Thus,

$$
P(Y = k) = \sum_{n=0}^{\infty} P(Y = k|X = n) P(X = n) = \sum_{n=k}^{\infty} \binom{n}{k}p^k(1-p)^{n-k} \cdot e^{-\lambda} \frac{\lambda^n}{n!}
$$

$$
= \frac{p^k\lambda^k}{k!} e^{-\lambda} \sum_{n=k}^{\infty} \frac{\lambda^{n-k}(1-p)^{n-k}}{(n-k)!}
$$

$$
= \frac{(p\lambda)^k}{k!} e^{-\lambda} e^{\lambda(1-p)} = e^{-\lambda p} \cdot \frac{(\lambda p)^k}{k!}.
$$

That is, the number of surviving turtles has the distribution of a Poisson-$\lambda p$ random variable.

□
**Exercise 13.7.** Let $X,Y$ be discrete random variables. Show that

$$f_{X,Y}(x,y) = f_{X|Y}(x|y)f_Y(y),$$

for all $y$ such that $f_Y(y) > 0$.

### 13.4. Sums of Independent Random Variables (Discrete Case)

Suppose that $X,Y$ are random variables on some probability space. One can define the random variable $Z(\omega) = X(\omega) + Y(\omega)$, or in short $Z = X + Y$. (This is always a random variable, because $Z = \phi(X,Y)$, where $\phi(x,y) = x + y$, and since $\phi$ is continuous it is measurable).

Suppose $X,Y$ are independent discrete random variables, with densities $f_X, f_Y$.

First, from independence, we have that

$$P(X = x, Y = y) = P(X = x)P(Y = y),$$

for all $x,y$. Thus, $f_{(X,Y)}(x,y) = f_X(x)f_Y(y)$.

Now, what is the density of $X + Y$? Well,

$$f_{X+Y}(z) = P(X + Y = z) = P\left(\bigcup_y \{X + Y = z, Y = y\}\right) = \sum_y P(X + Y = z, Y = y)$$

$$= \sum_y P(X = z - y)P(Y = y) = \sum_y f_X(z - y)f_Y(y).$$

This leads to the following definition:

**Definition 13.8.** Let $f,g$ be two real valued functions supported on a countable set $R \subset \mathbb{R}$. The (discrete) convolution of $f,g$, denoted $f * g$, is the real valued function $f * g : R \to \mathbb{R}$,

$$(f * g)(z) = \sum_{y \in R \cup (z-R)} f(z - y)g(y).$$

Note that $f * g = g * f$ by change of variables.

We conclude with the observation:

**Proposition 13.9.** Let $X,Y$ be two independent discrete random variables on some probability space. Let $Z = X + Y$. Then, $f_Z = f_X * f_Y$. 

Example 13.10. Consider $X_1, \ldots, X_n$ mutually independent random variables where $X_j \sim \text{Poi}(\lambda_j)$ for some $\lambda_1, \ldots, \lambda_n > 0$.

Let $Z = X_1 + \cdots + X_n$. What is the distribution of $Z$?

Well, since $f_{X_1}(x) = 0$ for $x < 0$,

$$f_{X_1+X_2}(z) = f_{X_1} \ast f_{X_2}(z) = \sum_{x=0}^{\infty} f_{X_1}(z-x) f_{X_2}(x) = \sum_{x=0}^{z} e^{-\lambda_1} \frac{\lambda_1^{z-x}}{(z-x)!} \cdot e^{-\lambda_2} \frac{\lambda_2^x}{x!} = e^{-(\lambda_1+\lambda_2)} \cdot \sum_{x=0}^{z} \binom{z}{x} \frac{\lambda_1^{z-x} \lambda_2^x}{x!} = e^{-(\lambda_1+\lambda_2)} \cdot \frac{(\lambda_1 + \lambda_2)^z}{z!}.$$  

This is the density of the Poisson distribution!

So $X_1 + X_2 \sim \text{Poi}(\lambda_1 + \lambda_2)$. Continuing inductively, we conclude that $Z \sim \text{Poi}(\lambda_1 + \lambda_2 + \cdots + \lambda_n)$.

That is: the sum of independent Poisson random variables is again a Poisson random variable.

Example 13.11. Let $X \sim \text{Poi}(\alpha)$ and $Y \sim \text{Poi}(\beta)$. Assume $X,Y$ are independent.

What is the distribution of $X|X+Y = n$?

Let’s calculate explicitly: Recall that $X + Y \sim \text{Poi}(\alpha + \beta)$. For $k \leq n$,

$$\Pr(X = k | X + Y = n) = \frac{\Pr(X = k) \Pr(Y = n-k)}{\Pr(X + Y = n)} = e^{-\alpha} \frac{\alpha^k}{k!} \cdot e^{-\beta} \frac{\beta^{n-k}}{(n-k)!} \cdot \left( e^{-(\alpha+\beta)} \frac{(\alpha + \beta)^n}{n!} \right)^{-1} = \binom{n}{k} \frac{\alpha^k \beta^{n-k}}{(\alpha + \beta)^n} = \binom{n}{k} \left( \frac{\alpha}{\alpha + \beta} \right)^k \cdot \left( 1 - \frac{\alpha}{\alpha + \beta} \right)^{n-k}.$$  

So $X$ conditioned on $X + Y = n$ has the Binomial-$(n, \alpha/(\alpha + \beta))$ distribution.
14.1. CONTINUOUS JOINT DISTRIBUTIONS

**Definition 14.1.** Let $X_1, \ldots, X_d$ be random variables on some probability space $(\Omega, \mathcal{F}, \mathbb{P})$. We say that $X_1, \ldots, X_d$ have an **absolutely continuous joint distribution** (or that $X = (X_1, \ldots, X_d)$ is a $\mathbb{R}^d$-valued absolutely continuous random variable) if there exists a non-negative integrable function $f_X = f_{X_1,\ldots,X_d} : \mathbb{R}^d \to \mathbb{R}^+$, such that for all $\vec{t} = (t_1, \ldots, t_d)$,

$$F_X(\vec{t}) = \int_{-\infty}^{t_1} \cdots \int_{-\infty}^{t_d} f_X(\vec{s})d\vec{s}.$$

Marginal distributions are as in the discrete case: If $X = (X_1, \ldots, X_d)$ is a joint random variable then the marginal distribution of $X_j$ is

$$F_{X_j}(t) = \lim_{t_k \to \infty : k \neq j} F_X(t_1, \ldots, t_{j-1}, t, t_{j+1}, \ldots, t_d).$$

Specifically, if $X$ is absolutely continuous, then

$$f_{X_j}(t) = \int \cdots \int f_X(t_1, \ldots, t_{j-1}, t, t_{j+1}, \ldots, t_d)dt_1 \cdots dt_{j-1} dt_{j+1} \cdots dt_d.$$

It is not true that if $X_1, \ldots, X_d$ are absolutely continuous, then necessarily $X = (X_1, \ldots, X_d)$ is absolutely continuous as a $\mathbb{R}^d$-valued random variable.

**Example 14.2.** If $X = Y$ are the same absolutely continuous random variable, then $(X, Y)$ is **not** absolutely continuous, since if it was, then for the Borel set $A = \{(x, y) \in \mathbb{R}^2 : x = y\}$,

$$1 = \int \int f_{X,Y}(x, y)dxdy = \int \int A f_{X,Y}(x, y)dxdy = 0$$

since $f_{X,Y}(x, y) = 0$ if $(x, y) \not\in A$ and since $A$ has measure 0 in the plane.
A more elementary way to see this (in the scope of these notes) is: If \( f_{X,Y} \) is the density of a pair \((X,Y)\), such that \( X = Y \) is an absolutely continuous random variable, then: Let \( g_s(x) = \int_{-\infty}^{s} f_{X,Y}(x,y)dy \). Then,

\[
F_{X,Y}(\infty, s) = \mathbb{P}[X \leq \infty, Y \leq s] = \mathbb{P}[X \leq s, Y \leq s] = F_{X,Y}(s, s).
\]

So

\[
\int_{-\infty}^{\infty} g_s(x)dx = F_{X,Y}(\infty, s) = F_{X,Y}(s, s) = \int_{-\infty}^{s} g_s(x)dx,
\]

which implies that

\[
\int_{s}^{\infty} g_s(x)dx = 0.
\]

Thus, \( g_s(x) = 0 \) for all (actually a.e.) \( x \geq s \) (since \( g_s \) is a non-negative function). Now, if \( x \geq s \) then

\[
0 = g_s(x) = \int_{-\infty}^{s} f_{X,Y}(x,y)dy,
\]

so \( f_{X,Y}(x,y) = 0 \) for all (actually a.e.) \( x \geq s, y \leq s \). Since this holds for all \( s \) we have that \( f_{X,Y} = 0 \) for all (or a.e.) \( x \geq y \). Exchanging the roles of \( X, Y \) we get that \( f_{X,Y} = 0 \) (a.e.), which is not a density function, contradiction!

\[\triangle \nabla \triangle\]

However, we have the following criterion for independence:

**Proposition 14.3.** \( X_1, \ldots, X_n \) are mutually independent absolutely continuous random variables if and only if \( X = (X_1, \ldots, X_n) \) is a \( \mathbb{R}^n \)-valued absolutely continuous random variable with density \( f_X(t_1, \ldots, t_n) = f_{X_1}(t_1) \cdots f_{X_n}(t_n) \).

**Proof.** The “if” part follows from Proposition 11.3.

For the “only if” part: Let \( f(t_1, \ldots, t_n) = f_{X_1}(t_1) \cdots f_{X_n}(t_n) \). We need to show that

\[
\mathbb{P}[X_1 \leq t_1, \ldots, X_n \leq t_n] = \int_{\mathbb{R}^n} f(t_1, \ldots, t_n)dt_1 \cdots dt_n.
\]

Indeed, this follows from independence. \[\square\]
14.2. Conditional Distributions (Continuous)

Recall that if \( X, Y \) are discrete random variables, then

\[
f_{X|Y}(x|y) = \mathbb{P}[X = x| Y = y].
\]

For continuous random variables this is not defined, since \( \mathbb{P}[Y = y] = 0 \). However, we still can define the following:

Let \( X, Y \) be absolutely continuous random variables. Let \( y \in \mathbb{R} \) be such that \( f_Y(y) > 0 \). Define a function \( f_{X|Y}(\cdot|y) : \mathbb{R} \to \mathbb{R}^+ \) by

\[
f_{X|Y}(x|y) := \frac{f_{X,Y}(x,y)}{f_Y(y)}.
\]

Then, for every \( y \) such that \( f_Y(y) > 0 \), this function is a density; indeed, recall that \( f_Y(y) = \int f_{X,Y}(x,y) \, dx \) so

\[
\int_{-\infty}^{\infty} f_{X|Y}(x|y) \, dx = \int \frac{f_{X,Y}(x,y)}{f_Y(y)} \, dx = 1.
\]

Thus, \( f_{X|Y}(\cdot|y) \) defines an absolutely continuous random variable, denoted \( X|Y = y \), and called \( X \) **conditioned on** \( Y = y \). \( f_{X|Y}(\cdot|y) \) is the **conditional density** of \( X \) conditioned on \( Y = y \). We have the identity

\[
f_{X,Y}(x,y) = f_{X|Y}(x|y) \cdot f_Y(y).
\]

For any Borel set \( B \in \mathcal{B} \), we define

\[
\mathbb{P}[X \in B|Y = y] := \mathbb{P}[(X|Y = y) \in B] = \int_B f_{X|Y}(x|y) \, dx.
\]

In many cases, the information is given as conditional densities, rather than joint densities.

**Example 14.4.** Let \( X, Y \) have joint density

\[
f_{X,Y}(x,y) = \begin{cases} 
\frac{1}{y} e^{-x/y} e^{-y} & x, y \geq 0 \\
0 & \text{otherwise}
\end{cases}
\]

Let’s compute the density of \( X|Y \): First, the marginal density of \( Y \) is

\[
f_Y(y) = \int_0^\infty \frac{1}{y} e^{-x/y} e^{-y} \, dx = e^{-y},
\]
for $y \geq 0$ and 0 otherwise. For $x, y \geq 0$,

$$f_{X|Y}(x|y) = \frac{\frac{1}{y} e^{-x/y} e^{-y}}{e^{-y}} = \frac{1}{y} e^{-x/y},$$

and $f_{X|Y}(x|y) = 0$ for $x < 0$. So, $X|Y = y \sim \text{Exp}(1/y)$.

Thus, for example,

$$P[X > 1|Y = y] = \int_1^\infty f_{X|Y}(x|y)dx = e^{-1/y}.$$

14.3. Sums (Continuous)

Let $X, Y$ be independent absolutely continuous random variables. Let $Z = X + Y$.

What is $F_Z$?

$$P[Z \leq z] = P[X + Y \leq z] = \int_{x+y\leq z} f_{X,Y}(x,y)dxdy$$

$$= \int_{-\infty}^{\infty} \int_{-\infty}^{z-y} f_X(x) f_Y(y) dxdy$$

$$= \int_{-\infty}^{\infty} F_X(z-y) f_Y(y) dy = F_X * f_Y(z).$$

Now, note that

$$F_Z(z) = \int_{-\infty}^{\infty} \int_{-\infty}^{z-y} f_X(x) f_Y(y) dxdy$$

$$= \int_{-\infty}^{z} \int_{-\infty}^{\infty} f_X(s-y) f_Y(y) dyds = \int_{-\infty}^{z} f_X * f_Y(s) ds.$$

Thus, $Z = X + Y$ is an absolutely continuous random variable with density $f_X * f_Y$.

Example 14.5. What is the distribution of $X + Y$, for independent $X, Y$, when:

- $X \sim N(0,1)$ and $Y \sim N(0,1)$.
- $X \sim \text{Exp}(\lambda)$ and $Y \sim \text{Exp}(\alpha)$.

In the normal case we have that $\frac{(x-y)^2 + y^2}{2} = \frac{x^2}{4} + (y - \frac{x}{2})^2$. So,

$$f_{X+Y}(x) = \frac{1}{2\pi} e^{-x^2/4} \cdot \int_{-\infty}^{\infty} e^{-(y-x/2)^2} dy = \frac{1}{2\pi} e^{-x^2/4} \cdot \sqrt{\pi} = \frac{1}{\sqrt{2\pi} \cdot \sqrt{2}} e^{-x^2/4}.$$ 

So $X + Y \sim N(0, \sqrt{2})$. 

\[ \triangle \nabla \triangle \]
In the exponential case for $x > 0$,

$$f_{X+Y}(x) = f_X * f_Y(x) = \alpha \lambda e^{-\lambda x} \int_{-\infty}^{\infty} 1_{[0,\infty)}(x-y) 1_{[0,\infty)}(y) e^{-(\alpha-\lambda)y} dy$$

$$= \frac{\alpha \lambda}{\alpha - \lambda} e^{-\lambda x} (1 - e^{-(\alpha-\lambda)x}) = \frac{\alpha \lambda}{\alpha - \lambda} (e^{-\lambda x} - e^{-\alpha x}).$$

Since $\mathbb{P}[X + Y < 0] = 0$ this is the density of $X + Y$. 

△ ▽ △
In this lecture we want to define the average value of a random variable $X$. If $X$ is discrete, it would be intuitive to define the average as $\sum_x P[X = x]x$. If $X$ is absolutely continuous, perhaps one would also think of defining the average analogously as $\int xf_X(x)dx$. But what about the general case?

We will follow Lebesgue integration theory for this task.

15.1. Preliminaries

Proposition 15.1. Let $X = (X_1, \ldots, X_n)$ and $Y = (Y_1, \ldots, Y_m)$ be random variables on some probability space $(\Omega, \mathcal{F}, \mathbb{P})$. Then, $X$ is independent of $Y$ if and only if for any measurable functions $g : \mathbb{R}^n \to \mathbb{R}$ and $h : \mathbb{R}^m \to \mathbb{R}$ we have that $g(X), h(Y)$ are independent.

Proof. Now, for the "only if" direction: Let $g,h$ be measurable functions as in the proposition. Let $A_1 \in \sigma(g(X)), A_2 \in \sigma(h(Y))$. Note that there exist Borel sets $B_1, B_2$ such that $A_1 = X^{-1}g^{-1}(B_1)$ and $A_2 = Y^{-1}h^{-1}(B_2)$. Since $g$ is measurable, $g^{-1}(B_1) \in \mathcal{B}^n$, and similarly, $h^{-1}(B_2) \in \mathcal{B}^m$. Thus, $A_1 \in \sigma(X)$ and $A_2 \in \sigma(Y)$. Thus, $A_1, A_2$ are independent. Since this holds for all $A_1, A_2$ we have that $g(X), h(Y)$ are independent.

For the "if" direction: Let $A_1 \in \sigma(X), A_2 \in \sigma(Y)$. So $A_1 = X^{-1}(B_1)$ and $A_2 = Y^{-1}(B_2)$ for some $B_1 \in \mathcal{B}^n$ and $B_2 \in \mathcal{B}^m$. Define $g(\vec{x}) = 1_{\{\vec{x} \in B_1\}}$ and $h(\vec{y}) = 1_{\{\vec{y} \in B_2\}}$. Since these are measurable functions, and since $A_1 \cap A_2 = \{X \in B_1, Y \in B_2\}$, we have that

$$\mathbb{P}[A_1 \cap A_2] = \mathbb{P}[g(X) = 1, h(Y) = 1] = \mathbb{P}[g(X) = 1|\mathbb{P}[h(Y) = 1] = \mathbb{P}[A_1]\mathbb{P}[A_2].$$

Since this holds for all $A_1, A_2$ we have that $X$ is independent of $Y$. \qed

For the TA
Example 15.2. The function \( g(x) = \lfloor x \rfloor \) is measurable, since for any \( b \in \mathbb{R} \),

\[ g^{-1}(-\infty, b] = \{ x : \lfloor x \rfloor \leq b \} = (-\infty, [b]) \in \mathcal{B}. \]

The function \( g(x) = c \cdot x \) is measurable, since it is continuous. The function \( g(x) = \max \{ x, 0 \} \) is measurable since \( g^{-1}(-\infty, b] = [0, b] \) if \( b \geq 0 \) and \( g^{-1}(-\infty, b] = \emptyset \) if \( b < 0 \).

In a similar way \( g(x) = \max \{ -x, 0 \} \) is measurable.

Thus, if \( X,Y \) are independent random variables, then \( X^+ := \max \{ X, 0 \}, Y^+ := \max \{ Y, 0 \} \) are independent, and \( X^+_n = 2^{-n} \lfloor 2^n X^+ \rfloor \) is independent of \( Y^+_n = 2^{-n} \lfloor 2^n Y^+ \rfloor \).

15.2. Simple Random Variables

Let us start with simple random variables:

**Definition 15.3.** Let \( (\Omega, \mathcal{F}, \mathbb{P}) \) be a probability space. \( X \) is a **simple** random variable if there exists a finite set of positive numbers \( A = \{a_1, \ldots, a_m \} \subset \mathbb{R} \) such that \( \Omega = \{X = a_1\} \cup \cdots \cup \{X = a_m\} \cup \{X = 0\} \).

Specifically, such a \( X \) is a discrete random variable. In this case it is intuitive that the average of \( X \) should be \( \sum_{k=1}^{m} a_k \mathbb{P}[X = a_k] \). Note that

\[ X = \sum_{k=1}^{m} a_k 1_{\{X = a_k\}}, \]

where all \( a_k > 0 \) and the events \( \{X = a_k\} \) are pairwise disjoint.

We want to say that simple random variables somehow approximate all random variables, so that we can define the average on these. This is the content of the following definitions and proposition.

**Definition 15.4.** Let \((\Omega, \mathcal{F}, \mathbb{P})\) be a probability space. A sequence of random variables \((X_n)_n\) is monotone non-decreasing if for all \( \omega \in \Omega \), and all \( n \), \( X_n(\omega) \leq X_{n+1}(\omega) \).

Let \( (\Omega, \mathcal{F}, \mathbb{P}) \) be a probability space. Let \((X_n)_n\) be a sequence of random variables. Suppose that for every \( \omega \in \Omega \) the limit \( X(\omega) := \lim_{n \to \infty} X_n(\omega) \) exists. Then, the
function $X$ is a random variable, since $X = \limsup X_n = \liminf X_n$, and since for any Borel set $B$,

$$X^{-1}(B) = \left\{ \omega : \liminf_{n \to \infty} X_n(\omega) \in B \right\} = \bigcup_{n} \bigcap_{k \geq n} \left\{ \omega : X_k(\omega) \in B \right\} = \liminf X^{-1}_n(B) \in \mathcal{F}.$$  

(In fact, $\lim X^{-1}_n(B) = X^{-1}(B)$.)

**Definition 15.5.** Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. Let $(X_n)_{n}$ be a sequence of random variables. Suppose that for every $\omega \in \Omega$ the limit $X(\omega) := \lim_{n \to \infty} X_n(\omega)$ exists. Then, the function $X$ is a random variable. In this case we say that $X_n$ converges everywhere to $X$.

If in addition, $(X_n)_{n}$ is a monotone non-decreasing sequence, we say that $X_n$ converges everywhere monotonely (from below) to $X$, and denote this by $X_n \nearrow X$.

Finally, a non-negative random variable $X$ is a random variable satisfying $X(\omega) \geq 0$ for all $\omega \in \Omega$. We write this as $X \geq 0$. In general, we write $X \leq Y$ if $X(\omega) \leq Y(\omega)$ for all $\omega \in \Omega$. (So, for example, a sequence $(X_n)_{n}$ is monotone non-decreasing if $X_n \leq X_{n+1}$ for all $n$.)

**Proposition 15.6.** Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, and $X$ a random variable. If $X$ is non-negative, then there exists a sequence $(X_n)_{n}$ of simple random variables such that $X_n \nearrow X$.

**Proof.** For all $n, k$ define $A_{n,k} := X^{-1}[k2^{-n}, (k+1)2^{-n})$, which is of course an event. Note that for all $n$,

$$\Omega = \bigcup_{k=0}^{2^n - 1} A_{n,k} \bigcup X^{-1}[n, \infty).$$

For all $n \geq 1$ define

$$X_n = \sum_{k=0}^{2^n - 1} 2^{-n}k1_{A_{n,k}} + n1_{X^{-1}[n, \infty)}.$$

[That is, we divide the interval $[0, n)$ into intervals of length $2^{-n}$, and replace $X$ with the leftmost value in that interval, cutting this after $n$. See Figure 4]

So $(X_n)_{n}$ is a sequence of simple random variables.
Note that $X_n(\omega) = \min\{n, 2^{-n}\lfloor 2^n X(\omega) \rfloor\}$. If $X(\omega) < n$ then

$$X_n(\omega) = 2^{-n} \lfloor 2^n X(\omega) \rfloor = 2^{-n} \lfloor \frac{1}{2} 2^{n+1} X(\omega) \rfloor \leq 2^{-(n+1)} \lfloor 2^{n+1} X(\omega) \rfloor = X_{n+1}(\omega).$$

(For all $x$, if $\lfloor x/2 \rfloor \leq x/2$ then $2 \lfloor x/2 \rfloor \leq x$.) If $n \leq X(\omega) < n+1$ then $2^{n+1} n \leq 2^{n+1} X(\omega)$ so

$$X_n(\omega) = n \leq 2^{-(n+1)} \lfloor 2^{n+1} X(\omega) \rfloor = X_{n+1}(\omega).$$

Finally, if $X(\omega) \geq n+1$, then $X_n(\omega) = n \leq n + 1 = X_{n+1}(\omega)$. So, $X_n(\omega) \leq X_{n+1}(\omega)$ for all $\omega$ and all $n$, and $(X_n)_n$ is a non-decreasing sequence.

Now to show that $\lim_n X_n(\omega) = X(\omega)$: Note that $|X(\omega) - 2^{-n} \lfloor 2^n X(\omega) \rfloor| \leq 2^{-n}$. Thus, for all $\omega \in \Omega$, $\lim_{n \to \infty} X_n(\omega) = X(\omega)$. \hfill $\square$

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{figure4.png}
\caption{$X_n$ and $X_{n+1}$}
\end{figure}
15.3. Expectation of Non-Negative Random Variables

We want to define the expectation on a random variable by approximating by simple random variables as in Proposition 15.6, and taking the limit. For this we have to make sure the limit exists and makes sense.

Let us start by defining the expectation for a simple random variable.

**Definition 15.7.** If \( X = \sum_{k=1}^{m} a_k 1_{X^{-1}(a_k)} \) is a simple random variable, define

\[
\mathbb{E}[X] := \sum_{k=1}^{m} a_k \mathbb{P}[X = a_k].
\]

**Proposition 15.8.** The above definition does not depend on the choice of the partition of \( \Omega \); that is, if \( X = \sum_{j=1}^{n} b_j 1_{B_j} \) where \( \Omega = \biguplus_{j=1}^{n} B_j \) are a partition of \( \Omega \), then \( \mathbb{E}[X] = \sum_{j=1}^{n} b_j \mathbb{P}[B_j] \).

**Proof.** Suppose that

\[
X = \sum_{k=1}^{m} a_k 1_{X^{-1}(a_k)} = \sum_{j=1}^{n} b_j 1_{B_j},
\]

where \( \Omega = \biguplus_{j=1}^{m} B_j \) are a partition of \( \Omega \) as in the statement of the proposition. We assume without loss of generality that \( a_m = 0 \) so \( \Omega = \biguplus_{k=1}^{m} \{X = a_k\} \). Then, for any \( k, j \), if \( \omega \in \{X = a_k\} \cap B_j \) then \( a_k = X(\omega) = b_j \). Otherwise, if \( \{X = a_k\} \cap B_j = \emptyset \) the \( \mathbb{P}[X = a_k, B_j] = 0 \). Thus, for all \( k, j \),

\[
a_k \mathbb{P}[X = a_k, B_j] = b_j \mathbb{P}[X = a_k, B_j].
\]

Since \( \Omega = \biguplus_{j=1}^{n} B_j = \biguplus_{k=1}^{m} \{X = a_k\} \), by the law of total probability we obtain,

\[
\mathbb{E}[X] = \sum_{k=1}^{m} a_k \mathbb{P}[X = a_k] = \sum_{k=1}^{m} \sum_{j=1}^{n} a_k \mathbb{P}[X = a_k, B_j]
= \sum_{j=1}^{n} \sum_{k=1}^{m} b_j \mathbb{P}[X = a_k, B_j] = \sum_{j=1}^{n} b_j \mathbb{P}[B_j].
\]

\( \square \)

**Proposition 15.9.** If \( X, Y \) are simple random variables, then:

- If \( X \leq Y \) then \( \mathbb{E}[X] \leq \mathbb{E}[Y] \).
- For all \( a \geq 0 \), \( \mathbb{E}[aX + Y] = a \mathbb{E}[X] + \mathbb{E}[Y] \).
Specifically we have for a simple random variable $X$,

$$
\mathbb{E}[X] = \sup \{ \mathbb{E}[S] : 0 \leq S \leq X, S \text{ is simple} \} .
$$

**Proof.** Assume that $X = \sum_k a_k 1_{\{X = a_k\}}$ and $Y = \sum_j b_j 1_{\{Y = b_j\}}$. If $\omega \in \Omega$ is such that $X(\omega) = a_k$ and $Y(\omega) = b_j$ then $a_k \leq b_j$. So, $a_k \mathbb{P}[X = a_k, Y = b_j] \leq b_j \mathbb{P}[X = a_k, Y = b_j]$. Thus,

$$
\mathbb{E}[X] = \sum_{k,j} a_k \mathbb{P}[X = a_k, Y = b_j] \leq \sum_{k,j} b_j \mathbb{P}[X = a_k, Y = b_j] = \mathbb{E}[Y].
$$

Now, note that if $Z = aX + Y$ for some $a \geq 0$, then

$$
Z = \sum_{k,j} (a \cdot a_k + b_j) 1_{\{X = a_k, Y = b_j\}}
$$

since $\Omega = \uplus_k \{X = a_k\} \uplus \{X = 0\}$ and $\Omega = \uplus_j \{Y = b_j\} \uplus \{Y = 0\}$. So

$$
\mathbb{E}[Z] = \sum_{k,j} (a \cdot a_k + b_j) \mathbb{P}[X = a_k, Y = b_j] = a \mathbb{E}[X] + \mathbb{E}[Y].
$$

$\square$

**Definition 15.10.** If $X$ is a general non-negative random variable, define

$$
\mathbb{E}[X] := \sup \{ \mathbb{E}[S] : 0 \leq S \leq X, S \text{ is simple} \} ,
$$

(where we allow the value $\mathbb{E}[X] = \infty$, and in this case we say $X$ has infinite expectation).

This definition coincides with the previous definition for simple random variables by the proposition above.

Note that the definition immediately implies that

$$
0 \leq X \leq Y \quad \Rightarrow \quad \mathbb{E}[X] \leq \mathbb{E}[Y]
$$

because the supremum for $Y$ is over a larger set.
15.4. Monotone Convergence

**Theorem 15.11** (Monotone Convergence). Let \((X_n)_n\) be a non-decreasing sequence of non-negative random variables. If \(X_n \nearrow X\) then \(E[X_n] \nearrow E[X]\).

**Proof.** The proof has four cases:

- **Case 1:** \(X_n\) are simple, \(X = 1_A\).
  For this case, let \(\epsilon > 0\), and set \(A_n := \{X_n > 1 - \epsilon\}\).
  If \(\omega \in A\) then \(X(\omega) = 1\), then since \(X_n(\omega) \nearrow X(\omega)\), then \(X_n(\omega) > 1 - \epsilon\) for some \(n\). On the other hand, if \(X_n(\omega) > 1 - \epsilon\), then \(X(\omega) > 1 - \epsilon\) and so \(X(\omega) = 1\) and \(\omega \in A\). Thus, \((A_n)_n\) is an increasing sequence of events, such that \(\lim n A_n = \bigcup_n A_n = A\).
  Since \(X_n\) are non-negative, \(X_n \geq (1 - \epsilon)\mathbf{1}_A\). Hence, \((1 - \epsilon)\mathbb{P}[A_n] \leq E[X_n] \leq \mathbb{E}[X] = \mathbb{P}[A]\). Continuity of probability finishes this case.

- **Case 2:** \(X_n, X\) simple.
  Assume \(X = \sum_k a_k \mathbf{1}_{\{X = a_k\}}\). Fix \(k\) and consider the sequence \(Y_n := a_k^{-1} \mathbf{1}_{\{X = a_k\}} X_n\).
  \(X_n \nearrow X\) implies that \(Y_n \nearrow \mathbf{1}_{\{X = a_k\}}\). By Case 1, since \(Y_n\) are simple, we have that \(E[Y_n] \nearrow \mathbb{P}[X = a_k]\).
  By linearity of expectation for simple random variables,
  \[
  E[X_n] = \sum_k a_k \cdot E[a_k^{-1} \mathbf{1}_{\{X = a_k\}} X_n] \nearrow \sum_k a_k \mathbb{P}[X = a_k] = E[X].
  \]

- **Case 3:** \(X_n\) simple, \(X\) general (non-negative).
  Let \(0 \leq S \leq X\) be a simple random variable, and \(Y_n := \min \{X_n, S\}\). \(X_n \nearrow X\) implies that \(Y_n \nearrow S\). So Case 2 and monotonicity of expectation for simple random variables give that \(E[X_n] \geq E[Y_n] \nearrow E[S]\). Thus, \(\lim_n E[X_n] \geq E[S]\) for all simple \(S \leq X\). Taking the supremum we are done.

- **Case 4:** \(X_n, X\) general.
  Define \(Y_n := \min \{2^{-n}[2^n X_n], n\}\). So \(Y_n\) are all simple, and \(Y_n \leq X_n \leq X\), and so \(E[Y_n] \leq E[X_n] \leq E[X]\). If we show that \(Y_n \nearrow X\) then we are done by Case 3.
  As in Proposition 15.6, \(Y_n \leq \min \{2^{-(n+1)}[2^{n+1} X_n], n+1\} \leq Y_{n+1}\), so \((Y_n)_n\) is a non-decreasing sequence.
For all $\omega$ and $n$ such that $X_n(\omega) < n$, we have that $|Y_n(\omega) - X_n(\omega)| \leq 2^{-n}$. So, for any $\omega$, and for any $\varepsilon > 0$, if $n > n_0(\omega, \varepsilon)$ is large enough, then $|Y_n(\omega) - X(\omega)| < \varepsilon$. That is, $Y_n \nearrow X$.

\[ \blacksquare \]

15.5. Linearity

**Lemma 15.12.** Let $X, Y$ be a non-negative random variables. Then, for any $a > 0$,

- $X \leq Y$ implies $\mathbb{E}[X] \leq \mathbb{E}[Y]$.
- $\mathbb{E}[aX + Y] = a \mathbb{E}[X] + \mathbb{E}[Y]$.

*Proof.* The first assertion is immediate from the definition, since if $0 \leq S \leq X$ is simple then $0 \leq S \leq Y$, so we are taking a supremum over a larger set.

Let $X_n \nearrow X$ and $Y_n \nearrow Y$ be sequences of simple random variables converging everywhere monotonely to $X, Y$ respectively, as in Proposition 15.6. So $aX_n + Y_n \nearrow aX + Y$.

Monotone convergence tells us that

$$
\mathbb{E}[X_n] \nearrow \mathbb{E}[X] \quad \mathbb{E}[Y_n] \nearrow \mathbb{E}[Y] \quad \text{and} \quad \mathbb{E}[aX_n + Y_n] \nearrow \mathbb{E}[aX + Y].
$$

Since for all $n$, $\mathbb{E}[aX_n + Y_n] = a \mathbb{E}[X_n] + \mathbb{E}[Y_n]$, we are done. \[ \blacksquare \]

15.6. Expectation

Finally, let us define the expectation of a general random variable.

First, some notation. Let $X$ be a random variable. Define $X^+ := \max \{X, 0\}$ and $X^- := \max \{-X, 0\}$. Note that $X^+, X^-$ are non-negative random variables, and that $X = X^+ - X^-$ and $|X| = X^+ + X^-$. 

**Definition 15.13.** Let $X$ be a random variable. If $\mathbb{E}[X^+] = \mathbb{E}[X^-] = \infty$ we say that the expectation of $X$ is not defined (or $X$ does not have expectation). If at least one of $\mathbb{E}[X^+], \mathbb{E}[X^-]$ is finite then define the **expectation** of $X$ by

$$
\mathbb{E}[X] := \mathbb{E}[X^+] - \mathbb{E}[X^-].
$$
15.6.1. **Properties of Expectation.** The most important property of expectation is linearity:

**Theorem 15.14.** Let $X, Y$ be random variables on some probability space $(\Omega, \mathcal{F}, \mathbb{P})$ such that $\mathbb{E}[X], \mathbb{E}[Y]$ exist. Then:

- For all $a \in \mathbb{R}$, $\mathbb{E}[aX + Y] = a \mathbb{E}[X] + \mathbb{E}[Y]$.
- $X \leq Y$ implies that $\mathbb{E}[X] \leq \mathbb{E}[Y]$.
- $|\mathbb{E}[X]| \leq \mathbb{E}[|X|]$.

**Proof.** For linearity, note that $(X + Y)^+ - (X + Y)^- = X + Y - X^- + Y^- - Y^-$, so $(X + Y)^+ + X^- + Y^- = (X + Y)^- + X^+ + Y^-$. Since both sides are linear combinations of non-negative random variables,

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

which implies that

$$\mathbb{E}[X + Y] = \mathbb{E}(X + Y)^+ - \mathbb{E}(X + Y)^- = \mathbb{E}X^+ - \mathbb{E}X^- + \mathbb{E}Y^+ - \mathbb{E}Y^- = \mathbb{E}X + \mathbb{E}Y.$$  

Also, for any $a > 0$, $(aX)^+ = aX^+$ and $(aX)^- = aX^-$. If $a < 0$ then $(aX)^+ = -aX^-$ and $(aX)^- = -aX^+$. Thus, in both cases $\mathbb{E}[aX] = a \mathbb{E}[X]$. This proves linearity.

For the second assertion, let $Z = Y - X$. So $Z \geq 0$, and thus $\mathbb{E}[Z] \geq 0$. Since $\mathbb{E}[Z] = \mathbb{E}[Y] - \mathbb{E}[X]$ by linearity, we are done.

Finally, for the last assertion

$$|\mathbb{E}[X]| = |\mathbb{E}[X^+] - \mathbb{E}[X^-]| \leq \mathbb{E}[X^+] + \mathbb{E}[X^-] = \mathbb{E}[|X|],$$

where we have used the fact that $X^+ \geq 0$ and $X^- \geq 0$. □
16.1. EXPECTATION - DISCRETE CASE

**Proposition 16.1.** Let \( X \) be a discrete random variable, with range \( R \) and density \( f_X \). Then,

\[
E[X] = \sum_{r \in R} rf_X(r).
\]

**Proof.** For all \( N \), let

\[
X^+_N := \sum_{r \in R \cap [0,N]} 1_{\{X=r\}} r \quad \text{and} \quad X^-_N := -\sum_{r \in R \cap [-N,0]} 1_{\{X=r\}} r.
\]

Note that \( X^+_N \nearrow X^+ \) and \( X^-_N \nearrow X^- \). Moreover, by linearity

\[
E[X^+_N] = \sum_{r \in R \cap [0,N]} \mathbb{P}[X = r] r \quad \text{and} \quad E[X^-_N] = -\sum_{r \in R \cap [-N,0]} \mathbb{P}[X = r] r.
\]

Using monotone convergence we get that

\[
E[X] = \lim_{N \to \infty} E[X^+_N] - \lim_{N \to \infty} E[X^-_N] = \sum_{r \in R} f_X(r) r.
\]

\[\square\]

**Examples:**

- Ber, Bin, Poi, Geo

**Example 16.2.** Let us calculate the expectations of different discrete random variables:

- If \( X \sim \text{Ber}(p) \) then

\[
E[X] = 1 \cdot \mathbb{P}[X = 1] + 0 \cdot \mathbb{P}[X = 0] = p.
\]
• If $X \sim \text{Bin}(n, p)$ then since $\binom{n}{k} \cdot k = n \cdot \binom{n-1}{k-1}$ for $1 \leq k \leq n$,

$$
\mathbb{E}[X] = \sum_{k=0}^{n} f_X(k) \cdot k = \sum_{k=0}^{n} \binom{n}{k} p^k (1-p)^{n-k} \cdot k
$$

$$
= np \cdot \sum_{k=1}^{n} \binom{n-1}{k-1} p^{k-1} (1-p)^{n-k} = np.
$$

An easier way, would be to note that $X = \sum_{k=1}^{n} X_k$ where $X_k \sim \text{Ber}(p)$ (and in fact $X_1, \ldots, X_n$ are independent). Thus, by linearity, $\mathbb{E}[X] = \sum_{k=1}^{n} \mathbb{E}[X_k] = np.$

• For $X \sim \text{Poi}(\lambda)$:

$$
\mathbb{E}[X] = \sum_{k=0}^{\infty} e^{-\lambda} \frac{\lambda^k}{k!} \cdot k = \lambda \cdot \sum_{k=1}^{\infty} e^{-\lambda} \frac{\lambda^{k-1}}{(k-1)!} = \lambda.
$$

• For $X \sim \text{Geo}(p)$: Note that for $g(x) = -(1-x)^k$, we have $\frac{\partial}{\partial p} g(x) = k(1-x)^{k-1}$.

$$
\mathbb{E}[X] = \sum_{k=1}^{\infty} f_X(k) \cdot k = \sum_{k=1}^{\infty} (1-p)^{k-1} p \cdot k
$$

$$
= -p \cdot \sum_{k=1}^{\infty} \frac{\partial}{\partial p} (1-p)^k = -p \cdot \frac{\partial}{\partial p} \left( \sum_{k=1}^{\infty} (1-p)^k \right)
$$

$$
= -p \cdot \frac{\partial}{\partial p} \left( \frac{1}{p} - 1 \right) = p \cdot \frac{1}{p^2} = \frac{1}{p}.
$$

**Do this one**

Another way: Let $E = \mathbb{E}[X]$. Then,

$$
E = p + \sum_{k=2}^{\infty} (1-p)^{k-1} pk = p + \sum_{k=2}^{\infty} (1-p)^{k-1} p(k-1) + \sum_{k=2}^{\infty} (1-p)^{k-1} p = p + (1-p)E + 1 - p.
$$

So $E = 1 + (1-p)E$ or $E = 1/p$.

**Example 16.3.** A pair of independent fair dice are tossed. What is the expected number of tosses needed to see Shesh-Besh?

Note that each toss of the dice is an independent trial, such that the probability of seeing Shesh-Besh is $2/36 = 1/18$. So if $X = \text{number of tosses until Shesh-Besh}$, then $X \sim \text{Geo}(1/18)$. Thus, $\mathbb{E}[X] = 18.$
16.1.1. **Function of a random variable.** Let $g : \mathbb{R}^d \to \mathbb{R}$ be a measurable function. Let $(X_1, \ldots, X_d)$ be a joint distribution of $d$ discrete random variables with range $R$ each. Then, $Y = g(X_1, \ldots, X_d)$ is a random variable. What is its expectation?

Well, first we need the density of $Y$: For any $y \in \mathbb{R}$ we have that

$$P[Y = y] = P[(X_1, \ldots, X_d) \in g^{-1}\{y\}] = \sum\limits_{(r_1, \ldots, r_d) \in R^d \cap g^{-1}\{y\}} P[(X_1, \ldots, X_d) = (r_1, \ldots, r_d)].$$

So, if $R_Y := g(R^d)$, then since

$$R^d = \bigcup_{y \in R_Y} R^d \cap g^{-1}\{y\},$$

and since $Y$ is discrete we get that

$$E[g(X_1, \ldots, X_d)] = \sum\limits_{y \in R_Y} y P[Y = y] = \sum\limits_{y \in R_Y} \sum\limits_{\vec{r} \in R^d \cap g^{-1}\{y\}} f_{X_1, \ldots, X_d}(\vec{r}) \cdot g(\vec{r}) = \sum\limits_{\vec{r} \in R^d} f_{X_1, \ldots, X_d}(\vec{r}) \cdot g(\vec{r}).$$

Specifically,

**Proposition 16.4.** Let $g : \mathbb{R}^d \to \mathbb{R}$ be a measurable function, and $X = (X_1, \ldots, X_d)$ a discrete joint random variable with range $R^d$. Then,

$$E[g(X)] = \sum\limits_{\vec{r} \in R^d} f_X(\vec{r}) \cdot g(\vec{r}).$$

**Example 16.5.** Manchester United plans to earn some money selling Wayne Rooney jerseys. Each jersey costs the club $x$ pounds, and is sold for $y > x$ pounds. Suppose that the number of people who want to buy jerseys is a discrete random variable with range $\mathbb{N}$.

What is the expected profit if the club orders $N$ jerseys?

How many jerseys should be ordered in order to maximize this profit?

**Solution.** Let $p_k = f_X(k) = P[X = k]$. Let $g_N : \mathbb{N} \to \mathbb{R}$ be the function that gets the number of people that want to buy jerseys and gives the profit, if the club ordered $N$
jerseys. That is,
\[ g_n(k) = \begin{cases} 
ky - Nx & k \leq N \\
N(y - x) & k > N 
\end{cases} \]

The expected profit is then
\[ \mathbb{E}[g_N(X)] = \sum_{k=0}^{\infty} p_k g_N(k) = \sum_{k=0}^{\infty} p_k N(y - x) - \sum_{k=0}^{N} p_k(N - k)y = N(y - x) - \sum_{k=0}^{N} p_k(N - k)y. \]

Call this \( G(N) := N(y - x) - \sum_{k=0}^{N} p_k(N - k)y. \) We want to maximize this as a function of \( N. \) Note that
\[ G(N + 1) - G(N) = y - x - \sum_{k=0}^{N} y p_k(N + 1 - N) = y - x - y \mathbb{P}[X \leq N]. \]

So \( G(N + 1) > G(N) \) as long as \( \mathbb{P}[X \leq N] < \frac{y - x}{y}, \) so the club should order \( N + 1 \) jerseys for the largest \( N \) such that \( \mathbb{P}[X \leq N] < \frac{y - x}{y}. \)

Example 16.6. Some more calculations:

- \( X \sim H(N, m, n) \) (recall that \( X \) is the number of “special” objects chosen, when choosing uniformly \( n \) objects out of \( N \) objects, with a total of \( m \) special objects).

  We use the fact that
  \[ \binom{m}{k} \cdot k = \binom{m-1}{k-1} \cdot m \quad \text{and} \quad \binom{N}{n} = \binom{N-1}{n-1} \cdot \frac{N}{n}. \]

  \[ \mathbb{E}[X] = \binom{N}{n}^{-1} \cdot \sum_{k=0}^{n \wedge m} k \cdot \binom{m}{k} \cdot \frac{(N - m)}{n - k} \]
  \[ = \binom{N-1}{n-1}^{-1} \cdot \frac{n}{N} \cdot m \cdot \sum_{k=1}^{m \wedge n} \binom{m-1}{k-1} \cdot \frac{(N - 1) - (m - 1)}{(n-1) - (k-1)} = \frac{nm}{N}. \]

- \( X \sim NB(m, p) \) (the number of trials until the \( m \)-th success). Using
  \[ \binom{k - 1}{m - 1} \cdot k = \binom{k}{m} \cdot m \]
  with \( j = k + 1 \) and \( n = m + 1, \)

  \[ \mathbb{E}[X] = \sum_{k=m}^{\infty} k \cdot \binom{k-1}{m-1} p^m(1 - p)^{k-m} = mp^{-1} \cdot \sum_{j=n}^{\infty} \binom{j-1}{n-1} p^n(1 - p)^{j-n} = \frac{m}{p}. \]
17.1. EXPECTATION - CONTINUOUS CASE

Our goal is now to prove the following theorem:

\[ \text{Goal is } E(g(X)) \]

**Theorem 17.1.** Let \( X = (X_1, \ldots, X_d) \) be an absolutely continuous random variable, and let \( g : \mathbb{R}^d \rightarrow \mathbb{R} \) be a measurable function. Then,

\[ E[g(X)] = \int_{\mathbb{R}^d} g(\mathbf{x}) f_X(\mathbf{x}) d\mathbf{x}. \]

The main lemma here is

**Lemma 17.2.** Let \( X = (X_1, \ldots, X_d) \) be an absolutely continuous random variable. Then, for any Borel set \( B \in \mathcal{B}^d \),

\[ P[X \in B] = \int_B f_X(\mathbf{x}) d\mathbf{x}. \]

**Proof.** Let \( Q(B) = \int_B f_X(\mathbf{x}) d\mathbf{x} \). Then \( Q \) is a probability measure on \((\mathbb{R}^d, \mathcal{B}^d)\), that coincides with \( P_X \) on the \( \pi \)-system of rectangles \((-\infty, b_1] \times \cdots \times (-\infty, b_d] \). Thus, \( P[X \in B] = P_X(B) = Q(B) \) for all \( B \in \mathcal{B}^d \). \( \square \)

**Remark 17.3.** We have not really defined the integral \( \int_B f_X(\mathbf{x}) d\mathbf{x} \). However, for our purposes, we can define it as \( P[X \in B] \), and note that this coincides with the Riemann integral on intervals.

**Proof of Theorem 17.1.** First assume that \( g \geq 0 \), so \( \mathbb{R}^d = g^{-1}[0, \infty) \). For all \( n \) define \( Y_n = 2^{-n} [2^n g(X)] \) which are discrete non-negative random variables.

First, we show that

\[ E[Y_n] = \int_{\mathbb{R}^d} 2^{-n} [2^n g(\mathbf{x})] f_X(\mathbf{x}) d\mathbf{x}. \]
Indeed, for \( n, k \geq 0 \) let \( B_{n,k} = g^{-1}[2^{-n}k, 2^{-n}(k + 1)) \in \mathcal{B}^d \).

Note that

\[
Y_n = \sum_{k=0}^{\infty} 2^{-n}k \mathbf{1}_{\{X \in B_{n,k}\}} \quad \text{and} \quad \mathbb{E}[Y_n] = \sum_{k=0}^{\infty} 2^{-n}k \mathbb{P}[X \in B_{n,k}].
\]

Now, since

\[
\mathbb{R}^d = g^{-1}[0, \infty) = \bigcup_{k=0}^{\infty} g^{-1}[2^{-n}k, 2^{-n}(k + 1)) = \bigcup_{k=0}^{\infty} B_{n,k},
\]

and since

\[
1_{B_{n,k}} 2^{-n} \lfloor 2^n g \rfloor f_X = 1_{B_{n,k}} 2^{-n} k f_X,
\]

we have by the lemma above that

\[
\int_{\mathbb{R}^d} 2^{-n} \lfloor 2^n g(\bar{x}) \rfloor f_X(\bar{x}) d\bar{x} = \int_{\mathbb{R}^d} \sum_{k=0}^{\infty} 1_{B_{n,k}} 2^{-n} \lfloor 2^n g(\bar{x}) \rfloor f_X(\bar{x}) d\bar{x} = \sum_{k=0}^{\infty} 2^{-n}k \int_{B_{n,k}} f_X(\bar{x}) d\bar{x}
\]

\[
= \sum_{k=0}^{\infty} 2^{-n}k \mathbb{P}[X \in B_{n,k}] = \mathbb{E}[Y_n].
\]

Here we have used the fact that

\[
\sum_{k=0}^{N} 1_{B_{n,k}} 2^{-n}k \not\geq \sum_{k=0}^{\infty} 1_{B_{n,k}} 2^{-n}k.
\]

Since \( |2^{-n}\lfloor 2^n g \rfloor - g| \leq 2^{-n} \), we get that \( |Y_n - g(X)| \leq 2^{-n} \). Thus,

\[
\left| \mathbb{E}[g(X)] - \int_{\mathbb{R}^d} g(\bar{x}) f_X(\bar{x}) d\bar{x} \right| \leq |\mathbb{E}[g(X)] - \mathbb{E}[Y_n]| + \left| \int_{\mathbb{R}^d} g(\bar{x}) f_X(\bar{x}) d\bar{x} - \int_{\mathbb{R}^d} 2^{-n} \lfloor 2^n g(\bar{x}) \rfloor f_X(\bar{x}) d\bar{x} \right|
\]

\[
\leq 2 \cdot 2^{-n} \to 0.
\]

This proves the theorem for non-negative functions \( g \).

Now, if \( g \) is a general measurable function, consider \( g = g^+ - g^- \). Since \( g^+, g^- \) are non-negative, we have that

\[
\mathbb{E}[g(X)] = \mathbb{E}[g^+(X) - g^-(X)] = \mathbb{E}[g^+(X)] - \mathbb{E}[g^-(X)] = \int_{\mathbb{R}^d} (g^+(\bar{x}) - g^-(\bar{x})) f_X(\bar{x}) d\bar{x}.
\]

\( \square \)
Corollary 17.4. Let $X$ be an absolutely continuous random variable with density $f_X$. Then,

$$E[X] = \int_{-\infty}^{\infty} t f_X(t) dt.$$ 

✓ Compare $\int tf_X(t) dt$ to $\sum r P[X = r]$ in the discrete case. This is another place where $f_X$ is “like” $P[X = \cdot]$ (although the latter is identically 0 in the continuous case, as we have seen).

Example 17.5. Expectations of some absolutely continuous random variables:

- $X \sim U[0, 1]: E[X] = \int_0^1 t dt = 1/2$.
- More generally, $X \sim U[a, b]$:
  $$E[X] = \int_a^b t \cdot \frac{1}{b-a} dt = \frac{1}{2(b-a)}(b^2 - a^2) = \frac{b+a}{2}.$$
- $X \sim \text{Exp}(\lambda)$: We use integration by parts, since $\int e^{-\lambda t} = -\lambda^{-1}e^{-\lambda t}$,
  $$E[X] = \int_0^\infty t \cdot \lambda e^{-\lambda t} dt = -te^{-\lambda t}\bigg|_0^\infty + \int_0^\infty e^{-\lambda t} dt = \frac{1}{\lambda}.$$
- $X \sim N(\mu, \sigma)$:
  $$E[X] = \int_{-\infty}^{\infty} \frac{t}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(t-\mu)^2}{2\sigma^2}\right) dt.$$ 
Change $u = t - \mu$ so $du = dt$,
  $$E[X] = \frac{1}{\sqrt{2\pi}\sigma} \cdot \int_{-\infty}^{\infty} u \exp\left(-\frac{u^2}{2\sigma^2}\right) du + \int_{-\infty}^{\infty} \mu f_X(t) dt.$$ 
Since the function $u \mapsto u \exp\left(-\frac{u^2}{2\sigma^2}\right)$ is an odd function, its integral is 0, so
  $$E[X] = \mu \int_{-\infty}^{\infty} f_X(t) dt = \mu.$$ 
A simpler way is to notice that $\frac{X-\mu}{\sigma} \sim N(0, 1)$ so
  $$E\left[\frac{X-\mu}{\sigma}\right] = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} te^{-t^2/2} dt = 0,$$ 
as an integral over an odd function. $E[X] = \mu$ follows from linearity of expectation.
Proposition 17.6. Let $X$ be an absolutely continuous random variable, such that $E[X]$ exists. Then,

$$E[X] = \int_0^\infty (\mathbb{P}[X > t] - \mathbb{P}[X \leq -t]) dt = \int_0^\infty (1 - F_X(t) - F_X(-t)) dt.$$  

Proof. Note that

$$\int_0^\infty \mathbb{P}[X > t] dt = \int_0^\infty \int_t^\infty f_X(s) ds dt = \int_0^\infty \int_0^s dt f_X(s) ds = \int_0^\infty s f_X(s) ds.$$  

Similarly,

$$\int_0^\infty \mathbb{P}[X \leq -t] dt = \int_0^\infty \int_{-\infty}^{-t} f_X(s) ds dt = \int_{-\infty}^0 \int_0^{-s} dt f_X(s) ds = \int_{-\infty}^0 -s f_X(s) ds.$$  

Subtracting both we have the result. \(\square\)

Exercise 17.7. Let $X$ be a discrete random variable, with range $\mathbb{Z}$ such that $E[X]$ exists. Then,

$$E[X] = \sum_{k=0}^\infty (\mathbb{P}[X > k] - \mathbb{P}[X < -k]).$$

Example 17.8. Let $X \sim N(0, 1)$. Compute $E[X^2]$.

By the above,

$$E[X^2] = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^\infty x^2 e^{-x^2/2} dx = \frac{1}{\sqrt{2\pi}} x e^{-x^2/2} \bigg|_{-\infty}^{\infty} + \frac{1}{\sqrt{2\pi}} \int_{-\infty}^\infty e^{-x^2/2} dx = 1$$

where we have used integration by parts, with $\frac{\partial}{\partial x} e^{-x^2/2} = -xe^{-x^2/2}$. \(\triangle \vee \triangle\)

17.2. Examples Using Linearity

Example 17.9 (Coupon Collector). Gilad collects “super-goal” cards. There are $N$ cards to collect altogether. Each time he buys a card, he gets one of the $N$ uniformly at random, independently.

What is the expected amount of cards Gilad needs to buy in order to collect all cards?

For $k = 0, 1, \ldots, N - 1$, let $T_k$ be the number of cards bought after getting the $k$-th new card, until getting the $(k + 1)$-th new card. That is, when Gilad has $k$ different cards, he buys $T_k$ more cards until he has $(k + 1)$ different cards.
If Gilad has $k$ different cards, then with probability $\frac{N-k}{N}$ he buys a card he does not already have. So, $T_k \sim \text{Geo}(\frac{N-k}{N})$.

Since the total number of cards Gilad buys until getting all $N$ cards is $T = T_0 + T_1 + T_2 + \cdots + T_{N-1}$, using linearity of expectation

$$E[T] = E[T_0] + E[T_1] + \cdots + E[T_{N-1}] = 1 + \frac{N}{N-1} + \cdots + N = N \cdot \sum_{k=1}^{N} \frac{1}{k}.$$  

\[\triangle \nabla \triangle\]

**Example 17.10** (Permutation fixed points). $n$ soldiers receive packages from their families. The packages get mixed up at the post office, so that each order is equally likely. Let $X$ be the number of soldiers that receive the correct package from their own family. What is the expectation of $X$?

Let us solve this by writing $X$ as a sum. Let $X_i$ be the indicator of the event that soldier $i$ receives the correct package. What is the probability that $X_i = 1$? There are $(n-1)!$ possible ordering for which this may happen, so $E[X_i] = \frac{(n-1)!}{n!} = \frac{1}{n}$.

Note that $X = \sum_{i=1}^{n} X_i$, so by linearity, $E[X] = n \cdot \frac{1}{n} = 1$.

Note also that the $X_i$’s are not independent, for example, if $i \neq j$,

$$P[X_i = 1, X_j = 1] = \frac{(n-2)!}{n!} = \frac{1}{n(n-1)} \neq \frac{1}{n^2}.$$  

\[\triangle \nabla \triangle\]

**Example 17.11.** We toss a die 100 times. What is the expected sum of all tosses?

Here it really begs to use linearity. If $X_k$ is the outcome of the $k$-th toss, and $X = \sum_{k=1}^{100} X_k$ then

$$E[X] = \sum_{k=1}^{100} E[X_k] = 100 \cdot \frac{7}{2} = 350.$$  

\[\triangle \nabla \triangle\]

**Example 17.12.** 200 random numbers are output by a computer, each distributed uniformly on $[0,1]$. What is their expected sum?

$$E[X] = 200 \cdot E[U[0,1]] = 200 \cdot \frac{1}{2} = 100.$$
Example 17.13. Let $X_n \sim U[0, 2^{-n}]$, for $n \geq 0$, and let $S_N = \sum_{k=0}^{N} X_k$. What is the expectation of $S_N$?

Linearity of expectation gives

\[ \mathbb{E}[S_N] = \sum_{k=0}^{N} \mathbb{E}[X_k] = \sum_{k=0}^{N} 2^{-(k+1)} = 1 - 2^{-(N+1)}. \]

Note that if $S_\infty = \sum_{k=0}^{\infty} X_k$, then $S_N \nearrow S_\infty$ so monotone convergence gives that $\mathbb{E}[S_\infty] = 1$.

17.3. A formula for general random variables

Theorem 17.14. Let $X$ be a random variable. Then

\[ \mathbb{E}[X^+] = \int_{0}^{\infty} (1 - F_X(t)) dt \quad \text{and} \quad \mathbb{E}[X^-] = \int_{0}^{\infty} F_X(-t) dt. \]

Thus, if $\mathbb{E}[X]$ exists then

\[ \mathbb{E}[X] = \int_{0}^{\infty} (1 - F_X(t) - F_X(-t)) dt. \]

Proof. First suppose that $X \geq 0$ and discrete. Then, if the range of $X$ is $R_X = \{0 = r_0 < r_1 < r_2 < \cdots \}$, we have for any $t \in [r_j, r_{j+1})$,

\[ \mathbb{P}[X > t] = \mathbb{P}[X \geq r_{j+1}] = \sum_{n=j+1}^{\infty} \mathbb{P}[X = r_n]. \]

Thus,

\[ \int_{r_j}^{r_{j+1}} (1 - F_X(t)) dt = (r_{j+1} - r_j) \cdot \sum_{n=j+1}^{\infty} \mathbb{P}[X = r_n]. \]

Summing over $j$ and interchange the sums we have

\[ \int_{0}^{\infty} (1 - F_X(t)) dt = \sum_{j=0}^{\infty} \int_{r_j}^{r_{j+1}} (1 - F_X(t)) dt = \sum_{j=0}^{\infty} \sum_{n=j+1}^{\infty} (r_{j+1} - r_j) \cdot \mathbb{P}[X = r_n] \]

\[ = \sum_{n=1}^{\infty} \mathbb{P}[X = r_n] \sum_{j=0}^{n-1} (r_{j+1} - r_j) = \sum_{n=1}^{\infty} r_n \mathbb{P}[X = r_n] = \mathbb{E}[X]. \]

Since for $t > 0$, the positivity of $X$ tells us that $F_X(-t) = 0$, we have the theorem for discrete non-negative random variables.
We also have the formula: for \( t \in (r_j, r_{j+1}] \),
\[
\mathbb{P}[X \geq t] = \mathbb{P}[X \geq r_{j+1}] = \sum_{n=j+1}^{\infty} \mathbb{P}[X = r_n].
\]

Thus,
\[
\int_{r_j}^{r_{j+1}} \mathbb{P}[X \geq t]dt = (r_{j+1} - r_j) \cdot \sum_{n=j+1}^{\infty} \mathbb{P}[X = r_n] = \int_{r_j}^{r_{j+1}} \mathbb{P}[X > t]dt,
\]
so
\[
(17.1) \quad \mathbb{E}[X] = \int_{0}^{\infty} \mathbb{P}[X > t]dt = \int_{0}^{\infty} \mathbb{P}[X \geq t]dt.
\]

Now, if \( X \geq 0 \) is any general non-negative random variable, let \( X_n := 2^{-n}[2^n X] \). So \( X_n \) is discrete, non-negative and \( |X_n - X| \leq 2^{-n} \). Also, because \( X_n \leq X \), for any \( t > 0 \),
\[
|F_{X_n}(t) - F_X(t)| = |\mathbb{P}[X_n \leq t] - \mathbb{P}[X \leq t]| = \mathbb{P}[X_n \leq t, X > t] \\
\leq \mathbb{P}[t < X \leq t + 2^{-n}] = F_X(t + 2^{-n}) - F_X(t) \to 0,
\]
as \( n \to \infty \) by right continuity. Since \( X_n \leq X_{n+1} \) we get that \( 1 - F_{X_n}(t) \nearrow 1 - F_X(t) \), and by monotone convergence,
\[
\mathbb{E}[X_n] = \int_{0}^{\infty} (1 - F_{X_n}(t))dt \nearrow \int_{0}^{\infty} (1 - F_X(t))dt.
\]
However, since \( |\mathbb{E}[X_n] - \mathbb{E}[X]| \leq 2^{-n} \) we obtain the theorem for general non-negative random variables.

Finally, if \( X = X^+ - X^- \) is any random variable, we have for all \( t \geq 0 \) that \( F_{X^+}(t) = \mathbb{P}[X^+ \leq t] = \mathbb{P}[0 \leq X \leq t] + \mathbb{P}[X < 0] = \mathbb{P}[X \leq t] = F_X(t) \), and since \( X^+ \geq 0 \),
\[
\mathbb{E}[X^+] = \int_{0}^{\infty} (1 - F_X(t))dt.
\]

Also, since \( X^- = (-X)^+ \), for all \( t \geq 0 \), we have \( \mathbb{P}[X^- \geq t] = \mathbb{P}[-X \geq t] = F_X(-t) \), so by (17.1),
\[
\mathbb{E}[X^-] = \int_{0}^{\infty} \mathbb{P}[X^- \geq t]dt = \int_{0}^{\infty} F_X(-t)dt.
\]
Example 17.15. Let $X$ be a random variable with distribution function

\[ F_X(t) = \begin{cases} 
0 & t < 0 \\
\frac{t+1}{2} & 0 \leq t < \frac{1}{2} \\
\frac{7}{8} & \frac{1}{2} \leq t < 10 \\
1 & 10 \leq t 
\end{cases} \]

Is $X$ discrete? Is $X$ absolutely continuous?

$X$ cannot be absolutely continuous, because it is not even continuous, since $0, \frac{1}{2}, 10$ are discontinuity points of $F_X$.

$X$ cannot be discrete because if $r$ is such that $\mathbb{P}[X = r] > 0$ then $r$ is a discontinuity point of $F_X$, so these can only be $0, \frac{1}{2}, 10$. However,

\[ \mathbb{P}[X \not\in \{0, \frac{1}{2}, 10\}] \geq \mathbb{P}[\frac{1}{8} < X \leq \frac{1}{4}] = F_X(\frac{1}{4}) - F_X(\frac{1}{8}) = \frac{5}{8} - \frac{9}{16} = \frac{1}{16} > 0. \]

Finally, let us compute the expectation of $X$. Since $X \geq 0$, we have that $X = X^+$, so

\[ \mathbb{E}[X] = \int_0^\infty (1 - F_X(t))\,dt = \int_0^{1/2} \frac{t-1}{2}\,dt + \int_{1/2}^{10} \frac{1}{8}\,dt \]

\[ = \frac{1}{2} (t - \frac{t^2}{2}) \bigg|_0^{1/2} + \frac{1}{8} \cdot \frac{10}{2} = \frac{1}{4} - \frac{1}{16} + \frac{10}{16} = \frac{11}{8}. \]
18.1. Jensen’s Inequality

Recall that a function $g : \mathbb{R} \to \mathbb{R}$ is convex on $[a, b]$ if for any $0 < \lambda < 1$, and any $x, y \in [a, b],
\begin{align*}
g(\lambda x + (1 - \lambda)y) &\leq \lambda g(x) + (1 - \lambda)g(y).
\end{align*}

For example, any twice differentiable function such that $g'' \geq 0$ is convex. e.g. $x^2$, $-\log x$.

Convex functions are continuous, and so always measurable.

First a small lemma regarding convex functions:

**Lemma 18.1.** Let $a < m < b$ and let $g : [a, b] \to \mathbb{R}$ be convex. Then, there exist $A, B \in \mathbb{R}$ such that $g(m) = Am + B$ and for all $x \in [a, b], g(x) \geq Ax + B$.

**Proof.** For all $a \leq x < m < y \leq b$ we have $m = \lambda x + (1 - \lambda)y$ where $\lambda = \frac{y - m}{y - x} \in (0, 1)$.

Thus, $g(m) \leq \lambda g(x) + (1 - \lambda)g(y)$ which implies that $\frac{g(m) - g(x)}{m - x} \leq \frac{g(y) - g(m)}{y - m}$. Thus, taking $\sup_{x < m} \frac{g(m) - g(x)}{m - x} \leq A \leq \inf_{y > m} \frac{g(y) - g(m)}{y - m}$, we have that for all $x < y \in [a, b], g(x) \geq A(x - m) + g(m) = Ax + (g(m) - Am)$.

**Theorem 18.2** (Jensen’s Inequality). Let $g : \mathbb{R} \to \mathbb{R}$ be convex on $[a, b]$. Let $X$ be a random variable such that $\mathbb{P}[a \leq X \leq b] = 1$. Then,
\begin{align*}
g(\mathbb{E}[X]) &\leq \mathbb{E}[g(X)].
\end{align*}

**Proof.** If $X$ is constant, then there is nothing to prove, so assume $X$ is not constant. Thus, $a < \mathbb{E}[X] < b$. Let $m = \mathbb{E}[X]$. Then, $g(X) \geq AX + B$ and $g(\mathbb{E}[X]) = A \mathbb{E}[X] + B$.
for some $A, B \in \mathbb{R}$. Specifically, since $g(X) \leq |A||X| + |B| \leq |A| \max \{|a|, |b|\} + |B|$ we get that $\mathbb{E}[g(X)] < \infty$, so $\mathbb{E}[g(X)]$ is well defined.

Now, since $AX + B \leq g(X)$,

$$g(\mathbb{E}[X]) = A\mathbb{E}[X] + B = \mathbb{E}[AX + B] \leq \mathbb{E}[g(X)].$$

\[\square\]

**Example 18.3.** Let $a_1, \ldots, a_n$ be $n$ positive numbers. Let $X$ be a discrete random variable with density $f_X(a_k) = \frac{1}{n}$.

Note that the function $g(x) = -\log x$ is a convex function on $(0, \infty)$ (since $g''(x) = x^{-2} > 0$). So Jensen’s inequality gives that

$$-\log \frac{1}{n} \sum_{k=1}^{n} a_k = -\log(\mathbb{E}[X]) \leq \mathbb{E}[-\log(X)] = -\frac{1}{n} \sum_{k=1}^{n} \log a_k.$$  

Exponentiating we get

$$\frac{1}{n} \sum_{k=1}^{n} a_k \geq \left( \prod_{k=1}^{n} a_k \right)^{1/n}.$$  

This is the arithmetic-geometric mean inequality.  

\[\triangle \nabla \triangle\]
Example 18.4. Let $p \geq 1$. The function $g(x) = |x|^p$ is convex, because away from 0, $g''(x) = p(p-1)|x|^{p-2}$, and at 0 we have $g(0) = 0 \leq \lambda g(x) + (1 - \lambda)g(y)$ for any $0 < \lambda < 1$ and $x, y$.

Thus, $\mathbb{E}[|X|^p] \geq |\mathbb{E}[X]|^p$ for any random variable $X$ such that $\mathbb{E}[X]$ is defined.

Furthermore, for $1 \leq q < p$, if $\mathbb{E}[|X|^p] < \infty$, then

$$\mathbb{E}[|X|^q] = (\mathbb{E}[|X|^p]^{p/q})^{q/p} \leq (\mathbb{E}[|X|^p])^{q/p} < \infty,$$

since the function $x \mapsto |x|^{p/q}$ is convex. △ ▽ △

18.2. Moments

Definition 18.5. Let $X$ be a random variable on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$.

The $n$-th moment of $X$ is defined to be $\mathbb{E}[X^n]$, if this exists and is finite. If this expectation does not exist (or is infinite), then we say that $X$ does not have a $n$-th moment.

Definition 18.6. Let $X$ be a random variable on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, such that $|\mathbb{E}[X]| < \infty$.

The variance of $X$ is defined to be

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

The standard deviation of $X$ is defined to be $\sqrt{\text{Var}[X]}$.

Proposition 18.7. $X$ has a second moment if and only if $|\mathbb{E}[X]| < \infty$ and $\text{Var}[X] < \infty$.

In fact, $\text{Var}[X] = \mathbb{E}[X^2] - (\mathbb{E}[X])^2$.

Proof. Using linearity of expectation,

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

So $\mathbb{E}[(X - a)^2] < \infty$ if and only if $-\infty < \mathbb{E}[X] < \infty$ and $\mathbb{E}[X^2] < \infty$. □

Example 18.8. Let us calculate some second moments and variances:

- $X \sim \text{Ber}(p)$:

$$\mathbb{E}[X^2] = \mathbb{P}[X = 1] \cdot 1^2 + \mathbb{P}[X = 0] \cdot 0^2 = p.$$ So $\text{Var}[X] = \mathbb{E}[X^2] - (\mathbb{E}[X])^2 = p(1-p)$. 

• $X \sim \text{Bin}(n, p)$:

We use the identity for all $2 \leq k \leq n$, $k(k-1)\binom{n}{k} = n(n-1)\binom{n-2}{k-2}$.

$$
\mathbb{E}[X(X-1)] = \sum_{k=0}^{n} k(k-1) \mathbb{P}[X = k] = \sum_{k=2}^{n} n(n-1)\binom{n-2}{k-2} p^{k-2} p^2 (1-p)^{n-k}
$$

$$
= p^2 n(n-1) \sum_{k=0}^{n-2} \binom{n-2}{k} p^k (1-p)^{n-2-k} = p^2 n^2 - p^2 n.
$$

By linearity of expectation,

$$
p^2 n(n-1) = \mathbb{E}[X^2] - \mathbb{E}[X] = \mathbb{E}[X^2] - np.
$$

So $\mathbb{E}[X^2] = p^2 n(n-1) + np$ and $\text{Var}[X] = np - p^2 n = np(1-p)$.

• $X \sim \text{Geo}(p)$:

$$
\mathbb{E}[X^2] = \sum_{k=1}^{\infty} k^2 (1-p)^{k-1} p = p + \sum_{k=2}^{\infty} (k-1+1)^2 (1-p)^{k-2} p(1-p)
$$

$$
= p + (1-p) \sum_{k=1}^{\infty} (k+1)^2 (1-p)^{k-1} p = p + (1-p) \mathbb{E}[(X+1)^2]
$$

$$
= p + (1-p) \mathbb{E}[X^2] + (1-p) + (1-p)2 \mathbb{E}[X] = 1 + \frac{2(1-p)}{p} + (1-p) \mathbb{E}[X^2].
$$

So

$$
\mathbb{E}[X^2] = \frac{1}{p} + \frac{2(1-p)}{p^2} = \frac{2 - p}{p^2}.
$$

Thus

$$
\text{Var}[X] = \mathbb{E}[X^2] - \mathbb{E}[X]^2 = \frac{1-p}{p^2}.
$$

• $X \sim \text{Poi}(\lambda)$:

$$
\mathbb{E}[X^2] = \sum_{k=0}^{\infty} k^2 \cdot e^{-\lambda} \frac{\lambda^k}{k!}
$$

$$
= \lambda \cdot \sum_{k=1}^{\infty} (k-1+1) e^{-\lambda} \frac{\lambda^{k-1}}{(k-1)!} = \lambda \cdot \sum_{k=0}^{\infty} (k+1) e^{-\lambda} \frac{\lambda^k}{k!}
$$

$$
= \lambda \cdot \mathbb{E}[X+1] = \lambda(\lambda + 1).
$$

Thus, $\text{Var}[X] = \mathbb{E}[X^2] - \mathbb{E}[X]^2 = \lambda$. 

• $X \sim \text{Exp}(\lambda)$: Use integration by parts:

$$
\mathbb{E}[X^2] = \int_0^\infty t^2 \lambda e^{-\lambda t} dt = -t^2 e^{-\lambda t}\bigg|_0^\infty + \int_0^\infty 2te^{-\lambda t} dt
$$

$$
= \frac{2}{\lambda} \mathbb{E}[X] = \frac{2}{\lambda^2}.
$$

So $\text{Var}[X] = \frac{2}{\lambda^2} - \frac{1}{\lambda^4} = \frac{1}{\lambda^2}$.

\[ \triangle \nabla \triangle \]

Exercise 18.9. Let $X$ be a random variable with $\text{Var}[X] < \infty$. Show that for any $a, b \in \mathbb{R}$,

$$
\text{Var}[aX + b] = a^2 \text{Var}[X].
$$

Solution. Let $\mu = \mathbb{E}[X]$. So $\mathbb{E}[aX + b] = a\mu + b$. Then,

$$
\text{Var}[aX + b] = \mathbb{E}[(aX + b - (a\mu + b))^2] = \mathbb{E}[(aX - a\mu)^2] = a^2 \text{Var}[X].
$$

\[ \square \]

Example 18.10. • $X \sim U[a, b]$: $Y = \frac{X - a}{b - a} \sim U[0, 1]$. (Prove this!) So

$$
\mathbb{E}[Y^2] = \int_0^1 s^2 ds = \frac{1}{3}.
$$

Thus $\text{Var}[Y] = \frac{1}{3} - \frac{1}{4} = \frac{1}{12}$. We have that

$$
\text{Var}[X] = \text{Var}[(b - a)Y + a] = (b - a)^2 \text{Var}[Y] = \frac{(b - a)^2}{12}.
$$

• $X \sim N(\mu, \sigma)$: Recall that $Y := (X - \mu)/\sigma$ has $N(0, 1)$ distribution. Thus, since $X = \sigma Y + \mu$, we have that $\text{Var}[X] = \sigma^2 \text{Var}[Y]$.

Now, using integration by parts, since $\frac{\partial}{\partial x} e^{-x^2/2} = -xe^{-x^2/2}$,

$$
\text{Var}[Y] = \mathbb{E}[Y^2] = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} t^2 e^{-t^2/2} dt
$$

$$
= -\frac{1}{\sqrt{2\pi}} \cdot te^{-t^2/2}\bigg|_{-\infty}^{\infty} + \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-t^2/2} dt = 1.
$$

So $\text{Var}[X] = \sigma^2$. (Note that this implies that $\mathbb{E}[X^2] = \sigma^2 + \mu^2$.)

\[ \triangle \nabla \triangle \]
19.1. Covariance

**Definition 19.1.** Let $X, Y$ be random variables on some probability space $(\Omega, \mathcal{F}, \mathbb{P})$ such that $\mathbb{E}[X], \mathbb{E}[Y], \mathbb{E}[XY]$ are finite. The covariance of $X$ and $Y$ is defined to be

$$\text{Cov}[X, Y] := \mathbb{E}[(X - \mathbb{E}[X]) \cdot (Y - \mathbb{E}[Y])].$$

The following are immediate:

**Proposition 19.2.** Let $X, Y, Z$ be random variables on some probability space $(\Omega, \mathcal{F}, \mathbb{P})$. Then,

- $\text{Cov}[X, Y] = \mathbb{E}[XY] - \mathbb{E}[X] \mathbb{E}[Y]$.
- $\text{Cov}[X, Y] = \text{Cov}[Y, X]$.
- $\text{Cov}[X, X] = \text{Var}[X] \geq 0$.

Where each equality holds if the relevant covariances are defined.

Note that Cov is almost an *inner product* on the vector space of random variables on $(\Omega, \mathcal{F}, \mathbb{P})$ with finite expectation.

19.2. Inner Products and Cauchy-Schwarz

**Reminder:**

Let $V$ be a vector space over $\mathbb{R}$. An inner-product on $V$ is a function $\langle \cdot, \cdot \rangle : V \times V \to \mathbb{R}$ such that

- Symmetry: For all $v, u \in V$: $\langle u, v \rangle = \langle v, u \rangle$. 
• Linearity: For all $\alpha \in \mathbb{R}$ and all $v, u, w \in V$: $< \alpha u + v, w > = \alpha < u, w > + < v, w >$.

• Positivity: For all $v \in V$, $< v, v > \geq 0$.

• Definiteness: For all $v \in V$: $< v, v > = 0$ if and only if $v = 0$.

For our purposes, we would like to replace the last condition by

• Definiteness': For all $v \in V$: $< v, v > = 0$ if and only if $< v, u > = 0$ for all $u \in V$.

A basic, but super important and fundamental result is the Cauchy-Schwarz inequality. (Denote $||v|| = \sqrt{<v, v>}$.)

**Theorem 19.3** (Cauchy-Schwarz Inequality). Let $< \cdot, \cdot >$ be an inner product on $V$ with the above modified definiteness condition. For all $v, u \in V$ we have

$$| < v, u > | \leq ||v|| \cdot ||u||.$$  

*Proof.* If $||u|| = 0$ then both sides are 0 (because of the modified definiteness condition), so we can assume without loss of generality that $||u|| > 0$. Set $\lambda = ||u||^{-2} < v, u >$. Then,

$$0 \leq < v - \lambda u, v - \lambda u > = < v, v > + \lambda^2 < u, u > - 2\lambda < v, u > = ||v||^2 - ||u||^{-2} < v, u >^2 $$

Multiplying by $||u||^2$ completes the proof. \[\square\]

19.2.1. **Two Inner Products.** Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. Consider the following vector space over $\mathbb{R}$: Let $L^2(\Omega, \mathcal{F}, \mathbb{P})$ be the space of all random variables $X$ with finite second moment.

In the exercise we show that if Cov$[X, X] = 0$ then Cov$[X, Y] = 0$ for all $Y$, and if $\mathbb{E}[X^2] = 0$ then $\mathbb{E}[XY] = 0$ for all $Y$. Thus, we have that both Cov$[X, Y]$ and $\mathbb{E}[XY]$ form (modified) inner products on $L^2(\Omega, \mathcal{F}, \mathbb{P})$.

✓ In order to get honest to goodness inner products, one needs to take random variables modulo measure 0 (for $\mathbb{E}[X, Y]$) and modulo measure 0 and additive constants for Cov.
We conclude:

**Theorem 19.4** (Cauchy-Schwarz inequality). Let $X,Y$ be random variables with finite second moment. Then,

\[ |\text{Cov}[X,Y]|^2 \leq \text{Var}[X] \cdot \text{Var}[Y] \quad \text{and} \quad |\mathbb{E}[XY]|^2 \leq \mathbb{E}[X^2] \cdot \mathbb{E}[Y^2]. \]

19.3. Correlation

**Definition 19.5.** If $X,Y$ are random variables such that $\text{Cov}[X,Y] = 0$ we say that $X,Y$ are **uncorrelated**. $X_1, \ldots, X_n$ are said to be **pairwise uncorrelated** if for any $k \neq j$, $X_k, X_j$ are uncorrelated.

We will soon see that if $X,Y$ are independent, they are uncorrelated. The opposite, however, is **not** true. Indeed,

**Example 19.6.** Let $X,Y$ have a joint distribution given by

\[
\begin{array}{c|ccc}
X \setminus Y & -1 & 0 & 1 \\
\hline
0 & 1/3 & 0 & 1/3 \\
1 & 0 & 1/3 & 0
\end{array}
\]

So $f_X(1) = 1/3$ and $f_Y(1) = f_Y(-1) = 1/3$. Hence,

\[
\mathbb{E}[XY] = \mathbb{P}[X = 1, Y = 1] - \mathbb{P}[X = 1, Y = -1] = 0,
\]
\[
\mathbb{E}[X] = 1/3 \quad \text{and} \quad \mathbb{E}[Y] = 1/3 - 1/3 = 0. \quad \text{So} \quad \text{Cov}[X,Y] = \mathbb{E}[XY] - \mathbb{E}[X] \cdot \mathbb{E}[Y] = 0. \quad \text{But} \quad X,Y \text{ are not independent, as}
\]
\[
\mathbb{P}[X = 1, Y = 1] = 0 \neq \frac{1}{3} \cdot \frac{1}{3} = \mathbb{P}[X = 1] \cdot \mathbb{P}[Y = 1].
\]

We show that independence is the fact that **any** functions of $X$ and $Y$ are uncorrelated.

**Theorem 19.7.** Let $X,Y$ be random variables on some probability space $(\Omega, \mathcal{F}, \mathbb{P})$. Then, $X,Y$ are independent, if and only if for any measurable functions $g,h$ such that $\text{Cov}[g(X), h(Y)]$ is defined, we have that $\text{Cov}[g(X), h(Y)] = 0$. 

△ △ △
Lemma 19.8. Let $X, Y$ be random variables with finite second moment. If $X, Y$ are independent then they are uncorrelated.

Proof. It suffices to show that if $X, Y$ are independent, then $E[XY] = E[X]E[Y]$.

If $X, Y$ are discrete random variables with range $R$,

$$E[XY] = \sum_{r, r' \in R} P[X = r, Y = r']r'r' = \sum_{r, r' \in R} P[X = r]P[Y = r']r'r' = E[X]E[Y].$$

Define $X_n^+ := 2^{-n}[2^n X^+]$ and $Y_n^+ := 2^{-n}[2^n Y^+]$. Since $X_n^+, Y_n^+$ are independent and discrete, we have that

$$E[X_n^+Y_n^+] = E[X_n^+]E[Y_n^+].$$

Since $X_n^+ \not	o X^+$ and $Y_n^+ \not	o Y^+$, and these are non-negative, we get by monotone convergence that

$$E[X^+Y^+] = \lim_{n \to \infty} E[X_n^+Y_n^+] = \lim_{n \to \infty} E[X_n^+]E[Y_n^+] = E[X^+]E[Y^+].$$

In a similar way, for $X_n^- := 2^{-n}[2^n X^-]$ and $Y_n^- := 2^{-n}[2^n Y^-]$, we have that $X_n^-, Y_n^-$ are independent, $X_n^-, Y_n^+$ are independent, and $X_n^+, Y_n^-$ are independent. Thus, $E[X_n^\xi Y_n^\zeta] = E[X_n^\xi]E[Y_n^\zeta]$ for any choice of $\xi, \zeta \in \{+, -\}$, and taking limits we get that

$$E[X^\xi Y^\zeta] = E[X^\xi]E[Y^\zeta].$$

Altogether,

= (E[X^+] - E[X^-]) \cdot (E[Y^+] - E[Y^-]) = E[X] \cdot E[Y].$$

□

Proof of Theorem 19.7. We repeatedly make use of the fact that $\text{Cov}[X, Y] = 0$ if and only if $E[XY] = E[X]E[Y]$. 
For the “if” direction: Let \( A_1 \in \sigma(X), A_2 \in \sigma(Y) \). So \( A_1 = X^{-1}(B_1), A_2 = Y^{-1}(B_2) \) for Borel sets \( B_1, B_2 \). For the functions \( g(x) = 1_{\{x \in B_1\}} \) and \( h(x) = 1_{\{x \in B_2\}} \) we have that

\[
P[A_1 \cap A_2] = \mathbb{E}[1_{\{X \in B_1, Y \in B_2\}}] = \mathbb{E}[g(X)h(Y)] = \mathbb{E}[g(X)]\mathbb{E}[h(Y)] = \mathbb{P}[A_1] \mathbb{P}[A_2].
\]

This was the “easy” direction.

Now, for the “only if” direction: We want to show that if \( X, Y \) are independent, then \( \mathbb{E}[g(X)h(Y)] = \mathbb{E}[g(X)]\mathbb{E}[h(Y)] \) for any measurable functions such that the above expectations exist. Since \( g(X), h(Y) \) are independent, the theorem follows from Lemma 19.8. \( \square \)

**Exercise 19.9.** Let \( X_1, \ldots, X_n \) be random variables with finite second moment. Show that

\[
\text{Var}[X_1 + \cdots + X_n] = \sum_{k=1}^{n} \text{Var}[X_k] + 2 \sum_{j<k} \text{Cov}[X_j, X_k].
\]

Deduce the Pythagorean Theorem: If \( X_1, \ldots, X_n \) are all pairwise uncorrelated, then

\[
\text{Var}[X_1 + \cdots + X_n] = \sum_{k=1}^{n} \text{Var}[X_k].
\]

The Pythagorean Theorem lets us calculate very easily the variance of a binomial random variable:

**Example 19.10.** Let \( X \sim \text{Bin}(n, p) \). We know that \( X = \sum_{k=1}^{n} Y_k \), where \( Y_1, \ldots, Y_n \) are independent \( \text{Ber}(p) \) random variables. So

\[
\text{Var}[X] = \sum_{k=1}^{n} \text{Var}[Y_k] = np(1 - p).
\]

The straightforward calculation would be more difficult. \( \triangle \nabla \triangle \)

19.4. Examples

**Example 19.11.** Let \( X \sim U[-1, 1] \) and \( Y = X^2 \).
What is the distribution of $Y$?

$$
P[Y \leq t] = P[X \in [-\sqrt{t}, \sqrt{t}]] = \begin{cases} \sqrt{t} & 0 \leq t \leq 1 \\ 0 & t < 0 \\ 1 & t \geq 1 \end{cases}
$$

Note that for

$$f_Y(t) = \begin{cases} \frac{1}{2} t^{-1/2} & t \in [0, 1] \\ 0 & t \not\in [0, 1] \end{cases}$$

we have that for any $t \in [0, 1]$,

$$\int_{-\infty}^{t} f_Y(s) ds = \int_{0}^{t} \frac{1}{2} s^{-1/2} ds = \sqrt{t} = F_Y(t),$$

so $Y$ is absolutely continuous with density $f_Y$.

Are $X,Y$ independent? Of course not, for example,

$$P[X \in [0, 1/2], Y \in [1/2, 1]] = 0 \neq \frac{1}{2} \cdot (1 - 2^{-1/2}) = P[X \in [0, 1]] \cdot P[Y \in [1/2, 1]].$$

What is Cov$(X,Y)$? Well,

$$E[XY] = E[X^3] = \int_{-1}^{1} t^3 \frac{1}{2} dt = 0.$$ 

Also, $E[X] = 0$ so

$$\text{Cov}(X,Y) = E[XY] - E[X] \cdot E[Y] = 0.$$ 

That is, $X,Y$ are uncorrelated.

What about $Z = X^{2n}$ for general $n$? In this case,

$$E[XZ] = E[X^{2n+1}] = \int_{-1}^{1} t^{2n+1} \frac{1}{2} dt = \frac{1}{2(2n+2)} t^{2n+2} \bigg|_{-1}^{1} = 0.$$ 

So $X, X^{2n}$ are uncorrelated for any $n$.

As for odd moments, if $W = X^{2n+1}$, then

$$E[XW] = \int_{-1}^{1} t^{2n+2} \frac{1}{2} dt = \frac{2}{2(2n+3)} = \frac{1}{2n+3},$$

So $X,W$ are not uncorrelated, because $\text{Cov}(X,W) = \frac{1}{2n+3}$.\[\triangle \nabla \triangle\]
Example 19.12. Let \( a \in (-1, 1) \) and

\[
A = \begin{bmatrix} 1 & a \\ a & 1 \end{bmatrix}
\]

So \(|A| = 1 - a^2\) and

\[
A^{-1} = \frac{1}{1-a^2} \cdot \begin{bmatrix} 1 & -a \\ -a & 1 \end{bmatrix}
\]

Let \((X,Y)\) be a two dimensional absolutely continuous random variable with density

\[
f_{X,Y}(t,s) = \frac{1}{2\pi \cdot \sqrt{|A|}} \cdot \exp \left( -\frac{1}{2} \langle A^{-1}(t,s), (t,s) \rangle \right) = \frac{1}{2\pi \cdot \sqrt{1-a^2}} \cdot \exp \left( -\frac{1}{2(1-a^2)} \cdot (t^2 + s^2 - 2ats) \right).
\]

First, let us show that \(f_{X,Y}\) is indeed a density. We will use the fact that \(t^2 + s^2 - 2ats = (t-as)^2 + s^2 - a^2s^2\).

\[
\int_{-\infty}^{\infty} f_{X,Y}(t,s) dt = \frac{1}{2\pi \cdot \sqrt{1-a^2}} \cdot \exp \left( -\frac{s^2(1-a^2)}{2(1-a^2)} \right) \cdot \int_{-\infty}^{\infty} \exp \left( -\frac{(t-as)^2}{2(1-a^2)} \right) dt
\]

\[
= \frac{1}{2\pi \cdot \sqrt{1-a^2}} \cdot e^{-s^2/2} \cdot \sqrt{2\pi(1-a^2)} = \frac{1}{\sqrt{2\pi}} \cdot e^{-s^2/2}.
\]

We have used that \(\frac{1}{\sqrt{2\pi(1-a^2)}} \cdot \exp \left( -\frac{(t-as)^2}{2(1-a^2)} \right)\) is the density of a \(N(as, \sqrt{1-a^2})\) random variable. So \(Y \sim N(0,1)\). Symmetrically, \(X \sim N(0,1)\).

Thus \(\int \int f_{X,Y} dtds = 1\) and \(f_{X,Y}\) is indeed a density. Moreover we have calculated the marginal densities of \(X\) and \(Y\).

Let us calculate

\[
\text{Cov}(X,Y) = \mathbb{E}[XY] = \int \int f_{X,Y}(t,s) t s dtds
\]

\[
= \int_{-\infty}^{\infty} s \cdot \frac{1}{\sqrt{2\pi}} \cdot e^{-s^2/2} \int_{-\infty}^{\infty} t \cdot \frac{1}{\sqrt{2\pi(1-a^2)}} \cdot \exp \left( -\frac{(t-as)^2}{2(1-a^2)} \right) dt ds
\]

\[
= \int_{-\infty}^{\infty} a s \cdot s f_Y(s) ds = \mathbb{E}[aY^2] = a.
\]

So now we also have that

\[
\text{Var}[X + Y] = \text{Var}[X] + \text{Var}[Y] + 2 \text{Cov}(X,Y) = 1 + 1 + 2a = 2(1 + a).
\]

Note that \(X,Y\) are independent if and only if \(\text{Cov}(X,Y) = 0\). \(\triangle \nabla \triangle\)
Example 19.13. A particle moves in the NE (nord-est) quadrant of the plane, $[0, \infty)^2$.

The direction $\theta \sim U[0, \pi/2]$ and the velocity $V \sim U[0, 1]$ independently.

What is the covariance of $X(t)$ and $V$ where the particle’s position at time $t$ is given by $(X(t), Y(t))$?

Note that for any $t$ $X(t) = V t \cos(\theta)$. So we want to calculate

$$E[X(t)V] - E[X(t)]E[V] = t E[V^2] \cdot E[\cos(\theta)] - t E[V] \cdot E[\cos(\theta)] = t \Var[V] \cdot E[\cos(\theta)].$$

We know that $\Var[V] = 1/12$. Also,

$$E[\cos(\theta)] = \int_0^{\pi/2} \frac{2}{\pi} \cos t dt = \frac{2}{\pi}.$$ 

So $\Cov(X(t), V) = \frac{t}{6\pi}$. △▽△

Example 19.14. Suppose $Y \sim \text{Exp}(\lambda)$ and for any $s > 0$, $X|Y = s \sim U[0, s]$. What is $\Cov(X, Y)$?

We can calculate

$$E[XY] = \int \int ts f_{X,Y}(t,s) dtds = \int_0^\infty s f_Y(s) \int t f_{X|Y}(t|s) dtds
= \int_0^\infty s f_Y(s) \cdot \frac{s}{2} ds = \frac{1}{2} \cdot E[Y^2] = \lambda^{-2}.$$ 

Also, $E[Y] = \lambda^{-1}$. If we consider the function $(t, s) \mapsto t$ we get

$$E[X] = \int \int t f_{X,Y}(t,s) dtds = \int_0^\infty f_Y(s) \int_0^s t f_{X|Y}(t|s) dtds = \int_0^\infty E[X|Y = s] f_Y(s) ds
= \int_0^\infty \frac{s}{2} f_Y(s) ds = \frac{1}{2} E[Y] = \frac{1}{2\lambda}.$$ 

So altogether,

$$\Cov(X, Y) = \frac{1}{\lambda^2} - \frac{1}{2\lambda} \cdot \frac{1}{\lambda} = \frac{1}{2\lambda^2}.$$ 

△▽△

19.5. Markov and Chebyshev inequalities

Theorem 19.15 (Markov’s inequality). Let $X \geq 0$ be a non-negative random variable. Then, for any $a > 0$,

$$P[X \geq a] \leq \frac{E[X]}{a}.$$
**Proof.** Consider the random variable $Y = a 1_{X \geq a}$. Note that $Y \leq X$ (here we use the non-negativity of $X$). So

$$
\mathbb{E}[X] \geq \mathbb{E}[Y] = a \mathbb{P}[X \geq a].
$$

$\square$

**Theorem 19.16** (Chebyshev’s inequality). Let $X$ be a random variable with finite second moment. Then, for any $a > 0$,

$$
\mathbb{P}[|X - \mathbb{E}[X]| \geq a] \leq \frac{\text{Var}[X]}{a^2}.
$$

**Proof.** Apply Markov to the non-negative random variable $Y = (X - \mathbb{E}[X])^2$, and note that $\{|X - \mathbb{E}[X]| \geq a\} = \{Y \geq a^2\}$. $\square$

If $X$ is a random variable with $\mathbb{E}[X] = \mu$ and $\text{Var}[X] = \sigma^2$, then Chebyshev’s inequality tells us that the probability that $X$ deviates from its mean $k\sigma$ (i.e. $k$ standard deviations) is at most $1/k^2$.

**Example 19.17.** The average wage in Eurasia is 1000 a month.

- A person is chosen at random. What is the probability that this person earns at least 10,000 a month?
  
  By Markov’s inequality, if $X$ is the random person’s salary, then

  $$
  \mathbb{P}[X \geq 10,000] \leq \frac{\mathbb{E}[X]}{10,000} = \frac{1}{10}.
  $$

- The average wage for an Inner Party member is 10,000 a month, and the standard deviation of the wage is 100. What is the probability that a randomly chosen Inner Party member earns at least 11,000?
  
  Here we have that $\mathbb{E}[X] = 10,000$ and $\text{Var}[X] = 10,000$. So

  $$
  \mathbb{P}[X \geq 11,000] \leq \mathbb{P}[|X - \mathbb{E}[X]| \geq 1,000] \leq \frac{\text{Var}[X]}{1,000^2} = \frac{1}{100}.
  $$

$\triangle \nabla \triangle$

**Example 19.18.** Let $X \sim N(0, \sigma)$. Chebyshev’s inequality tells us that $\mathbb{P}[|X| > k\sigma] \leq k^{-2}$. 

However because $\sigma^{-1}X \sim N(0,1)$, we can calculate for $a = \frac{t}{\sigma}$,

$$
\mathbb{P}[|X| > t] = \mathbb{P}[\sigma^{-1}X > a] + \mathbb{P}[\sigma^{-1}X < -a] = \int_{-\infty}^{-a} f_{\sigma^{-1}X}(s)ds + \int_{a}^{\infty} f_{\sigma^{-1}X}(s)ds
$$

$$
= 2 \int_{a}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-s^2/2}ds \leq 2 \int_{a}^{\infty} \frac{s}{a} \cdot \frac{1}{\sqrt{2\pi}} e^{-s^2/2}ds
$$

$$
= \frac{2}{a \cdot \sqrt{2\pi}} \cdot (-e^{-s^2/2}) \bigg|_{a}^{\infty} = \frac{2}{a \cdot \sqrt{2\pi}} \cdot e^{-a^2/2}.
$$

If we plug in $t = k\sigma$ we get

$$
\mathbb{P}[|X| > k\sigma] \leq \frac{2}{\sqrt{2\pi}k^2} \cdot e^{-k^2/2} \ll k^{-2},
$$

which is a much better bound. \(\triangle \nabla \triangle\)

19.6. The Weierstrass Approximation Theorem

Here is a probabilistic proof by Bernstein for the famous Weierstrass Approximation Theorem.

**Theorem 19.19** (Weierstrass Approximation Theorem). Let $f : [0,1] \to \mathbb{R}$ be continuous. For every $\varepsilon > 0$ there exists a polynomial $p(x)$ such that $\sup_{x \in [0,1]} |p(x) - f(x)| < \varepsilon$. (That is, the polynomials are dense in $L^\infty([0,1])$.)

**Proof.** By classical analysis, $f$ is uniformly continuous and bounded on $[0,1]$; that is, there exists $M > 0$ such that $\sup_{x \in [0,1]} |f(x)| \leq M$ and for every $\varepsilon > 0$ there exists $\delta = \delta(\varepsilon) > 0$ such that if $|x - y| \leq \delta$ then $|f(x) - f(y)| < \frac{\varepsilon}{2}$.

Fix $\varepsilon > 0$. Let $\delta = \delta(\varepsilon)$. Let $x \in (0,1)$. Consider $B \sim \text{Bin}(n, x)$. Chebyshev’s inequality guaranties that

$$
\mathbb{P}[|B - nx| \geq n^{2/3}] \leq \text{Var}[B] \cdot n^{-4/3} = x(1 - x) \cdot n^{-1/3} \leq \frac{1}{4n^{1/3}}.
$$

Thus, we may choose $n = n(\varepsilon, M)$ large enough so that this probability is at most $\frac{\varepsilon}{4M}$ for any $x$. Assume also that $n^{-1/3} < \delta$.

Define the polynomial

$$
p(x) := \sum_{j=0}^{n} \binom{n}{j} x^j (1 - x)^{n-j} f \left( \frac{j}{n} \right).
$$

That is, $p(x) = \mathbb{E}[f(B/n)]$ where $B \sim \text{Bin}(n, x)$. 
Let \( A = \{ \left| \frac{B}{n} - x \right| > \delta \} \). So
\[
P[A] \leq P[|B - nx| > n^{2/3}] \leq \frac{\varepsilon}{4M}.
\]

Thus,
\[
\left| p(x) - f(x) \right| = \left| E[f(B/n) - f(x)] \right|
\leq E[|f(B/n) - f(x)| \cdot 1_A] + E[|f(B/n) - f(x)| \cdot 1_{A^c}]
\leq M \cdot P[A] + \frac{\varepsilon}{2} \cdot P[A^c] \leq \frac{3\varepsilon}{4} < \varepsilon.
\]

This bound is independent of \( x \). \( \square \)

\section{The Paley-Zygmund Inequality}

\[ \textbf{Theorem 19.20 (Paley-Zygmund Inequality).} \] Let \( X \geq 0 \) be a non-negative random variable with finite variance. Then, for any \( \alpha \in [0,1] \),
\[
P[X > \alpha E[X]] \geq (1 - \alpha)^2 \cdot \frac{(E[X])^2}{E[X^2]}.
\]
Specifically,
\[
P[X > 0] \geq \frac{(E[X])^2}{E[X^2]}.
\]

\textit{Proof.} Note that
\[
X = X1_{\{X \leq \alpha E[X]\}} + X1_{\{X > \alpha E[X]\}}.
\]

Since \( X \geq 0 \) we have that \( X1_{\{X \leq \alpha E[X]\}} \leq \alpha E[X] \) and by Cauchy-Schwarz,
\[
(E[X1_{\{X > \alpha E[X]\}}])^2 \leq E[X^2] \cdot P[X > \alpha E[X]].
\]

Combining these we have,
\[
E[X] \leq \alpha E[X] + \sqrt{E[X^2]} \cdot P[X > \alpha E[X]],
\]
which implies that
\[
(1 - \alpha)^2(E[X])^2 \leq E[X^2] \cdot P[X > \alpha E[X]].
\]
\( \square \)
Example 19.21 (Erdős-Rényi random graph). Suppose that \( n \) students are on Facebook. For every two different students, say \( x, y \), they become friends with probability \( p \), all pairs of students being independent.

We say that students \( x \) and \( y \) are connected if there is some sequence of students \( x = x_1, \ldots, x_k = y \) such that \( x_j \) and \( x_{j+1} \) are friends for every \( j \).

What is the probability that some student has no friends? What is the probability that all students are connected?

Suppose \( \{1, 2, \ldots, n\} \) is the set of students. Let \( I_{x,y} \) be the indicator of the event that \( x \) and \( y \) are friends. So \( (I_{x,y})_{x \neq y} \) are \( \binom{n}{2} \) independent Bernoulli-\( p \) random variables.

Let \( X \) be the number of students with no friends; the isolated students.

Let us begin by calculating the expected number of isolated students: A student \( x \) is isolated if \( I_{x,y} = 0 \) for all \( y \neq x \). So this happens with probability \( (1 - p)^{n-1} \). Thus, by linearity,

\[
\mathbb{E}[X] = \sum_x \mathbb{P}[x \text{ is isolated}] = n(1 - p)^{n-1} =: \mu.
\]

Let us also calculate the second moment of \( X \): First, for \( x \neq y \), is both are isolated then for all \( z \notin \{x, y\} \) we have \( I_{x,z} = 0 \) and \( I_{y,z} = 0 \). Also, \( I_{x,y} = 0 \). So,

\[
\mathbb{P}[x \text{ is isolated}, y \text{ is isolated}] = (1 - p)^{2(n-2)+1} = (1 - p)^{2(n-1)}(1 - p)^{-1}.
\]

Thus,

\[
\mathbb{E}[X^2] = \mathbb{E} \sum_{x,y} 1_{\{x \text{ is isolated}, y \text{ is isolated}\}} = \sum_x \mathbb{P}[x \text{ is isolated}] + \sum_{x \neq y} \mathbb{P}[x \text{ is isolated}, y \text{ is isolated}] = \mu + n(n - 1)(1 - p)^{2(n-1)}(1 - p)^{-1} = \mu + \mu^2 \cdot (1 - p)^{-1} \cdot (1 - \frac{1}{n}).
\]

Specifically,

\[
\frac{(\mathbb{E}[X])^2}{\mathbb{E}[X^2]} = \left( \mu^{-1} + (1 - p)^{-1} \cdot (1 - \frac{1}{n}) \right)^{-1}.
\]

We will make use of the inequalities \( e^{-\frac{p}{1-p}} \leq 1 - p \leq e^{-p} \) valid for all \( p \leq \frac{1}{2} \).
Now, if \( p \geq \frac{(1+\varepsilon) \log n}{n-1} \), then

\[
E[X] \leq ne^{-p(n-1)} = n^{-\varepsilon} \to 0,
\]

so by Markov’s inequality,

\[
P[X \geq 1] \leq E[X] \to 0.
\]

That is, with high probability all students are connected.

On the other hand, if \( p \leq \frac{(1-\varepsilon) \log n}{n-1} \), then

\[
\mu \geq n \exp \left( -\frac{1}{1-p} \cdot (1 - \varepsilon) \log n \right) \to \infty
\]

(because \( \frac{1}{1-p} (1 - \varepsilon) < 1 \) for large enough \( n \)). Thus, by the Paley-Zygmund inequality,

\[
P[X > 0] \geq \frac{(E[X])^2}{E[X^2]} = \left( \mu^{-1} + (1 - p)^{-1} \cdot (1 - \frac{1}{n}) \right)^{-1} \to 1.
\]

That is, with high probability there exists a student that has no friends. △ ▽ △
20.1. CONVERGENCE OF RANDOM VARIABLES

As in the case of sequences of numbers, we would like to talk about convergence of random variables. There are many ways to say that a sequence of random variables approximates some limiting random variable. We will talk about 4 such possibilities.

20.2. A.S. AND $L^p$ CONVERGENCE

We have already seen some type of convergence; namely, $X_n \to X$ if $X_n(\omega) \to X(\omega)$ for all $\omega \in \Omega$. This is a strong type of convergence; we are usually not interested in changes that have probability 0, so we will define the following type of convergence:

Let $(X_n)_n$ be a sequence of random variables on some probability space $(\Omega, \mathcal{F}, \mathbb{P})$. We have already seen that $\lim \inf X_n, \lim \sup X_n$ are both random variables. Thus, the set

$$\{\omega : \lim \inf X_n(\omega) = \lim \sup X_n(\omega)\}$$

is an event in $\mathcal{F}$. Hence we can define:

**Definition 20.1.** Let $(X_n)_n$ be a sequence of random variables on $(\Omega, \mathcal{F}, \mathbb{P})$. Let $X$ be a random variable.

We say that $(X_n)$ converges to $X$ **almost surely**, or a.s., denoted

$$X_n \xrightarrow{\text{a.s.}} X$$

if $\mathbb{P}[\lim X_n = X] = 1$; that is, if

$$\mathbb{P}\left[\left\{\omega : \lim_{n \to \infty} X_n(\omega) = X(\omega)\right\}\right] = 1.$$

Another type of convergence is a type of average convergence:
Definition 20.2. Let $p > 0$. Let $(X_n)_n$ be a sequence of random variables on $(\Omega, \mathcal{F}, \mathbb{P})$ such that $\mathbb{E}[|X_n|^p] < \infty$ for all $n$. Let $X$ be a random variable such that $\mathbb{E}[|X|^p] < \infty$.

We say that $(X_n)$ converges to $X$ in $L^p$, denoted

$$X_n \overset{L^p}{\to} X$$

if

$$\lim_{n \to \infty} \mathbb{E}[|X_n - X|^p] = 0.$$

In the exercises we will show that for $0 < p < q$, $L^q$ convergence implies $L^p$ convergence.

We will also show that the limit is unique; that is

Exercise 20.3. Suppose that $X_n \overset{a.s.}{\to} X$ and $X_n \overset{L^p}{\to} Y$. Show that $\mathbb{P}[X = Y] = 1$.

Suppose that $X_n \overset{L^p}{\to} X$ and $X_n \overset{L^q}{\to} Y$. Show that $\mathbb{P}[X = Y] = 1$.

Example 20.4. Fix $p > 0$. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be uniform measure on $[0,1]$. For all $n$ let $X_n$ be the discrete random variable

$$X_n(\omega) = \begin{cases} n^{1/p} & \omega \in [0,1/n] \\ 0 & \text{otherwise} \end{cases}.$$ 

So $X_n$ has density

$$f_{X_n}(s) = \begin{cases} \frac{1}{n} & s = n^{1/p} \\ 1 - \frac{1}{n} & s = 0 \\ 0 & \text{otherwise} \end{cases}.$$ 

Note that

$$\mathbb{E}[|X_n|^q] = \frac{1}{n}|n^{1/p}|^q = n^{q/p-1}.$$ 

Thus, if $0 < q < p$, then $\mathbb{E}[|X_n - 0|^q] \to 0$, so $X_n \overset{L^q}{\to} 0$.

However, for $q \geq p$ we have that $\mathbb{E}[|X_n - 0|^q] \geq 1$ for all $n$, so $X_n \not\overset{L^q}{\to} 0$ (and thus $X_n \not\overset{L^p}{\to} X$ for any $X$, since the limit must be unique).

We claim that $X_n \overset{a.s.}{\to} 0$. Indeed, let $\omega \in [0,1]$. If $\omega > 0$ the for all $n > \frac{1}{\omega}$, we get that $X_n(\omega) = 0$, and thus $X_n(\omega) \to 0$. Thus,

$$\mathbb{P}[\{\omega : X_n(\omega) \not\to 0\}] \leq \mathbb{P}[\{0\}] = 0,$$
so $X_n \overset{a.s.}{\rightarrow} 0$.

This example has shown that

- We can have $L^q$ convergence without $L^p$ convergence, for $p > q$.
- We can have a.s. convergence without $L^p$ convergence.

Example 20.5. For all $n$ let $X_n$ be mutually independent Ber$(1/n)$ random variables.

Thus, for all $n$ and all $p > 0$,

$$\mathbb{E}[|X_n|^p] = \frac{1}{n} \rightarrow 0 \quad \text{so} \quad X_n \overset{L^p}{\rightarrow} 0.$$  

On the other hand, let $A_n = \{|X_n| > 1/2\}$. So $(A_n)_n$ is a sequence of mutually independent events, and $\mathbb{P}[A_n] = 1/n$. Thus, since $\sum_n \mathbb{P}[A_n] = \infty$, using the second Borel-Cantelli Lemma,

$$\mathbb{P}[\limsup_n A_n] = 1.$$  

That is, with probability 1, there exist infinitely many $n$ such that $X_n > 1/2$. Thus,

$$\mathbb{P}[\limsup_n X_n \geq 1/2] = 1.$$  

So we cannot have that $X_n \overset{a.s.}{\rightarrow} 0$. Since the limit must be unique, we cannot have $X_n \overset{a.s.}{\rightarrow} X$ for any $X$.

We conclude that: It is possible for $(X_n)$ to converge in $L^p$ for all $p$, but not to converge a.s.

20.3. Convergence in Probability

Definition 20.6. Let $(X_n)_n$ be a sequence of random variables on $(\Omega, \mathcal{F}, \mathbb{P})$. We say that $(X_n)$ converges in probability to $X$, denoted

$$X_n \overset{P}{\rightarrow} X$$

if for all $\varepsilon > 0$,

$$\lim_{n \rightarrow \infty} \mathbb{P}[|X_n - X| > \varepsilon] = 0.$$
Example 20.7. Suppose that \( X_n \overset{a.s.}{\to} X \). That is,
\[
P[|X_n - X| \to 0] = 1.
\]
Fix \( \varepsilon > 0 \). For any \( n \) let \( A_n = \{|X_n - X| > \varepsilon\} \). Note that if \( A_n \) occurs for infinitely many \( n \), then \( \limsup_n |X_n - X| \geq \varepsilon > 0 \). Thus, by Fatou’s Lemma,
\[
0 \leq \limsup_n P[A_n] = P[\limsup_n A_n] = P[A_n \ i.o.] \leq P[\limsup_n |X_n - X| > 0] = 0.
\]
Thus,
\[
\lim_n P[|X_n - X| > \varepsilon] = \limsup_n P[A_n] = 0,
\]
and so \( X_n \overset{P}{\to} X \).

That is: a.s. convergence implies convergence in probability. \( \bigtriangleup \ \nabla \ \bigtriangleup \)

Example 20.8. Assume that \( X_n \overset{L^p}{\to} X \). Then, for any \( \varepsilon > 0 \), by Markov’s inequality,
\[
P[|X_n - X| > \varepsilon] \leq \frac{E[|X_n - X|^p]}{\varepsilon^p} \to 0.
\]
Thus, \( X_n \overset{P}{\to} X \).

That is, convergence in \( L^p \) implies convergence in probability. \( \bigtriangleup \ \nabla \ \bigtriangleup \)

The following is in the exercises:

Exercise 20.9. Let \( X_n \overset{P}{\to} X \) and \( X_n \overset{P}{\to} Y \). Show that \( P[X = Y] = 1 \).
21.1. The Law of Large Numbers

In this section we will prove

**Theorem 21.1** (Strong Law of Large Numbers). Let \( X_1, X_2, \ldots, X_n, \ldots \), be a sequence of mutually independent random variables, such that \( \mathbb{E}[X_n] = 0 \) and \( \sup_n \mathbb{E}[X_n^2] < \infty \).

For each \( N \) let

\[
S_N = \sum_{n=1}^{N} X_n.
\]

Then,

\[
\frac{S_N}{N} \xrightarrow{a.s.} 0.
\]


**Lemma 21.2** (Kolmogorov’s inequality). Let \( X_1, X_2, \ldots, X_n, \ldots \), be a sequence of mutually independent random variables, such that \( \mathbb{E}[X_n] = 0 \) and \( \mathbb{E}[X_n^2] < \infty \). For each \( N \) let

\[
S_N = \sum_{n=1}^{N} X_n \quad \text{and} \quad M_N = \max_{1 \leq n \leq N} |S_n|.
\]

Then, for any \( \lambda > 0 \),

\[
\mathbb{P}[M_N \geq \lambda] \leq \frac{\sum_{n=1}^{N} \mathbb{E}[X_n^2]}{\lambda^2}.
\]

**Proof.** Fix \( \lambda > 0 \) and let

\[
A_n = \{S_1 < \lambda, S_2 < \lambda, \ldots, S_{n-1} < \lambda, S_n \geq \lambda\}.
\]

We will make a few observations.
First note that since \((X_n)\) are mutually independent, they are uncorrelated, so
\[
\mathbb{E}[S_N] = 0 \quad \text{and} \quad \text{Var}[S_N] = \mathbb{E}[S_N^2] = \sum_{n=1}^{N} \mathbb{E}[X_n^2].
\]

Next, note that \(M_N \geq \lambda\) if and only if there exists \(1 \leq n \leq N\) such that \(|S_n| \geq \lambda\) and \(|S_k| < \lambda\) for all \(1 \leq k < n\). That is, \(M_N \geq \lambda\) implies that there exists \(1 \leq n \leq N\) such that \(S_n^2 \cdot 1_{A_n} \geq \lambda^2\).

Since \((X_1, \ldots, X_n)\) are independent of \((X_{n+1}, \ldots, X_N)\), we have that the random variable \(S_n \cdot 1_{A_n}\) is independent of \(X_{n+1} + X_{n+2} + \cdots + X_N = S_N - S_n\). Thus, for any \(1 \leq n \leq N\) we have that
\[
\mathbb{E}[S_N^2 \cdot 1_{A_n}] = \mathbb{E}[(S_N - S_n + S_n)^2 \cdot 1_{A_n}] = \mathbb{E}[(S_N - S_n)^2 \cdot 1_{A_n}] + \mathbb{E}[S_n^2 \cdot 1_{A_n}] + 2 \mathbb{E}[(S_N - S_n) \cdot S_n \cdot 1_{A_n}] \\
\geq \mathbb{E}[S_n^2 \cdot 1_{A_n}],
\]
where we have used that
\[
\mathbb{E}[(S_N - S_n) \cdot S_n \cdot 1_{A_n}] = \mathbb{E}[S_N - S_n] \cdot \mathbb{E}[S_n \cdot 1_{A_n}] = 0
\]
by independence.

We now have that
\[
\mathbb{E}[S_N^2] \geq \mathbb{E}[S_N^2 \cdot 1_{\{M_N \geq \lambda\}}] = \mathbb{E}\left[\sum_{n=1}^{N} 1_{A_n} S_n^2 \right] \geq \sum_{n=1}^{N} \mathbb{E}[S_n^2 \cdot 1_{A_n}]
\]
Note that by Boole’s inequality, since \(M_N \geq \lambda\) implies that there exists \(1 \leq n \leq N\) such that \(S_n^2 \cdot 1_{A_n} \geq \lambda^2\),
\[
\mathbb{P}[M_N \geq \lambda] \leq \sum_{n=1}^{N} \mathbb{P}[S_n^2 \cdot 1_{A_n} \geq \lambda^2] \leq \frac{1}{\lambda^2} \sum_{n=1}^{N} \mathbb{E}[S_n^2 \cdot 1_{A_n}] \leq \frac{1}{\lambda^2} \mathbb{E}[S_N^2].
\]


**Lemma 21.3.** Suppose that \(x_1, \ldots, x_n, \ldots\) is a sequence of numbers such that \(\sum_{n=1}^{\infty} \frac{x_n}{n}\) converges. Then,
\[
\lim_{N \to \infty} \frac{1}{N} \sum_{n=1}^{N} x_n = 0.
\]
Proof. Let \( s_N = \sum_{n=1}^{N} \frac{x_n}{N} \). So \( s_N \to s := \sum_{n=1}^{\infty} \frac{x_n}{n} \). Thus, also \( \frac{1}{N} \sum_{n=1}^{N} s_n \to s \). Note that for all \( n \), \( x_n = n(s_n - s_{n-1}) \) (where we set \( s_0 = 0 \)). Thus,

\[
\frac{1}{N} \sum_{n=1}^{N} x_n = \frac{1}{N} (N s_N + ((N - 1) - N) s_{N-1} + ((N - 2) - (N - 1)) s_{N-2} + \cdots + (1 - 2) s_1)
\]

\[
= s_N - \frac{N - 1}{N} \cdot \frac{1}{N - 1} \sum_{n=1}^{N-1} s_n \to s - s = 0.
\]

\[\square\]


Lemma 21.4. Let \( X_1, X_2, \ldots, X_n, \ldots \) be a sequence of mutually independent random variables such that \( E[X_n] = 0 \) for all \( n \). If \( \sum_{k=1}^{\infty} \text{Var}[X_k] < \infty \) then

\[
\mathbb{P}\left[ \sum_{k=1}^{\infty} X_k \text{ converges } \right] = 1.
\]

Proof. Let \( S_n = \sum_{k=1}^{n} X_k \).

For every \( \omega \in \Omega \), \( \sum_{k=1}^{\infty} X_k(\omega) \) converges if and only if \( S_n(\omega) \) converges, which is if and only if \( (S_n(\omega))_n \) form a Cauchy sequence. Thus, it suffices to show that

\[
\mathbb{P}[\text{\( (S_n)_n \) is a Cauchy sequence }] = 1.
\]

Let \( n > 0 \), and consider

\[
S_{n+m} - S_n = X_{n+1} + X_{n+2} + \cdots + X_{n+m}.
\]

Kolmogorov’s inequality gives that

\[
\mathbb{P}\left[ \max_{1 \leq m \leq N} |S_{n+m} - S_n| \geq \lambda \right] \leq \frac{1}{\lambda^2} \cdot \sum_{k=1}^{N} E[X^2_{n+k}] \leq \frac{1}{\lambda^2} \cdot \sum_{k \geq 1} E[X^2_{n+k}].
\]

Taking \( N \to \infty \) on the left-hand side, using continuity of probability for the increasing sequence of events,

\[
\mathbb{P}[\sup_{m \geq 1} |S_{n+m} - S_n| \geq \lambda] \leq \frac{1}{\lambda^2} \cdot \sum_{k \geq 1} E[X^2_{n+k}].
\]
This last term is the tail of the convergent series $\sum_k \mathbb{E}[X_k^2]$. Thus, for any $r > 0$ there exists $n(r)$ so that $\sum_{k \geq n} \mathbb{E}[X_k^2] < r^{-3}$, which implies that

$$\mathbb{P}[\inf_n \sup_m |S_{n+m} - S_n| > r^{-1}] \leq \mathbb{P}[\sup_{m \geq 1} |S_{n(r)+m} - S_{n(r)}| > r^{-1}] < r^{-1}.$$  

Now, the events $\{\inf_n \sup_{m \geq 1} |S_{n+m} - S_n| > r^{-1}\}$ are increasing in $r$, so taking $r \to \infty$ with continuity of probability,

$$\mathbb{P}[\inf_n \sup_{m \geq 1} |S_{n+m} - S_n| > 0] \leq 0.$$

That is,

$$\mathbb{P}[\inf_n \sup_{m \geq 1} |S_{n+m} - S_n| = 0] = 1.$$

Now let $\omega \in \Omega$ be such that $\inf_n \sup_{m \geq 1} |S_{n+m}(\omega) - S_n(\omega)| = 0$. This implies that for all $\varepsilon > 0$ there exists $N = N(\varepsilon, \omega)$ such that

$$\sup_{m \geq 1} |S_{N+m}(\omega) - S_N(\omega)| < \varepsilon/2.$$  

Thus, for any $n > N(\varepsilon, \omega)$ and any $m \geq 1$,

$$|S_{n+m}(\omega) - S_n(\omega)| \leq |S_{N+n+m-N}(\omega) - S_N(\omega)| + |S_{N+n-N}(\omega) - S_N(\omega)| < \varepsilon.$$

That is, for any $\varepsilon > 0$ there exists $N = N(\varepsilon, \omega)$ such that for all $n > N(\varepsilon, \omega)$ and all $m \geq 1$, $|S_{n+m}(\omega) - S_n(\omega)| < \varepsilon$; that is, for this $\omega$, $(S_n(\omega))_n$ is a Cauchy sequence. We have shown that

$$\{\omega : \inf_n \sup_{m \geq 1} |S_{n+m}(\omega) - S_n(\omega)| = 0\} \subset \{\omega : (S_n(\omega))_n \text{ is a Cauchy sequence }\}.$$

Since the first event has probability 1, so does the second event, and we are done. \qed

The proof of the law of large numbers is now immediate:

**Proof of Theorem 21.1** By Kronecker’s Criterion is suffices to show that

$$\mathbb{P} \left[ \sum_{n=1}^{\infty} \frac{X_n}{n} \text{ converges} \right] = 1.$$  

By the above Lemma it suffices to show that

$$\sum_{n=1}^{\infty} \text{Var}[X_n/n] < \infty.$$
Since \( \text{Var}[X_n/n] = \frac{1}{n^2} \), this is immediate. \( \square \)

**Example 21.5.** The Central Bureau of Statistics wants to assess how many citizens use the train system. They poll \( N \) randomly and independently chosen citizens, and set \( X_j = 1 \) if the \( j \)-th citizen uses the train, and \( X_j = 0 \) otherwise. Any citizen is equally likely to be chosen, and all are independent.

As \( N \to \infty \) the sample mean

\[
\frac{1}{N} \sum_{j=1}^{N} X_j
\]

converges to \( E[X_j] = E[X_1] \) which is the probability that a randomly chosen citizen uses the train system. Since we choose each citizen uniformly, this probability is exactly the number of citizens that use the train, divided by the total number of citizens.

Thus, for large \( N \),

\[
C \cdot \frac{1}{N} \sum_{j=1}^{N} X_j
\]

is a good approximation for the number of citizens that use the train system, where \( C \) is the number of citizens. \( \triangle \nabla \triangle \)
22.1. CONVERGENCE IN DISTRIBUTION

Previously we talked about types of convergence that required the sequence and the limit to be defined on the same probability space. We now look at a type of convergence which does not have this requirement.

**Definition 22.1.** Let \((X_n)\) be a sequence of random variables. We say that \((X_n)\) **converges in distribution** to \(X\), denoted

\[ X_n \xrightarrow{\mathcal{D}} X, \]

if for all \(t \in \mathbb{R}\) such that \(F_X(t)\) is continuous at \(t\),

\[ \lim_{n \to \infty} F_{X_n}(t) = F_X(t). \]

That is, the distribution functions of \(X_n\) converge pointwise to the distribution function of \(X\).

**Remark 22.2.** Note that the convergence is required only at continuity points of \(F_X\).

**Example 22.3.** Let \(X_n \sim U[0, 1/n]\). Then \(X_n \xrightarrow{\mathcal{D}} 0\). Indeed, the distribution function of 0 is \(F_0(t) = 1\) for \(t \geq 0\) and \(F_X(t) = 0\) for \(t < 0\). Note that

\[ F_{X_n}(t) = \begin{cases} 
0 & t < 0 \\
nt & 0 \leq t \leq 1/n \\
1 & t > 1/n
\end{cases} \]
Since 0 is not a continuity point for $F_0$, we don’t care about it. For $t < 0$ we have that $F_{X_n}(t) = 0 = F_X(t)$. For $t > 0$ we have that for large enough $n > 1/t$, $F_{X_n}(t) = 1 = F_X(t)$.

\[ \triangledown \triangledown \triangledown \]

Note that for the other types of convergence we needed all random variables to live on the same space. For convergence in distribution, this is not required. However, this is the weakest kind of convergence, as can be seen by the following proposition.

**Proposition 22.4.** Convergence in probability implies convergence in distribution.

**Proof.** Let $X_n \xrightarrow{P} X$. Fix $t \in \mathbb{R}$.

First, for all $\varepsilon > 0$,
\[
P[X_n \leq t] = P[X_n \leq t, |X_n - X| \leq \varepsilon] + P[X_n \leq t, |X_n - X| > \varepsilon] \\
\leq P[X \leq t + \varepsilon] + P[|X_n - X| > \varepsilon] \rightarrow P[X \leq t + \varepsilon].
\]
So $\limsup_n F_{X_n}(t) \leq F_X(t + \varepsilon)$ for all $\varepsilon$. Since $F_X$ is right-continuous, we get that $\limsup_n F_{X_n}(t) \leq F_X(t)$.

On the other hand, for any $\varepsilon > 0$,
\[
P[X \leq t - \varepsilon] = P[X \leq t - \varepsilon, |X_n - X| \leq \varepsilon] + P[X \leq t - \varepsilon, |X_n - X| > \varepsilon] \\
\leq P[X_n \leq t] + P[|X_n - X| > \varepsilon].
\]
Taking $\liminf$ we get that $F_X(t - \varepsilon) \leq \liminf_n F_{X_n}(t)$. Now, if $F_X$ is continuous at $t$, then taking $\varepsilon \rightarrow 0$ gives
\[
F_X(t) \leq \liminf_n F_{X_n}(t) \leq \limsup_n F_{X_n}(t) \leq F_X(t).
\]
So $X_n \xrightarrow{D} X$. \qed

### 22.2. Approximating by Smooth Functions

**Lemma 22.5.** Let $(X_n)_n$ be be a sequence of random variables. Let $C^3_b$ be the space of all three times continuously differentiable functions on $\mathbb{R}$ with bounded derivatives. If for every $\phi \in C^3_b$,
\[
E[\phi(X_n)] \rightarrow E[\phi(X)]
\]
then $X_n \xrightarrow{D} X$.

Proof. We want to choose a $C^3$ function $\psi$ that will approximate the step function $1_{(-\infty,0]}$. For this we choose a $C^3$ function that is 1 on $(-\infty,0]$, decreasing on $[0,1]$ and 1 on $[1,\infty)$. Call this function $\psi$. (An explicit construction can be found in Section 22.8 below.)

For every $t \in \mathbb{R}$ and $\varepsilon > 0$ define

$$\psi_{t,\varepsilon}(x) = \psi(\varepsilon^{-1}(x - t)).$$

Note that for every $t \in \mathbb{R}$ and $\varepsilon > 0$ we have

$$\psi_{t-\varepsilon,\varepsilon}(x) \leq 1_{(-\infty,t]}(x) \leq \psi_{t,\varepsilon}(x).$$

For convergence in distribution we need to show that for every $t \in \mathbb{R}$ such that $F_X$ is continuous at $t$,

$$\lim_{n \to \infty} \mathbb{E}[1_{(-\infty,t]}(X_n)] = \mathbb{E}[1_{(-\infty,t]}(X)].$$

Fix $t \in \mathbb{R}$. For any $\varepsilon > 0$ we have that

\[
\mathbb{E}[1_{(-\infty,t]}(X_n)] - \mathbb{E}[1_{(-\infty,t]}(X)] \\
\leq \mathbb{E}[\psi_{t,\varepsilon}(X_n)] - \mathbb{E}[1_{(-\infty,t]}(X)] \to \mathbb{E}[\psi_{t,\varepsilon}(X)] - \mathbb{E}[1_{(-\infty,t]}(X)] \\
\leq \mathbb{E}[1_{(-\infty,t+\varepsilon]}(X)] - \mathbb{E}[1_{(-\infty,t]}(X)] = \mathbb{P}[X \in (t, t+\varepsilon)] = F_X(t+\varepsilon) - F_X(t).
\]

Similarly,

\[
\mathbb{E}[1_{(-\infty,t]}(X)] - \mathbb{E}[1_{(-\infty,t]}(X_n)] \\
\leq \mathbb{E}[1_{(-\infty,t]}(X)] - \mathbb{E}[\psi_{t-\varepsilon,\varepsilon}(X_n)] \to \mathbb{E}[1_{(-\infty,t]}(X)] - \mathbb{E}[\psi_{t-\varepsilon,\varepsilon}(X)] \\
\leq \mathbb{E}[1_{(-\infty,t]}(X)] - \mathbb{E}[1_{(-\infty,t-\varepsilon]}(X)] = \mathbb{P}[X \in (t-\varepsilon,t)] = F_X(t) - F_X(t-\varepsilon).
\]

We conclude that

$$\limsup_{n \to \infty} \mathbb{E}[1_{(-\infty,t]}(X_n)] - \mathbb{E}[1_{(-\infty,t]}(X)] \leq |F_X(t+\varepsilon) - F_X(t)| + |F_X(t) - F_X(t-\varepsilon)|$$

for all $\varepsilon > 0$.

For any $t$ that is a continuity point of $F_X$ the above tends to 0 as $\varepsilon \to 0$. $\square$
22.3. Central Limit Theorem

We will now build the tools for the following very important approximation theorem:

**Theorem 22.6 (Central Limit Theorem).** Let $X_1, X_2, \ldots, X_n, \ldots,$ be mutually independent random variables, such that $\mathbb{E}[X_n] = 0$ and $\mathbb{E}[X_n^2] = 1$. Let

$$S_N = \sum_{n=1}^{N} X_n.$$

Then,

$$\frac{S_N}{\sqrt{N}} \xrightarrow{d} N(0,1).$$

That is, for all $t \in \mathbb{R},$

$$\mathbb{P}[S_N \leq t\sqrt{N}] \to \int_{-\infty}^{t} \frac{1}{\sqrt{2\pi}} e^{-s^2/2} ds.$$

In the exercises we will show that

**Exercise 22.7.** Let $X_1, X_2, \ldots, X_n, \ldots,$ be mutually independent random variables, such that $\mathbb{E}[X_n] = \mu$ and $\text{Var}[X_n] = \sigma^2$. Let

$$S_N = \sum_{n=1}^{N} X_n.$$

Then,

$$\frac{S_N - N\mu}{\sqrt{N}\sigma} \xrightarrow{d} N(0,1).$$

✓ That is, no matter what distribution we start with - we can approximate the sum of many independent trials by a normal random variable.

22.4. Uses for the CLT

**Example 22.8.** Every student’s grade in a course has expectation 80 and standard deviation 20. Give an approximation of the average grade for 100 students, if all students are independent. What is a good estimate for the probability of the average grade to be below 70?

If $X_k$ is the grade of the $k$-th student, then the average grade is $M := \frac{1}{100} \sum_{k=1}^{100} X_k$. 
The central limit theorem tells us that 
\[ \frac{1}{10} \cdot 100 \cdot (100 - 100 \cdot 80) = \frac{1}{2} (M - 80) \] 
is very close to a \( N(0, 1) \) random variable. Thus,
\[
P[M < 70] = P\left[ \frac{1}{2} (M - 80) < -5 \right] \approx P[N(0, 1) < -5] = \int_{-\infty}^{-5} \frac{1}{\sqrt{2\pi}} e^{-s^2/2} ds = ... 
\]
which can be calculated using a standard normal distribution table.

**Example 22.9.** \( X_1, X_2, \ldots \), are independent random variables with the same distribution, such that \( E[X_n] = 60 \) and \( \text{Var}[X_n] = 25 \).

Show that we can take \( N \) large enough so that the empirical average \( \frac{1}{N} \sum_{k=1}^{N} X_k \) is between 55 and 65 with probability greater than 0.99.

How large should we take \( N \) if we want to ensure this?

Let \( S_N = \sum_{k=1}^{N} X_k \). The central limit theorem tells us that the distribution of \( \frac{1}{5\sqrt{N}} (S_N - 60N) \) converges to the distribution of a \( N(0, 1) \) random variable. Thus,
\[
P[55 < \frac{1}{N} S_N < 65] = P[-\sqrt{N} < \frac{S_N - 60N}{5\sqrt{N}} < \sqrt{N}] \approx P[-\sqrt{N} < N(0, 1) < \sqrt{N}]. 
\]
The right hand side above
\[
P[-\sqrt{N} < \frac{S_N - 60N}{5\sqrt{N}} < \sqrt{N}] 
\]
goes to 1. So there is some large enough \( N \) so that this is greater than 0.99.

**Example 22.10.** \( X_1, X_2, \ldots \), are independent random variables with the same distribution, such that \( E[X_n] = \mu \) and \( \text{Var}[X_n] = 1 \).

Show that there exists \( N_0 \) and \( c > 0 \) such that for all \( N > N_0 \),
\[
P[|S_N - N\mu| \leq \frac{1}{2} \sqrt{N}] \geq c, 
\]
where \( S_N = \sum_{k=1}^{N} X_k \).

Note that we cannot use Kolmogorov here, because that would only give
\[
P[\max_{1 \leq n \leq N} |S_n - n\mu| > \frac{1}{2} \sqrt{N}] \leq 4. 
\]
However, the central limit theorem tells us that
\[
P[|S_N - N\mu| \leq \frac{1}{2} \sqrt{N}] \to P[|N(0, 1)| \leq \frac{1}{2}]. 
\]
If we take $c = \frac{1}{2} \mathbb{P}[|N(0,1)| \leq \frac{1}{2}]$, we get that for all large enough $N$,

$$\mathbb{P}[|S_N - N\mu| \leq \frac{1}{2}\sqrt{N}] > c.$$ 

Example 22.11. The government decides to test if the lab workers at the central lab of Kupat Cholim are adequate. They are given an exam, and graded between 0 and 100. The exam is such so that the expectation of a random worker’s grade is 80 and standard deviation is 20.

If the lab has 10000 workers, give a good approximation of the probability that the average grade in the lab is smaller than 70?

In this case, we can let $X_1, \ldots, X_{10000}$ be the grades of the workers, so $\mathbb{E}[X_j] = 80$ and $\text{Var}[X_j] = 400$. If $A = \frac{1}{10000} \sum_{j=1}^{10000} X_j$ is the average grade, the central limit theorem tells us that

$$\mathbb{P}[A \leq 70] = \mathbb{P}[100(A - 80) \leq -10 \cdot 100] = \mathbb{P}\left[\frac{1}{100 \cdot 20} \sum_{j=1}^{10000} (X_j - 80) \leq -50\right]$$

is very close to

$$\mathbb{P}[N(0,1) \leq -50] = \int_{-\infty}^{-50} \frac{1}{\sqrt{2\pi}} e^{-s^2/2} ds.$$ 

Example 22.12. Suppose there are 1.2 million people that will actually vote in the elections. We want to forecast how many mandates party A will receive in the elections. Each mandate is worth $1 \cdot 10^6 / 120 = 10^4$ people. There are $a$ people that will actually vote for party A, so party A will get $10^{-4}a$ mandates.

Suppose we choose a random voter uniformly at random. The probability that she votes for party A is then $p = \frac{a}{1.2 \cdot 10^6}$, and the variance is $p(1-p)$.

If we repeat this experiment 900 times, we can sum up the number of people who said they would vote for party A, and the average of this number should be close to $p$. So if $X_j$ is the indicator of the event that the $j$-th voter polled said she would vote for party A, we get that

$$120 \cdot \frac{1}{900} \sum_{j=1}^{900} X_j$$
is close to $120p = 1.2 \cdot 10^6 \cdot 10^{-4}p = 10^{-4}a$, which is the number of mandates party A gets.

How good is this approximation? What is the number of mandates so that this is close to the real number with 95% confidence?

For any $k$, we have that

$$\Pr[\left| 120 \cdot \frac{1}{900} \sum_{j=1}^{900} X_j - 10^{-4}a \right| > k] = \Pr[\left| \frac{1}{900} \sum_{j=1}^{900} (X_j - p) \right| > \frac{k}{120}]$$

$$= \Pr[\left| \frac{1}{36} \sum_{j=1}^{900} (X_j - p) \right| > \frac{k}{4}]$$

which is close to

$$\Pr[|N(0, 1)| > \frac{k}{4\sqrt{p(1-p)}}]$$

by the central limit theorem. If we take $k \geq 4\sqrt{p(1-p)}$ then this is at most

$$2 \int_{t}^{\infty} \frac{1}{\sqrt{2\pi}} \cdot e^{-s^2/2} ds \leq 2 \int_{t}^{\infty} \frac{s}{t \sqrt{2\pi}} \cdot e^{-s^2/2} ds = \frac{\sqrt{2}}{t \sqrt{\pi}} e^{-t^2/2}.$$

For $t = 2.05$ this is smaller than 0.05. Since $p(1-p) \leq 1/4$ we can take $k = 5 > 2 \cdot 2.05 \geq 4\sqrt{p(1-p)} \cdot 2.05$ and then

$$\Pr[|120 \cdot \frac{1}{900} \sum_{j=1}^{900} X_j - 10^{-4}a| > k] < 0.05.$$

$\triangle \triangledown \triangle$

**Example 22.13.** Let $(X_n)_n$ be independent identically distributed random variables, let $(Y_n)_n$ be independent identically distributed random variables, such that $\mathbb{E}[X_n] = \mathbb{E}[Y_n]$ for all $n$, and for all $n$:

$$\text{Var}[X_n] = \text{Var}[Y_n] = 1 \quad \text{and} \quad \text{Cov}(X_n, Y_n) = \frac{1}{2}.$$

Let $S = \sum_{j=1}^{10^4} X_j$ and $T = \sum_{j=1}^{10^4} Y_j$. Give a good approximation for $\Pr[S > T]$.

We have that

$$S - T = \sum_{j=1}^{10^4} (X_j - Y_j),$$
where \((X_n - Y_n)_n\) are independent with mean 0 and

\[ \text{Var}[X_j - Y_j] = \text{Var}[X_j] + \text{Var}[Y_j] - 2 \text{Cov}(X_j, Y_j) = 1. \]

So

\[ \mathbb{P}[S > T] = \mathbb{P}[S - T > 0] = \mathbb{P}\left[\frac{1}{100}(S - T) > 0\right] \approx \mathbb{P}[N(0, 1) > 0] = \frac{1}{2}. \]

Example 22.14. A gambler plays a game repeatedly where at each game he earns Geo(0.1) Shekel and loses Poi(10) Shekel, both wins and losses independent.

- What is a good approximation for the probability that he has more than 20 Shekel after 100 games?
- What is a good approximation for the probability that he has exactly 0 Shekel?

The amount of money after 100 games is

\[ M := \sum_{j=1}^{100} X_j - Y_j \]

where \(X_j \sim \text{Geo}(0.1)\) and \(Y_j \sim \text{Poi}(10)\) all independent. Since \(E[X_j - Y_j] = 0\) and since

\[ \text{Var}[X_j - Y_j] = \text{Var}[X_j] + \text{Var}[Y_j] = 100 \cdot 0.9 + 10 = 100, \]

we have using the central limit theorem

\[ \mathbb{P}[M > 20] = \mathbb{P}\left[\frac{1}{100}M > 0.2\right] \approx \mathbb{P}[N(0, 1) > 0.2] \]

For \(\mathbb{P}[M = 0]\) we cannot use the approximation \(\mathbb{P}[N = 0] = 0!\) So we use the following trick: Since \(M\) takes on only integer values, \(M = 0\) if and only if \(M \in [-1/2, 1/2]\). So,

\[ \mathbb{P}[M = 0] = \mathbb{P}\left[\frac{1}{100}M \in [-\frac{1}{200}, \frac{1}{200}]\right] \approx \mathbb{P}\left[-\frac{1}{200} \leq N(0, 1) \leq \frac{1}{200}\right] \]
22.5. Lindeberg’s Method

A “hands-on” proof of the CLT (that does not use the Levy Continuity Theorem) was given by Lindeberg. The main idea is: If $Z_1, \ldots, Z_n$ are independent $N(0, 1)$ random variables, then $n^{-1/2}(Z_1 + \cdots + Z_n) \sim N(0, 1)$. So we can change $X_1, X_2, \ldots, X_n$ into $Z_1, \ldots, Z_n$ one by one, each time comparing the sums. Hopefully each of the $n$ differences will not be large.

First, a technical lemma regarding $C^3$ functions.

Lemma 22.15. For any $C^3$ function $\phi$ and any $x, y \in \mathbb{R}$,

$$|\phi(x + y) - \phi(x) - \phi'(x)y - \frac{1}{2}\phi''(x)y^2| \leq \min \left\{ ||\phi''||_\infty |y|^2, ||\phi'''||_\infty |y|^3 \right\}.$$

Proof. For any $y > 0$, by Taylor’s theorem, expanding around $x$,

$$|\phi(x + y) - \phi(x) - \phi'(x)y - \frac{1}{2}\phi''(x)y^2| \leq \frac{1}{6}||\phi'''||_\infty \cdot y^3.$$

The second order Taylor expansion

$$|\phi(x + y) - \phi(x) - \phi'(x)y| \leq \frac{1}{2}||\phi''||_\infty \cdot y^2,$$

with the triangle inequality gives that

$$|\phi(x + y) - \phi(x) - \phi'(x)y - \frac{1}{2}\phi''(x)y^2| \leq |\phi(x + y) - \phi(x) - \phi'(x)y| + \frac{1}{2}||\phi''||_\infty \cdot y^2 \leq ||\phi''||_\infty \cdot y^2.$$

If $y < 0$ apply the previous argument to the function $\phi(-x)$, and note that the absolute values of the derivatives of the two functions are always the same.

Lemma 22.16. Let $A, X, Z$ be independent random variables with $\mathbb{E}[A] = \mathbb{E}[X] = \mathbb{E}[Z] = 0$ and $\mathbb{E}[X^2] = \mathbb{E}[Z^2]$. Then for any $C^3$ function $\phi$ and any $\epsilon, \delta > 0$,

$$\left| \mathbb{E}[\phi(A + \delta X)] - \mathbb{E}[\phi(A + \delta Z)] \right| \leq M_\phi \delta^2 \cdot \left( \epsilon \mathbb{E}[X^2] + \epsilon \mathbb{E}[Z^2] + \mathbb{E}[Z^2 \mathbb{1}_{\{|Z| > \epsilon \delta - 1\}}] + \mathbb{E}[X^2 \mathbb{1}_{\{|X| > \epsilon \delta - 1\}}] \right),$$

where $M_\phi = \max \{ ||\phi''||_\infty, ||\phi'''||_\infty \}$. 

We gain $\epsilon$ in the first term, and the $\epsilon^2 \mathbb{1}_{\{|X| > \epsilon \delta - 1\}}$ term for the second term.
Proof. Set

\[
Y = \left( \phi(A + \delta X) - \phi(A) - \phi'(A) \cdot \delta X - \frac{1}{2} \phi''(A) \cdot \delta^2 X^2 \right)
- \left( \phi(A + \delta Z) - \phi(A) - \phi'(A) \cdot \delta Z - \frac{1}{2} \phi''(A) \cdot \delta^2 Z^2 \right).
\]

Since \(\mathbb{E}[X] = \mathbb{E}[Z]\) and \(\mathbb{E}[X^2] = \mathbb{E}[Z^2]\) we get that

\[
\mathbb{E}[Y] = \mathbb{E}[\phi(A + \delta X)] - \mathbb{E}[\phi(A + \delta Z)].
\]

Also, for \(Y_n = \phi(A + \delta X) - \phi(A) - \phi'(A) \cdot \delta X - \frac{1}{2} \phi''(A) \cdot \delta^2 X^2\),

\[
|\mathbb{E}[Y_n]| = |\mathbb{E}[Y_n 1_{\{|X| \geq \epsilon \delta^{-1}\}}] + \mathbb{E}[Y_n 1_{\{|X| \leq \epsilon \delta^{-1}\}}]|
\leq ||\phi''||_{\infty}^2 \mathbb{E}[X^2 1_{\{|X| \geq \epsilon \delta^{-1}\}}] + ||\phi'''||_{\infty} \mathbb{E}[|X|^3 1_{\{|X| \leq \epsilon \delta^{-1}\}}]
\leq ||\phi''||_{\infty} \mathbb{E}[X^2 1_{\{|X| \geq \epsilon \delta^{-1}\}}] + \epsilon ||\phi'''||_{\infty} \mathbb{E}[X^2].
\]

We are now ready to prove the Central Limit Theorem.

**Theorem 22.17** (Central Limit Theorem). Let \((X_n)_n\) be independent identically distributed random variables with \(\mathbb{E}[X_k] = 0\) and \(\mathbb{E}[X_k^2] = 1\). Let \(S_n = \sum_{k=1}^n X_k\) and \(Z \sim N(0,1)\). Then, for any \(C^3_o\) function \(\phi\),

\[
\mathbb{E}[\phi(n^{-1/2}S_n)] \to \mathbb{E}[\phi(Z)],
\]

so consequently, \(n^{-1/2}S_n \xrightarrow{D} Z\).

Proof. Let \((Z_n)_n\) be independent \(N(0,1)\) random variables, also independent of \((X_n)_n\).

Note that \(n^{-1/2} \sum_{k=1}^n Z_k \sim N(0,1)\).

Fix \(n\) and let \(0 \leq k \leq n\). Define

\[
A_k = X_1 + \cdots + X_k + Z_{k+1} + \cdots Z_n.
\]

So \(S_n = A_n\) and \(n^{-1/2} A_0 \sim N(0,1)\). We can write the telescopic sum

\[
\mathbb{E}[\phi(Z)] - \mathbb{E}[\phi(n^{-1/2}S_n)] = \sum_{k=0}^{n-1} \mathbb{E}[\phi(n^{-1/2}A_k)] - \mathbb{E}[\phi(n^{-1/2}A_{k+1})].
\]

So it suffices to bound the increments \(|\mathbb{E}[\phi(n^{-1/2}A_k)] - \mathbb{E}[\phi(n^{-1/2}A_{k+1})]|\).
For $0 \leq k \leq n - 1$ let $A = n^{-1/2}(X_1 + \cdots + X_k + Z_{k+2} + \cdots + Z_n)$. So $n^{-1/2}A_k = A + n^{-1/2}Z_{k+1}$ and $n^{-1/2}A_{k+1} = A + n^{-1/2}X_{k+1}$. Thus, for any $\varepsilon > 0$,

$$
\left| E[\phi(n^{-1/2}A_k)] - E[\phi(n^{-1/2}A_{k+1})] \right| \leq M_{\phi} \cdot n^{-1} \cdot \left( 2\varepsilon + E[Z_{k+1}^21_{\{|Z_{k+1}|>\varepsilon\sqrt{n}\}}] + E[X_{k+1}^21_{\{|X_{k+1}|>\varepsilon\sqrt{n}\}}] \right).
$$

Summing over $k$ we get for any $\varepsilon > 0$,

$$
\left| E[\phi(Z)] - E[\phi(n^{-1/2}S_n)] \right| \leq M_{\phi} \cdot \left( 2\varepsilon + E[Z_2^21_{\{|Z|>\varepsilon\sqrt{n}\}}] + E[X_2^21_{\{|X|>\varepsilon\sqrt{n}\}}] \right).
$$

Taking $n \to \infty$, and recalling that $\varepsilon > 0$ was arbitrary, it suffices to show that for any random variable $X$ with finite variance, the quantity $E[X^21_{\{|X|>\varepsilon\sqrt{n}\}}] \to 0$.

Indeed, the sequence $(X^21_{\{|X|\leq\varepsilon\sqrt{n}\}})_n$ is monotonically increasing to $X^2$. So monotone convergence implies that

$$
E[X^21_{\{|X|>\varepsilon\sqrt{n}\}}] = E[X^2] - E[X^21_{\{|X|\leq\varepsilon\sqrt{n}\}}] \to 0.
$$

\[\Box\]

22.6. Characteristic Function

The characteristic function is actually just the Fourier transform of a random variable.

22.6.1. Preliminaries. Let $(X,Y)$ be a two-dimensional random variable. We can think of $(X,Y)$ as a complex valued random variable and write $(X,Y) = X + iY$. Define the expectation of a complex valued random variable to be $E[X + iY] = E[X] + iE[Y]$ as long as these expectations are all finite.

Note that if $g : \mathbb{R} \to \mathbb{C}$ is a measurable function, then we can write $g = \text{Reg} + \text{Img}$, and these are also measurable. Moreover, if $X, Y$ are independent, and $g, h : \mathbb{R} \to \mathbb{C}$ are measurable, then $g(X), h(Y)$ are independent, since $\text{Reg}(X), \text{Img}(X)$ are independent of $\text{Re}(Y), \text{Im}(Y)$.

**Definition 22.18.** Let $X$ be a random variable. Define a function $\varphi_X : \mathbb{R} \to \mathbb{C}$ by

$$
\varphi_X(t) := E[e^{itX}] = E[\cos(tX)] + iE[\sin(tX)].
$$

$\varphi_X$ is called the **characteristic function** of $X$. 

Note that \( \phi_X \) is always defined since \( \mathbb{E}[|\cos(tX)|] \leq 1 \) and \( \mathbb{E}[|\sin(tX)|] \leq 1 \). Moreover, we get that (as a complex number) \( |\phi_X(t)| \leq 1 \).

Note that if \( X \) is discrete with density \( f_X \) then
\[
\phi_X(t) = \sum_r f_X(r) e^{itr}.
\]

If \( X \) is absolutely continuous with density \( f_X \) then
\[
\phi_X(t) = \int_{-\infty}^{\infty} e^{its} f_X(s) ds.
\]
(This latter object is the Fourier transform of \( f_X \).)

**Example 22.19.** Let’s calculate some characteristic functions:

- \( X \sim \text{Ber}(p) \):
  \[
  \phi_X(t) = pe^{it} + (1-p) = 1-p(e^{it}).
  \]

- \( Z = X + Y \) for \( X, Y \) independent:
  \[
  \phi_Z(t) = \mathbb{E}[e^{itX}e^{itY}] = \mathbb{E}[e^{itX}]\mathbb{E}[e^{itY}] = \phi_X(t) \cdot \phi_Y(t).
  \]

- \( Y = cX \) for some real \( c \in \mathbb{R} \):
  \[
  \phi_Y(t) = \mathbb{E}[e^{itcX}] = \phi_X(tc).
  \]

- \( X \sim \text{Bin}(n,p) \): We can write \( X = \sum_{k=1}^{n} X_j \) where \( (X_k)_{k=1}^{n} \) are independent Ber\( (p) \). Thus,
  \[
  \phi_X(t) = \prod_{k=1}^{n} \phi_{X_k}(t) = (1-p(1-e^{it}))^n.
  \]

- \( X \sim \text{Geo}(\lambda) \):
  \[
  \phi_X(t) = \sum_{k=1}^{\infty} (1-p)^{k-1}pe^{itk} = \frac{pe^{it}}{1-(1-p)e^{it}}.
  \]

- \( X \sim \text{Poi}(\lambda) \):
  \[
  \phi_X(t) = \sum_{k=0}^{\infty} \frac{e^{-\lambda} \lambda^k}{k!} e^{itk} = e^{-\lambda(1-e^{it})}.
  \]
• $X \sim U[a, b]$:

$$
\phi_X(t) = \int_{-\infty}^{\infty} e^{its} f_X(s) ds = \frac{1}{b-a} \int_a^b e^{its} ds = \frac{1}{it(b-a)} (e^{itb} - e^{ita}).
$$

• $X \sim \text{Exp}(\lambda)$:

$$
\phi_X(t) = \int_0^{\infty} \lambda e^{-\lambda s} e^{its} ds = \lambda \frac{\lambda}{\lambda - it}.
$$

• $X \sim N(\mu, \sigma)$: Write $Y = (X - \mu)/\sigma$. So $Y \sim N(0, 1)$. We use the fact that $s \mapsto \sin(ts)e^{-s^2/2}$ is an odd function, so its integral over $\mathbb{R}$ is 0. Since $e^{its} = \cos(ts) + i \sin(ts)$ we get that

$$
\phi_Y(t) = \int_{-\infty}^{\infty} 1 \sqrt{2\pi} e^{-s^2/2} e^{its} ds = \int_{-\infty}^{\infty} \sqrt{2\pi} e^{-s^2/2} \cos(ts) ds.
$$

Let $\chi_s(t) = e^{-s^2/2} \cos(ts)$. Since $|\chi_s(t)| \leq 1$, since $\chi_s'(t) = -e^{-s^2/2} \sin(ts)$ which is continuous in $t$ for every $s$, since for any $t$

$$
\int_\mathbb{R} \int_{-\epsilon}^\epsilon |\chi_s'(t + \eta)| d\eta ds \leq \int_\mathbb{R} |s| e^{-s^2/2} ds < \infty,
$$

and since

$$
\int_\mathbb{R} \chi_s'(t) ds = -\int_{-\infty}^{\infty} \sin(ts) se^{-s^2/2} ds = \sin(ts)e^{-s^2/2} \bigg|_{-\infty}^{\infty} - \int_{-\infty}^{\infty} t \cos(ts) e^{-s^2/2} ds = -t \int_\mathbb{R} \chi_s(t) ds,
$$

which is continuous in $t$, we can differentiate $\phi_Y$ under the integral to get

$$
\phi'_Y(t) = \frac{1}{\sqrt{2\pi}} \int_\mathbb{R} \chi_s'(t) = -t \cdot \frac{1}{\sqrt{2\pi}} \int_\mathbb{R} \chi_s(t) = -t \phi_Y(t).
$$

Dividing by $\phi_Y$ we get that

$$
\frac{d}{dt} (\log \phi_Y(t)) = \frac{\phi'_Y(t)}{\phi_Y(t)} = -t.
$$

Integrating, $\log \phi_Y(t) = -\frac{t^2}{2} + C$ and since $\phi_Y(0) = 1$ we get that $C = 0$ and $\phi_Y(t) = e^{-t^2/2}$. This completes the calculation for $Y \sim N(0, 1)$. For $X = \sigma Y + \mu$ we get

$$
\phi_X(t) = \mathbb{E}[e^{it\sigma Y}]e^{it\mu} = e^{it\mu} \phi_Y(t) = e^{it\mu} e^{-\sigma^2 t^2/2}.
$$

$\triangle \nabla \triangle$
Lemma 22.20. Let $X$ be a random variable with $\mathbb{E}[X] = 0$ and $\mathbb{E}[X^2] = 1$. Then, as $t \to 0$,

$$\phi_X(t) = 1 - \frac{t^2}{2} + o(t^2).$$

Proof. As in Lemma 22.15 we get that for any $x \in \mathbb{R}$ and $t > 0$,

$$\left| e^{-itx} - 1 - itx + \frac{t^2}{2}x^2 \right| \leq \min \left\{ t^2x^2, t^3|x|^3 \right\}.$$

Thus, for any $\varepsilon > 0$,

$$\left| \mathbb{E}[e^{-itX}] - 1 - \frac{t^2}{2} \right| \leq \mathbb{E}\left[ \min \left\{ t^2X^2, t^3|X|^3 \right\} \right] \leq \frac{t^2}{2} \mathbb{E}[X^21_{\{|X|\geq t\varepsilon\}}] + \varepsilon t^2.$$

We have already seen that for $0 < \eta < 1$, as $t \to 0$, $\mathbb{E}[X^21_{\{|X|\geq t\eta^{-1}\}}] \to 0$ by monotone convergence. So choosing $\varepsilon = t^\eta$, as $t \to 0$,

$$\phi_X(t) = 1 - \frac{t^2}{2} + o(t^2).$$

22.7. Lévy’s Continuity Theorem

Perhaps the main theorem using characteristic functions is a theorem of Paul Lévy giving a necessary and sufficient condition for convergence in distribution.

Theorem 22.21 (Lévy’s Continuity Theorem). Let $(X_n)_n$ be a sequence of random variables, and let $X$ be another random variable. Then, $X_n \xrightarrow{D} X$ if and only if $\phi_{X_n}(t) \to \phi_X(t)$ for all $t \in \mathbb{R}$.

22.7.1. Proof of CLT. We can use Lévy’s Continuity Theorem to give a different proof of the Central Limit Theorem.

Proof of Theorem 22.6. Let $\phi(t) = \phi_{X_n}(t)$. Note that

$$\phi_{S_n/\sqrt{N}}(t) = \phi(t/\sqrt{N})^N.$$

As $N \to \infty$, we have that

$$\phi(t/\sqrt{N}) = 1 - \frac{t^2}{2N} + o(t^2N^{-1}).$$
as $N \to \infty$. Thus,

$$\lim_{N \to \infty} \phi(t/\sqrt{N})^N = e^{-t^2/2}. $$

So $\phi_{S_N/\sqrt{N}}(t) \to e^{-t^2/2}$ which is the characteristic function of a $N(0, 1)$ random variable. By Lévy’s Continuity Theorem we get that $S_N/\sqrt{N} \xrightarrow{D} N(0, 1)$. \hfill \square

### 22.8. An Explicit Approximation of the Step Function

In this section we construct a function $\psi$, which is $C^3_b$, takes values 1 on $(-\infty, 0]$, decreasing on $[0, 1]$ and values 0 on $[1, \infty)$. Some consideration leads us to understand that we would like $\psi' < 0$ on $(0, 1)$, and so we search for a function such that $\psi''(x) = -\psi''(1-x)$ on $[0, 1]$, vanishes at the endpoints 0, 1 and obtains local minimum or maximum at the endpoint 0, 1. Some thought leads to the function $g(x) = x^2(1-x)^2(1-2x)$. Integrating and taking a derivative we have that for

$$h(x) = \frac{x^4}{12} - \frac{x^5}{5} + \frac{x^6}{6} - \frac{x^7}{21},$$

the derivatives on $(0, 1)$ satisfy $h'(x) = \frac{1}{3}x^3(1-x)^3$, $h''(x) = g(x)$ and $h'''(x) = g'(x) = 2x - 12x^2 + 20x^3 - 10x^4$.

Set $A = 420 = 7 \cdot 5 \cdot 3 \cdot 4$. Because $h(1) = A^{-1}$ and $h(0) = 0$ we take

$$\psi(x) = \begin{cases} 
1 - A \cdot h(x) & x \in [0, 1] \\
1 & x < 0 \\
0 & x > 1 
\end{cases}$$

We have that on $(0, 1), \psi' = -Ah' < 0$, also $\psi'' = -Ah''$ and $\psi''' = -Ah'''$. These three derivatives all vanish at 0 and 1 so $\psi$ is $C^3_b$. 


Figure 6. The second derivative $-20 \cdot g(x)$ and the function $1 - A \cdot h(x)$. 
23.1. Concentration of Measure

Suppose \((X_n)_n\) are a sequence of i.i.d. random variables with \(P[X_n = 1] = P[X_n = -1] = 1/2\). Let

\[ S_n = \sum_{k=1}^{n} X_k. \]

So \(E[S_n] = 0\) and \(\text{Var}[S_n] = E[S_n^2] = n\).

Chebyshev’s inequality tells us that

\[ P[|S_n| \geq \lambda \sqrt{n}] \leq \lambda^{-2}. \]

However, the central limit theorem gives that for large \(n\),

\[
\begin{align*}
P[|S_n| \geq \lambda \sqrt{n}] &\approx P[|N(0, 1)| \geq \lambda] = \int_{\lambda}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-s^2/2} ds + \int_{-\infty}^{-\lambda} \frac{1}{\sqrt{2\pi}} e^{-s^2/2} ds \\
&\leq \sqrt{2/\pi} \cdot \lambda^{-1} \int_{\lambda}^{\infty} s e^{-s^2/2} ds = \sqrt{2/\pi} \lambda^{-1} \cdot (-e^{-s^2/2}) \bigg|_{\lambda}^{\infty} \\
&= \sqrt{2/\pi} \lambda^{-1} e^{-\lambda^2/2}.
\end{align*}
\]

This decays much faster than \(\lambda^{-2}\). However, we don’t know how good our approximation by a standard normal is.

Because \((X_n)_n\) are independent, we can obtain an very nice concentration result using a smart trick by Bernstein.

**Theorem 23.1** (Bernstein’s Inequality). *Let \((X_n)_n\) be independent random variables such that for all \(n\), \(|X_n| \leq 1\) a.s. and \(E[X_n] = 0\). Let \(S_n = \sum_{k=1}^{n} X_k\). Then for any \(\lambda > 0\),

\[ P[S_n \geq \lambda] \leq \exp \left( -\frac{\lambda^2}{2n} \right), \]
and consequently,
\[ \mathbb{P}(|S_n| \geq \lambda) \leq 2 \exp \left( -\frac{\lambda^2}{2n} \right). \]

**Proof.** There are two main ideas:

The first idea, is that for a random variable \( X \) with \( \mathbb{E}[X] = 0 \) and \( |X| \leq 1 \) a.s. one has \( \mathbb{E}[e^{\alpha X}] \leq e^{\alpha^2/2} \). Indeed, \( g(x) = e^{\alpha x} \) is a convex function, so for \( |x| \leq 1 \) we can write \( x = \beta \cdot 1 + (1 - \beta) \cdot (-1) \), where \( \beta = \frac{x+1}{2} \), and
\[
e^{\alpha x} \leq \beta e^{\alpha} + (1 - \beta) e^{-\alpha} = \cosh(\alpha) + x \sinh(\alpha).
\]
(Here \( 2 \cosh(\alpha) = e^\alpha + e^{-\alpha} \) and \( 2 \sinh(\alpha) = e^\alpha - e^{-\alpha} \).) Thus, because \( \mathbb{E}[X] = 0 \), and using \( (2k)! \geq 2^k k! \),
\[
\mathbb{E}[e^{\alpha X}] \leq \cosh(\alpha) + \mathbb{E}[X] \sinh(\alpha) = \cosh(\alpha)
\leq \sum_{k=0}^{\infty} \frac{\alpha^{2k}}{(2k)!} \leq \sum_{k=0}^{\infty} \frac{\alpha^{2k}}{2^k k!} = e^{\alpha^2/2}.
\]

For the second idea, due to Sergei Bernstein, one applies the Chebyshev / Markov inequality to the non-negative random variable \( e^{\alpha X} \), and then optimizes on \( \alpha \).

In our case, using independence,
\[
\mathbb{E}[e^{\alpha S_n}] = \prod_{k=1}^{n} \mathbb{E}[e^{\alpha X_k}] \leq e^{\alpha^2 n/2}.
\]

Now apply Markov’s inequality to the non-negative random variable \( e^{\alpha S_n} \) to get for any \( \alpha > 0 \),
\[
\mathbb{P}[S_n \geq \lambda] = \mathbb{P}[e^{\alpha S_n} \geq e^{\alpha \lambda}] \leq \exp \left( \frac{1}{2} n \alpha^2 - \alpha \lambda \right).
\]
Optimizing over \( \alpha \) we get that for \( \alpha = \lambda/n \),
\[
\mathbb{P}[S_n \geq \lambda] \leq \exp \left( -\frac{\lambda^2}{2n} \right).
\]
\( \square \)

**Example 23.2.** Suppose a gambler plays a game repeatedly and independently, such that with probability \( p < 1/2 \) he earns one dollar and loses one dollar with probability \( 1 - p \).
For every $n$ let $X_n$ be the amount he won in the $n$-th game. So $S_n = \sum_{k=1}^{n} X_k$ is the total amount he wins after $n$ games.

Note that $E[X_k] = 2p - 1 < 0$, so $Y_k = \frac{1}{2}(X_k - (2p - 1))$ has expectation 0 and $|Y_k| \leq 1$. By Bernstein’s inequality

$$
P[S_n > 0] = P\left[\sum_{k=1}^{n} Y_k > \frac{n}{2}(1 - 2p)\right] \leq \exp\left(-\frac{n(1 - 2p)^2}{8}\right).$$

So the probability of winning is very small even if the house has a very small advantage.

$\triangle \nabla \triangle$
Exercise 24.1. We have $N$ urns. Urn number $k$ contains $k$ white balls and $N-k$ black balls.

An urn is chosen randomly, all urns equally likely. Then, we start removing balls from that urn, all balls equally likely, all balls independent, returning the removed ball to the urn after removing it. Let $A$ be the event that the first $n$ balls removed are white. Let $B$ be the event that the $n+1$-th ball removed is black.

Calculate $P[A], P[B \cap A], P[B|A]$.

Now calculate the same if the process is that each time a random urn is chosen independently, and then a ball is removed and returned to and from that urn.

Solution to Exercise 24.1. For the first scenario, let $C_k$ be the event the $k$-th urn was chosen. Independence gives us that

$$P[A|C_k] = \left(\frac{k}{N}\right)^n$$
$$P[B \cap A|C_k] = \left(\frac{k}{N}\right)^n \cdot \frac{N-k}{N}.$$

Thus by the law of total probability,

$$P[A] = \frac{1}{N^{n+1}} \sum_{k=1}^{N} k^n$$
$$P[B \cap A] = \frac{1}{N^{n+2}} \sum_{k=1}^{N} k^n (N-k).$$

Thus,

$$P[B|A] = 1 - \frac{1}{N} \cdot \frac{\sum_{k=1}^{N} k^{n+1}}{\sum_{k=1}^{N} k^n}.$$  

In the second scenario, the probability that a white ball is chosen each time is

$$\frac{1}{N} \sum_{k=1}^{N} \frac{k}{N} = \frac{1}{N^2} \cdot \binom{N+1}{2} = \frac{N+1}{2N} := p.$$
All balls are independent, so
\[ P[A] = p^n \quad P[B \cap A] = p^n(1 - p) \quad P[B|A] = (1 - p). \]

\[ \square \]

**Exercise 24.2.** Random variables \( X, Y \) are equally distributed if \( F_X = F_Y \). Show that if \( X, Y \) are equally distributed then \( \mathbb{E}[X] = \mathbb{E}[Y] \) if the expectations exist, and \( \mathbb{E}[X] \) does not exist if and only if \( \mathbb{E}[Y] \) does not exist. (Hint: Start with simple, go through non-negative, continue to general. First show that \( P_X = P_Y \).

**Solution to Exercise 24.2.** The first observation is that \( P[X \in B] = P[Y \in B] \) for any Borel set \( B \in \mathcal{B} \). Indeed, the probability measures \( P_X, P_Y \) agree on the \( \pi \)-system \( \{(-\infty, t] : t \in \mathbb{R}\} \) by definition, and this \( \pi \)-system generates \( \mathcal{B} \), so \( P_X, P_Y \) agree on all \( \mathcal{B} \).

Now, note that if \( g : \mathbb{R} \to \mathbb{R} \) is a measurable function, then if \( X, Y \) are equally distributed, then so are \( g(X), g(Y) \). This is because
\[ P[g(X) \leq t] = P[X \in g^{-1}(-\infty, t]] = P[Y \in g^{-1}(-\infty, t]] = P[g(Y) \leq t]. \]

The next step is to prove the claim for simple random variables: If \( X, Y \) are simple and equally distributed then
\[ P[X = r] = F_X(r) - F_X(r^-) = F_Y(r) - F_Y(r^-) = P[Y = r]. \]

Let \( R \) be a range for \( X \) and \( Y \). Linearity of expectation now gives us that
\[ \mathbb{E}[X] = \sum_{r \in R} r P[X = r] = \sum_{r \in R} r P[Y = r] = \mathbb{E}[Y]. \]

Now assume that \( X, Y \geq 0 \). Let \( g_n(r) = \min\{2^{-n} \lfloor 2^n r \rfloor, n\} \). So \( 0 \leq g_n(X) \nearrow X \) and \( 0 \leq g_n(Y) \nearrow Y \). Monotone convergence gives that
\[ \mathbb{E}[X] = \lim_n \mathbb{E}[g_n(X)] = \lim_n \mathbb{E}[g_n(Y)] = \mathbb{E}[Y], \]
where we have used the fact that \( g_n(X), g_n(Y) \) are equally distributed simple random variables for all \( n \). Note that the above also holds if \( \mathbb{E}[X] = \mathbb{E}[Y] = \infty \).
Finally, for general equally distributed $X,Y$: $X^+, Y^+$ are equally distributed and non-negative. Thus, $E[X^+] = E[Y^+]$. Similarly for $E[X^-] = E[Y^-]$. Thus,

$$E[X] = E[X^+] - E[X^-] = E[Y^+] - E[Y^-] = E[Y],$$

if at least one of $E[X^+] = E[Y^+]$ or $E[X^-] = E[Y^-]$ is finite. If they are both infinite, then both $E[X]$ and $E[Y]$ do not exist. \(\Box\)

**Exercise 24.3.** $X \sim U[-1,1]$. Calculate:

- $P[|X| > 1/2]$ .
- $P[\sin(\pi X/2) > 1/2]$.

**Solution to Exercise 24.3.** Note that $\{|X| > 1/2\} = \{X > 1/2\} \cup \{X < -1/2\}$. So,

$$P[|X| > 1/2] = P[X > 1/2] + P[X < -1/2] = \frac{1}{4} + \frac{1}{4} = \frac{1}{2}.$$

Since $\sin$ is increasing in $[-\pi/2, \pi/2]$ and since $\sin(\pi/6) = 1/2$, we get that

$$P[\sin(\pi X/2) > 1/2] = P[\pi X/2 > \pi/6] = P[X > 1/3] = 1/3.$$

\(\Box\)

**Exercise 24.4.** $B, C$ are independent and $B \sim \text{Exp}(\lambda)$, $C \sim U[0,1]$. What is the probability that the equation $x^2 + 2Bx + C = 0$ has two different real solutions.

**Solution to Exercise 24.4.** We are looking for the probability that

$$P[4B^2 - 4C > 0] = P[B^2 > C].$$

For any $t > 0$,

$$P[B^2 \leq t] = P[B \leq \sqrt{t}] = \int_0^{\sqrt{t}} \lambda e^{-\lambda s} ds = \int_0^t \frac{1}{2\sqrt{u}} \lambda e^{-\lambda \sqrt{u}} du,$$

so $B^2$ is absolutely continuous with this density. Note that for all $s > 0$,

$$\int_s^\infty f_{B^2}(u) du = \int_s^\infty \frac{1}{2\sqrt{u}} \lambda e^{-\lambda \sqrt{u}} du = \int_s^\infty e^{-\lambda t} dt = e^{-\lambda \sqrt{s}}.$$

Since $B^2, C$ are independent, we have that

$$f_{B^2|C}(t|s) = f_{B^2}(t).$$
Thus,
\[
\mathbb{P}[B^2 > C] = \int_0^1 \int_s^\infty f_{B^2,C}(t,s)dt\,ds = \int_0^1 \int_s^\infty f_{B^2|C}(t|s)dt\,ds = \int_0^1 e^{-\lambda \sqrt{s}}\,ds
\]
\[
= 2 \int_0^1 u e^{-\lambda u}du = -\frac{2u}{\lambda} e^{-\lambda u}\bigg|_0^1 + 2 \int_0^1 e^{-\lambda u}du
\]
\[
= -\frac{2}{\lambda} e^{-\lambda} - \frac{2}{\lambda^2}(e^{-\lambda} - 1) = \frac{2}{\lambda^2} \left(1 - (1 + \lambda)e^{-\lambda}\right).
\]

We have now a non-trivial inequality: \(\frac{2}{\lambda^2} \left(1 - (1 + \lambda)e^{-\lambda}\right) \leq 1\), which is equivalent to
\[
e^\lambda - (1 + \lambda) \leq \frac{\lambda^2}{2} \cdot e^\lambda.
\]

This indeed holds as \(2k! \leq (k + 2)!\) for all \(k\) and
\[
e^\lambda - (1 + \lambda) = \sum_{k=0}^\infty \frac{\lambda^{k+2}}{(k+2)!} \leq \sum_{k=0}^\infty \frac{\lambda^k}{k!} \cdot \frac{\lambda^2}{2}.
\]

\(\square\)

**Exercise 24.5.** Let \((X,Y)\) be jointly absolutely continuous with
\[
f_Y(s) = \begin{cases} 
C(s^3 - s - 1) & s \in [2,4] \\
0 & \text{otherwise}
\end{cases}
\]
and for all \(s \in [2,4],\)
\[
f_{X|Y}(t|s) = \begin{cases} 
C(s) \cos\left(\frac{\pi}{2s}t\right) & t \in [0,s] \\
0 & \text{otherwise}
\end{cases}
\]

Calculate:
\begin{itemize}
  \item \(C\) and \(C(s)\) for all \(s \in [2,4].\)
  \item \(\mathbb{E}[Y], \mathbb{E}[X|Y = s], \mathbb{E}[X].\)
  \item \(\text{Var}[Y], \text{Var}[X].\)
  \item \(\text{Cov}(X,Y)\) and \(\text{Var}[X + Y].\)
\end{itemize}

**Solution to Exercise 24.5.** For \(C:\)
\[
1 = C \int_2^4 (s^3 - s - 1)\,ds = C \cdot \left(\frac{4^4}{4} - \frac{4^2}{2} - 4 - \frac{2^4}{4} + \frac{2^2}{2} + 2\right) = C \cdot (60 - 6 - 2) = C \cdot 52.
\]
For $C(s)$:

\[
1 = C(s) \int_0^s \cos(\frac{\pi}{2s} t) dt = C(s) \cdot \frac{2s}{\pi} \sin(\frac{\pi}{2s} t) \bigg|_0^s
\]

\[
= C(s) \cdot \frac{2s}{\pi},
\]

So $C(s) = \frac{\pi}{2s}$.

We also have

\[
\mathbb{E}[Y] = 52^{-1} \int_2^4 (s^4 - s^2 - s) ds = \frac{5^5 - 2^5}{5 \cdot 52} - \frac{3^3 - 2^3}{3 \cdot 52} - \frac{2^2 - 2^2}{2 \cdot 52} = 5212^{12}_{30}.
\]

For $s \in [2, 4]$, using the fact that $(x \sin x + \cos x)' = \sin x + x \cos x - \sin x = x \cos x$ we have

\[
\mathbb{E}[X|Y = s] = C(s) \int_0^s t \cos(C(s) t) dt = C(s)^{-1} \int_0^{\pi/2} t \cos t dt
\]

\[
= C(s)^{-1} \cdot (\frac{\pi}{2} - 1) = (1 - \frac{2}{\pi}) \cdot s.
\]

To calculate $\mathbb{E}[X]$ we use the fact that $f_{X,Y}(t, s) = f_Y(s) f_{X|Y}(t|s)$ for all $s \in [2, 4]$ and $f_{X,Y}(t, s) = 0$ otherwise. Thus, with the function $(t, s) \mapsto t$,

\[
\mathbb{E}[X] = \int_2^4 f_Y(s) \int_0^s t f_{X|Y}(t|s) dtds = (1 - \frac{2}{\pi}) \cdot 52^{-1} \cdot \int_2^4 s f_Y(s) ds = (1 - \frac{2}{\pi}) \cdot \mathbb{E}[Y].
\]

Now, similarly

\[
\mathbb{E}[XY] = \int_2^4 s f_Y(s) \int_0^s t f_{X|Y}(t|s) dtds = (1 - \frac{2}{\pi}) \cdot \mathbb{E}[Y^2],
\]

so we get that

\[
\text{Cov}(X, Y) = (1 - \frac{2}{\pi}) \cdot \mathbb{E}[Y^2] - (1 - \frac{2}{\pi}) \cdot \mathbb{E}[Y] \cdot \mathbb{E}[Y] = (1 - \frac{2}{\pi}) \cdot \text{Var}[Y].
\]

Exercise 24.6. A chicken lays an egg of random size. The size of the egg has the distribution of $|N|$ where $N \sim N(0, 5)$. For any $s > 0$, given that an egg is of size $s$, the time until the egg hatches is distributed $\text{Exp}(s^{-2})$. What is the expected time for an egg to hatch?
Solution to Exercise 24.6. Let \( S \) be the size and \( T \) the time. We are given that \( S \sim |N(0, 5)| \). That is, for any \( s > 0 \),
\[
\mathbb{P}[S \leq s] = \mathbb{P}[N(0, 5) \in [-s, s)] = \int_{-s}^{s} \frac{1}{\sqrt{2\pi}5}e^{-t^2/50}dt = 2 \int_0^{s} \frac{1}{\sqrt{2\pi}5}e^{-t^2/50}dt.
\]
So \( S \) is absolutely continuous with density
\[
f_S(s) = \begin{cases} 
\frac{\sqrt{2}}{5\sqrt{\pi}} e^{-s^2/50} & s > 0 \\
0 & s \leq 0
\end{cases}
\]
So we can calculate the expectation of \( T \):
\[
\mathbb{E}[T] = \int \int tf_{T,S}(t,s)dtds = \int_0^{\infty} f_S(s) \int_0^{\infty} ts^{-2}e^{-s^2-2t}dtds \\
= \int_0^{\infty} s^2 f_Y(s) ds = \mathbb{E}[S^2] = \mathbb{E}[N(0, 5)^2] = 25.
\]
\[\square\]

Exercise 24.7. Let \( X \) be a discrete random variable with range \( \mathbb{Q} \). Suppose that \( \mathbb{E}[X] = \mu, \mathbb{E}[X^2] = \sigma \) and \( \mathbb{E}[X^3] = \rho \). Let \( Y \) be a discrete random variable defined by \( Y|X = q \sim \text{Poi}(q^2) \).

Calculate \( \text{Cov}(X, Y) \).

Solution to Exercise 24.7. Note that for any \( q \in \mathbb{Q} \),
\[
q^2 = \mathbb{E}[Y|X = q] = \sum_{k=0}^{\infty} ke^{-q^2}q^{2k}/k!.
\]
Note that for \( q \in \mathbb{Q} \) and \( k \in \mathbb{N} \),
\[
f_{Y,X}(k, q) = f_X(q)e^{-q^2}q^{2k}/k!.
\]
So
\[
\mathbb{E}[XY] = \sum_{q \in \mathbb{Q}} \sum_{k=0}^{\infty} qk f_{Y,X}(k, q) = \sum_{q \in \mathbb{Q}} q^3 f_X(q) = \rho,
\]
and
\[
\mathbb{E}[Y] = \sum_{q \in \mathbb{Q}} \sum_{k=0}^{\infty} k f_{Y,X}(k, q) = \sum_{q \in \mathbb{Q}} q^2 f_X(q) = \sigma.
\]
Thus, \( \text{Cov}(X, Y) = \rho - \sigma \mu \). \[\square\]