Important concepts from today: independence (especially for more than two random variables), uncorrelated doesn’t mean independent, scalings of normal random variables.

1. Review of last time

Last time, we talked about what it means for one measure to be absolutely continuous with respect to the other:

\( \mu \ll \nu \) means that if \( \nu(A) = 0 \), then also necessarily \( \mu(A) = 0 \).

If the only set \( A \) such that \( \nu(A) = 0 \) is \( A = \emptyset \), then \( \mu \ll \nu \) for any other measure \( \mu \) defined on the same space. For example, any measure on the nonnegative integers is absolutely continuous with respect to the Poisson.

When \( \mu \) and \( \nu \) are measures on \( \mathbb{R} \), we also saw what absolute continuity implies for the CDFs: if \( \mu \ll \nu \) then

1. whenever the CDF of \( \nu \) is flat, the CDF of \( \mu \) is also flat;
2. whenever the CDF of \( \mu \) has a jump, the CDF of \( \nu \) also has a jump (but it may jump by a different amount).

We stated that if \( \mu \ll \nu \), then we can write \( \mu(A) = \int_A h(x) \, d\nu(x) \), where the function \( h(x) = d\mu/d\nu(x) \) is called the Radon-Nikodym derivative. When \( \nu \) is the Lebesgue measure, the derivative is the ordinary derivative of the CDF, and is also called the pdf (probability density function). When \( \mu \) and \( \nu \) both have densities, the derivative \( d\mu/d\nu \) is the ratio of the densities.

On discrete sets (finite, integers, etc.),

\[
\frac{d\mu}{d\nu}(x) = \frac{\mu(A)}{\nu(A)}, \quad A \text{ is the smallest measurable set containing } x.
\]

Often, \( A = \{x\} \), but if \( \{x\} \) isn’t measurable on its own, you take the smallest set containing it that is measurable. This method doesn’t work for continuous sets like \( \mathbb{R} \), because one-point sets have measure 0, and the fraction 0/0 wouldn’t be defined. Instead, you have to take the limit over small sets \( A \) containing \( x \).

**Example 1.** Let \( S = \{1, 2, 3, 4, 5, 6\} \). Let \( \mathcal{F} \) be the \( \sigma \)-algebra containing the sets \( \{1\}, \{2\}, \{3, 5\}, \) and \( \{4, 6\} \), as well as their arbitrary unions. (Thus, \( \mathcal{F} \) has \( 2^4 = 16 \) elements.)

Let \( \mu \) be the measure given by \( \mu(\{1\}) = 0, \mu(\{2\}) = \mu(\{3, 5\}) = \mu(\{4, 6\}) = 1/3 \). Let \( \nu \) be the measure given by \( \nu(\{1\}) = 0, \nu(\{2\}) = 1/3, \nu(\{3, 5\}) = 0, \nu(\{4, 6\}) = 2/3 \).

The only sets of \( \mu \)-measure 0 are \( \emptyset \) and \( \{1\} \). Since these also have \( \nu \)-measure 0, we see \( \nu \ll \mu \). We also see \( \mu \nleq \nu \), since \( \nu(\{3, 5\}) = 0 \) but \( \mu(\{3, 5\}) > 0 \).
The Radon-Nikodym derivative $d\nu/d\mu$ must be a measurable function, that is, it will take one value on 3 and 5, and one value on 4 and 6. To compute it, we write

\[
\frac{d\nu}{d\mu}(1) = \frac{\nu(\{1\})}{\mu(\{1\})} = 0/0 = \text{not defined}
\]

\[
\frac{d\nu}{d\mu}(2) = \frac{\nu(\{2\})}{\mu(\{2\})} = 1/3 = 1
\]

\[
\frac{d\nu}{d\mu}(3) = \frac{\nu(\{3,5\})}{\mu(\{3,5\})} = 0/1/3 = 0
\]

\[
\frac{d\nu}{d\mu}(4) = \frac{\nu(\{4,6\})}{\mu(\{4,6\})} = 2/3/1/3 = 2
\]

\[
\frac{d\nu}{d\mu}(5) = \frac{\nu(\{3,5\})}{\mu(\{3,5\})} = 0
\]

\[
\frac{d\nu}{d\mu}(6) = \frac{\nu(\{4,6\})}{\mu(\{4,6\})} = 2.
\]

The derivative is not defined at 1, but $\{1\}$ has measure 0, which means that it “doesn’t happen”.

2. Independent random variables

Two random variables $X, Y$ defined on the same probability space are called independent if for any (measurable) events $A, B \subset T$,

\[
P(X \in A \text{ and } Y \in B) = P(X \in A)P(Y \in B).
\]

Let’s do some examples, and see why this should mean that $X$ and $Y$ are independent.

Example 2. (1) The canonical example is tossing coins, or rolling dice. You toss two coins, and whether the second coin comes up heads doesn’t depend on what the first coin did. Therefore, the probability that the first coin comes up heads and the second comes up tails, is $P(H,T) = P(H)P(T)$. Or, you roll a die twice, and the probability that you roll a 3 and then a 5 is $P(3,5) = P(3)P(5)$.

Let’s say this formally. When we roll a die twice, we are looking at a random variable defined on $T = \{1,2,3,4,5,6\}^2$, the set of pairs of numbers 1 through 6. The two random variables we’re looking at are

\[
X((s_1,s_2)) = s_1, \quad Y((s_1,s_2)) = s_2.
\]

The probability that $X \in A$ and $Y \in B$ (for example, $X = 3$ and $Y = 5$) is

\[
P(X \in A \text{ and } Y \in B) = \frac{\#(s_1 \in A \text{ and } s_2 \in B)}{\#(s \in S)^2} = \frac{\#(s_1 \in A \text{ and } s_2 \in B)}{\#(s \in S)^2}.
\]

If we’re looking at the uniform probability measure, i.e. all pairs of outcomes are equally likely, then this probability is simply

\[
P((s_1, s_2) \in A \times B) = \frac{\#(\text{pairs } (s_1, s_2) \text{ in } A \times B)}{\#(\text{pairs } (s_1, s_2) \in S = \{1, \ldots, 6\}^2)} = \frac{\#(s_1 \in A) \cdot \#(s_2 \in B)}{\#(s \in S)^2}.
\]
But the right-hand side is also
\[
\frac{\#(s_1 \in A) \cdot \#(s_2 \in B)}{\#(s \in S)^2} = \frac{\#(s_1 \in A)}{\#(s \in S)} \cdot \frac{\#(s_2 \in B)}{\#(s \in S)} = \mathbb{P}(s_1 \in A)\mathbb{P}(s_2 \in B) = \mathbb{P}(X \in A)\mathbb{P}(Y \in B).
\]

In our example \(A = \{3\}\), \(B = \{5\}\), this is of course \(1/36 = 1/6 \cdot 1/6\), but fundamentally this is just generalizing the idea that the number of outcomes in the product set \(A \times B\) is the product of the number of outcomes in \(A\) and the number of outcomes in \(B\).

(2) In general, when I have two random variables \(X, Y : S \rightarrow T\), I can define a pair of independent random variables
\[
\tilde{X} : S \times S \rightarrow T, \quad \tilde{X}((s_1, s_2)) = X(s_1), \quad \tilde{Y} : S \times S \rightarrow T, \quad \tilde{Y}((s_1, s_2)) = Y(s_2).
\]
That is, I make \(\tilde{X}\) and \(\tilde{Y}\) be functions of independent components of \((s_1, s_2) \in S \times S\). Notice that \(\tilde{X}\) and \(\tilde{Y}\) are defined on the set \(S \times S – \text{not}\) the original set \(S\)!

In particular, while \(X\) and \(\tilde{X}\) have the same distribution (meaning \(\mathbb{P}(X \in A) = \mathbb{P}(-X \in A)\) always), they are not the same random variable!

(3) Often when \(X\) and \(Y\) are independent random variables defined on a space \(S\), it’s possible to think of \(S\) as a product set, but sometimes it’s easier not to do that. For example, let \(X, Y : [0, 1] \rightarrow \mathbb{R}\) be given by
\[
X(s) = \begin{cases} 0, & s < \frac{1}{4} \\ 1, & 1/2 \leq s \leq 1 \end{cases}, \quad Y(s) = \begin{cases} 0, & s < \frac{1}{4} \\ 1, & 1/4 \leq s < 1/2 \\ 0, & 1/2 \leq s < 3/4 \\ 1, & 3/4 \leq s \leq 1. \end{cases}
\]

Let’s check that they are independent. Since \(X\) and \(Y\) take on two values each (0 and 1), and each with probability 1/2, it’s enough to check that \(\mathbb{P}(X = 0, Y = 0) = 1/4\). (This one probability determines all the rest of them, because for example, \(\mathbb{P}(X = 0, Y = 1) = \mathbb{P}(X = 0) - \mathbb{P}(X = 0, Y = 0)\).)

Checking:
\[
\mathbb{P}(X = 0, Y = 0) = \mathbb{P}(s < 1/4) = 1/4.
\]

Here we can also think of \(X\) as reading off the first digit of \(s\) in binary notation, and \(Y\) is reading off the second digit of \(s\) in binary notation. In this sense, they are defined on a product space, but we didn’t need to use that to show the independence.

(4) If \(X\) and \(Y\) are independent, and \(g_1\) and \(g_2\) are measurable functions, then \(g_1(X)\) and \(g_2(Y)\) are independent. Indeed:
\[
\mathbb{P}(g_1(X) \in A, g_2(Y) \in B) = \mathbb{P}(X \in g^{-1}(A), Y \in g^{-1}(B)),
\]
and it’s pretty clear from here.

(5) Reminder: independent random variables must be defined on the same probability space. It doesn’t matter if they take the same values.

(6) Reminder: you can also have variables defined on a product space that aren’t independent. For example, let \(S = \{U, D\}^2\), let \(X : S \rightarrow \mathbb{R}\) count the number of \(D\)’s, and let \(Y : S \rightarrow \mathbb{R}\) be the indicator of whether the second letter is a \(D\). These are not independent: if \(X = 2\) then necessarily \(Y = 1\), so
\[
\mathbb{P}(X = 2, Y = 1) \neq \mathbb{P}(X = 2)\mathbb{P}(Y = 1),
\]
meaning \(X\) and \(Y\) are not independent.
Another way of thinking about independence is “does knowing the value of \( X \) tell you anything about the value of \( Y \)?” If yes, the random variables aren’t independent. (Unless \( Y \) is a constant.) This way of thinking about independence leads to conditional expectations, which we’ll talk about later in the course.

We now have three concepts under our belt that we can apply to a pair of random variables:

- we can test whether they’re independent;
- we can test whether they are measurable with respect to each other;
- and we can test whether their distributions are absolutely continuous with respect to each other.

Let’s look at how (and whether) these three are related.

- For absolute continuity, it’s enough to look at each of \( X \) and \( Y \) separately; measurability and independence both require us to look at the values \( X(\omega) \) and \( Y(\omega) \) simultaneously, on the same \( \omega \). That is, measurability and independence are about the relationship between the measurable (identifiable) sets \( \sigma(X) \) and \( \sigma(Y) \), whereas absolute continuity is about the values that \( X \) and \( Y \) take.
- Measurability has nothing to do with the distribution or the probability measure (beyond knowing which sets are measurable). Absolute continuity needs to know which sets have non-zero measure. Independence requires knowing the full (joint) distribution of \( X \) and \( Y \): I can’t tell whether two random variables are independent without knowing their probability distribution.
- Independence is essentially the opposite of measurability. If \( X \) is \( \sigma(Y) \)-measurable, then knowing the value of \( Y \) completely determines the value of \( X \).\(^1\) If \( X \) and \( Y \) are independent, then knowing the value of \( Y \) should tell me absolutely nothing about the value of \( X \). Note that if \( Y \) is a constant random variable, then it’s both \( \sigma(X) \)-measurable and independent of \( X \), for any \( X \); but that’s the only case when you can have both.

Independence can be extended to more than two random variables. When we say that \( X_1, X_2, \ldots, X_n \) are independent, we mean

\[
P(X_1 \in A_1, X_2 \in A_2, \ldots, X_n \in A_n) = P(X_1 \in A_1)P(X_2 \in A_2)\cdots P(X_n \in A_n).
\]

This implies that every pair among these random variables is independent, but it also says more than that.

For an infinite collection of random variables, we say that they are independent when every finite subcollection is independent. That is, we don’t look at infinite products above. (Infinite products involving probabilities tend to be equal to zero anyway, although they don’t have to be.)

How do we use independence? Here are two of the most important examples.

**Theorem 2.1.** If \( X_1, X_2, \ldots \) are independent and all have densities, then their joint density is the product of the individual densities:

\[
f(x_1, x_2, \ldots, x_n)(a_1, \ldots, a_n) = f_{X_1}(a_1)f_{X_2}(a_2)\cdots f_{X_n}(a_n).
\]

\(^1\)Strictly speaking, it only determines where in the \( \sigma \)-field of the image the value of \( X \) falls, but usually that means you can figure out the value itself.
Joint densities are similar to ordinary densities:

\[ \Pr(X_1 \in (a_1 + \frac{1}{2} \epsilon_1), X_2 \in (a_2 + \frac{1}{2} \epsilon_2), \ldots, X_n \in (a_n + \frac{1}{2} \epsilon_n)) \approx f(x_1, x_2, \ldots, x_n)(a_1, \ldots, a_n) \times \epsilon_1 \epsilon_2 \ldots \epsilon_n. \]

We don’t usually talk about multidimensional CDFs (because it can be hard to say what the set of points less than a certain vector is). But if you define it correctly, then

\[ f(x_1, x_2, \ldots, x_n)(a_1, \ldots, a_n) = \frac{\partial^n}{\partial x_1 \partial x_2 \ldots \partial x_n} F_{x_1, \ldots, x_n}(x_1 = a_1, \ldots, x_n = a_n). \]

As on the real line, the density is also the Radon-Nikodym derivative of the distribution of \((X_1, \ldots, X_n)\) with respect to the Lebesgue measure on \(\mathbb{R}^n\) (the \(n\)-dimensional volume).

**Theorem 2.2.** If \(X_1, X_2, \ldots, X_n\) are independent, then for any real-valued functions \(g_1, g_2, \ldots, g_n\)

\[ \mathbb{E}[g_1(X_1)g_2(X_2) \ldots g_n(X_n)] = \mathbb{E}[g_1(X_1)]\mathbb{E}[g_2(X_2)] \ldots \mathbb{E}[g_n(X_n)] \]

provided all of these expectations are defined. In particular, if \(X_i\) are real-valued,

\[ \mathbb{E}[X_1X_2 \ldots X_n] = \mathbb{E}[X_1]\mathbb{E}[X_2] \ldots \mathbb{E}[X_n]. \]

In general, expectation of the product is not the product of expectations; but for independent random variables it is. Notice that it’s still the case that \(\mathbb{E}[g(X)] \neq g(\mathbb{E}[X])\).

**Warning:** The converse is not true: \(\mathbb{E}[X \cdot Y] = \mathbb{E}[X]\mathbb{E}[Y]\) does not mean that \(X\) and \(Y\) are independent. For example, suppose \(X\) is uniform on the set \(\{1, 2, \ldots, 9, 10\}\), and suppose that \(Y = \pm X\) with probability 1/2 for each of the possibilities. Then \(X\) and \(Y\) are definitely not independent: knowing \(Y\) tells you the value of \(X\) exactly (so \(X\) is \(\sigma(Y)\)-measurable), and knowing \(X\) severely restricts the possibilities for \(Y\). On the other hand, \(XY\) can be either \(X^2\) or \(-X^2\) with probability 1/2 each, so when I compute \(\mathbb{E}[XY]\) I will get 0, which is also \(\mathbb{E}[X]\mathbb{E}[Y]\).

Another way of phrasing that is that “uncorrelated does not imply independent”. The covariance of two random variables is

\[ \text{Cov}(X, Y) = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y], \]

(correlation is covariance divided by the two standard deviations). So if \(X\) and \(Y\) are independent, then \(\text{Cov}(X, Y) = 0\); but if \(\text{Cov}(X, Y) = 0\) then \(X\) and \(Y\) may still be dependent.

(Although it is actually true that if expectation of the product of functions is equal to the product of expectations of functions for all functions, then the random variables are independent.)

**Proposition 2.3.** Here’s an alternative expression for the covariance:

\[ \text{Cov}(X, Y) = \mathbb{E}[(X - \mathbb{E}X)(Y - \mathbb{E}Y)]. \]

**Proof.** Expand the product, and recall that expectation is linear:

\[ \mathbb{E}[XY - \mathbb{E}[X] \cdot Y - X \cdot \mathbb{E}[Y] + \mathbb{E}[X]\mathbb{E}[Y]] = \mathbb{E}[XY] - 2\mathbb{E}[X]\mathbb{E}[Y] + \mathbb{E}[X]\mathbb{E}[Y] \]

\[ = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]. \]  

This shows you that covariance isn’t about the mean of the random variables, it’s about oscillations about that mean.

Here are some other properties of covariance:
(1) If \( X \) is a constant random variable, and \( Y \) is any other variable, then \( \text{Cov}(X, Y) = 0 \). This follows directly from independence. (There’s a partial converse: if \( \text{Var}(X) = \text{Cov}(X, X) \) is zero, then \( X = \mathbb{E}[X] \) is a constant with probability 1.)

(2) Covariance is \textit{bilinear}: \( \text{Cov}(aX + bY, cZ + dW) = ac \text{Cov}(X, Z) + bc \text{Cov}(Y, Z) + ad \text{Cov}(X, W) + bd \text{Cov}(Y, W) \). (Expand the product and check!)

(3) As a consequence of this, \( \text{Cov}(X + a, Y + b) = \text{Cov}(X, Y) \) for any constants \( a \) and \( b \). We just saw one example of this when \( a = -\mathbb{E}[X], b = -\mathbb{E}[Y] \): the covariance of \( X \) and \( Y \) is equal to the covariance of those random variables \textit{centered} to have mean zero.

3. Normal random variables

The normal random variable \( N(\mu, \sigma^2) \) has mean \( \mu \), variance \( \sigma^2 \), and pdf

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

The \textit{standard} normal random variable is \( N(0, 1) \).

The normal random variable will be important to our course, essentially for the following reason.

Let \( X_1, X_2, \ldots \) be independent real-valued random variables (hence defined on the same probability space), and suppose they have the same law. Such variables are called \textit{iid}, meaning \textit{independent and identically distributed}. Let’s also suppose that \( \mathbb{E}[X_i] = \mu \) and \( \text{Var}(X_i) = \sigma^2 \). Then the average of many of the \( X_i \),

\[
\frac{1}{n} \sum_{i=1}^{n} X_i \to \mu, \quad \text{as } n \to \infty.
\]

that is, the average of many iid random variables converges to their mean. This is the law of large numbers. (We should be a little more precise by what we mean about convergence here, because we’re working with random variables. The correct statement here is that we have almost sure convergence, i.e. for almost all \( \omega \in \Omega \), \( \lim_{n \to \infty} n^{-1}(\sum_{i=1}^{n} X_i(\omega)) = \mu \).

However, suppose we’re interested in the deviations from that mean, then we have the (much harder!) central limit theorem:

\textbf{Theorem 3.1} (Central limit theorem).

\[
\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{X_i - \mu}{\sigma} \implies N(0, 1), \quad \text{as } n \to \infty.
\]

Here, what we mean is that the random variables on the left converge to the random variable on the right \textit{in distribution}: if we plot the CDF of the left-hand side as a function of \( n \), it will converge to the CDF of the \( N(0, 1) \) random variable. I’ve used two different arrows (\( \to \) and \( \implies \)) to differentiate the two notions of convergence.

There’s actually a more general statement of the central limit theorem, which allows the \( X_i \) to have different means and variances, and even allows them to be slightly dependent, but we won’t go there. In practice, the central limit theorem says that anything that’s the sum of many small independent contributions should be approximately normal. For example, people’s height is influenced by many factors and is approximately normal; size of apples in an orchard tends to be approximately normal; and so on. (Another reason the normal distribution appears often is that it turns out to be a very nice mathematical object for proving results.)
Let’s come back to normal variables. Here’s one fundamental result about normal random variables:

**Proposition 3.2.** If $X$ is a standard normal variable, then $Y = \sigma X + \mu$ is a $N(\mu, \sigma^2)$ random variable.

Note, it’s clear that $\mathbb{E}[Y] = \mu$ and $\text{Var}(Y) = \sigma^2$. What’s not clear is that the entire distribution has the right functional form.

**Proof.** To check that the random variables are equal, we need to compare their CDFs.

\[
F_Y(a) = \mathbb{P}(Y \leq a) = \mathbb{P}(X < \frac{a - \mu}{\sigma}) = F_X\left(\frac{a - \mu}{\sigma}\right) = \int_{-\infty}^{(a-\mu)/\sigma} \frac{1}{\sqrt{2\pi}} e^{-x^2/2} \, dx
\]

On the other hand, for a $N(\mu, \sigma^2)$ random variable, we have

\[
F_{N(\mu, \sigma^2)}(a) = \int_{-\infty}^{a-\mu/\sigma} \frac{1}{\sqrt{2\pi} \sigma} \exp\left(-\frac{1}{2} \left(\frac{x - \mu}{\sigma}\right)^2\right) \, dx
\]

Let’s change variables in this second integral to $u = (x - \mu)/\sigma$, then $du = dx/\sigma$ so $dx = \sigma du$. Thus,

\[
F_{N(\mu, \sigma^2)}(a) = \int_{-\infty}^{(a-\mu)/\sigma} \frac{1}{\sqrt{2\pi} \sigma} \exp\left(-\frac{1}{2} u^2\right) \sigma du,
\]

which after canceling the $\sigma$ is exactly what we wanted to get. \qed

We will sometimes refer to the CDF of a standard normal variable as $\Phi$:

\[
\Phi(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-u^2/2} \, du.
\]

The fact that we can express *all* normal random variables through the standard normal variable means that many computations involving normal random variables can be expressed through $\Phi$ (the CDF of the standard normal), $\mu$, and $\sigma$. 

---

7