MA 2631 Probability Theory Lecture Notes

Binan Gu

October 7, 2025

Contents

Part I
Fundamentals of Probability

1 Lecture I: Examples and Intuitions

The first lecture is meant to introduce the notion of events at a fundamental level. It raises questions about the parameters that underscore the events, the size of the sample space, and the true notion of a probability. Before we dive into the theoretic setup, we first consider a few simple experiments and their outcomes. Along the way, we learn how to count with some organization.

1.1 Examples

Example 1.1 (Routing)**.** Suppose there are three routes between cities A and B, and four routes between cities B and C. How many routes in total can one go from A to C through B?

Solution.

$$3 \times 4 = 12$$

No surprise here.

Example 1.2 (Rolling a die twice)**.**

- How many combinations are there that sum to 7? Suppose the order DOES matter.
$$\{1,6\},\{2,5\},\{3,4\},\{4,3\},\{5,2\},\{6,1\}.$$
- How many distinct combinations?
$$\{1,6\},\{2,5\},\{3,4\}.$$
- How many combinations are there in total?
$$6 \times 6 = 36,$$
i.e., each roll has 6 possible outcomes, and the two rolls are independent (what does this even mean?).
- How many combinations are there where two dice show the same value?
$$\{i,i\}, i = 1,\ldots,6$$

- How many combinations are there that show different faces?

$$6 \times 5 = 30,$$

i.e., the first roll has 6 possible outcomes, but the second roll, since we restrict it to be different from the first, only has 5 possible outcomes.

You may also think that out of a total of 36 rolls, we eliminate the ones that have the same value, per second last item, namely, 6 of them. We end up with, not surprisingly, 30 combinations that yield distinct faces.

Example 1.3 (The World Series)**.** Two baseball teams A and B play a series of 7 games. The team that wins 4 games wins the series.

- How many cases are there where team A wins the series in exactly 5 games?

Solution.

$$BAAAA, \quad ABAAA, \quad AABAA, \quad AAABA.$$

Does $AAAAB$ count? No, there is no need to play the last game since A has won 4 already.

- How about 6 games?

See Homework. You should be expecting a lot more cases than the previous experiment.

Remark 1.1. *So, why is there all of a sudden many more cases when the series lasts exactly* 6 *games? What* ***fundamentally*** *has changed? Meanwhile, is this a smart way to count?*

In our earlier examples, we saw how the size of the sample space depends on the situation we are modeling. For instance, in a 7–game series, a win "in 5 games" or "in 6 games" corresponds to different event spaces: in the 5–game case we only count sequences where one team wins exactly 4 of the first 5 games, while in the 6–game case we count sequences where the team wins exactly 4 of the first 6, with the 6th being decisive. The point is that even within the same underlying rules of the sport, the question we ask dictates which cases are included.

Example 1.4 (Chess Endgame)**.** A similar, but deeper, idea arises when thinking about strategic games such as chess. Consider the following situation: you have reached an endgame in which you could reasonably force a draw, but a win is possible if — and only if — your opponent makes a significant blunder, say by losing a rook.

- If your goal is **to win the game**, then your relevant sample space consists of lines of play in which your opponent either does or does not make this blunder. Your "event of interest" (a win) is embedded in a space where the opponent's mistake is one possible outcome (and its subsequent moves).
- If your goal is **simply to advance in the tournament**, and a draw is sufficient, then your relevant sample space looks quite different. You will choose safe moves that secure the draw. In this framing, the opponent's blunder no longer belongs to the space you are modeling, since you will not allow play to proceed in a way that depends on it.

The rules of chess have not changed, but the sample space we use to model probability has changed because our perspective – our intent – has changed.

Remark 1.2. *This illustrates a central point: probability theory is not just about counting outcomes, but about constructing an appropriate event space relative to the question we ask. Just as "win in 5" and "win in 6" require different sample spaces, "play for a win" and "play for a draw" do too.*

2 Lecture II: Probability Space

Building on the examples from last class, we now formalize our language and precisely define what we meant by "events", "experiment" and ultimately, the so-called "universe" that was often mentioned in class.

2.1 Sample Space of an Experiment

Definition 2.1. *The **sample space** of an experiment, denoted by Ω, is the set of all possible outcomes of that experiment.*

Definition 2.2. *An **event** is any collection (subset) of outcomes contained in the sample space Ω. An event is said to be **simple** if it consists of exactly one outcome and **compound** if it consists of more than one outcome.*

Example 2.1 (The World Series revisited)**.**

- One experiment may be "Play the World Series between team A and team B, until one team wins 4 games".
- $AABAA$ is an outcome.
- $\{AABAA\}$ is a simple event. Note the difference from the last item. Here, it is precise to say that

" The singleton set AABAA is a simple event, containing the outcome AABAA. "

- $\{AABAA, AAABA, ABAAA\}$ is a (compound) event – can you describe this in plain English?

  " The event where team A wins the series in 5 games, and wins game 1. "

2.2 Basic Set Theory

An event is nothing but a set, so we should formalize all discussions involving the sample space using standard set theory notations and language.

2.2.1 Notations

To shorten terms in written format, we adhere to the following notation standard. They should follow from your previous calculus classes.

Symbol	Meaning
$\{\ldots,\ldots,\ldots\}$	set of numbers
$\{\cdots \mid \cdots\}$	set of numbers such that
$\in$	is an element of
$\notin$	is not an element of
$\subset$	is a proper subset of
$\subseteq$	is a subset of
$\bigcup$	union
$\bigcap$	intersection
$\emptyset$	the empty set
$\mathbb{R}$	all real numbers
$\mathbb{R}^+$	all positive real numbers
$\mathbb{R}^n = \mathbb{R} \times \cdots \times \mathbb{R}$	n-tuples of elements of $\mathbb{R}$
$\mathbb{Z}$	all whole numbers
$\mathbb{N}$ or $\mathbb{Z}^+$	all natural numbers (positive whole numbers)

For example, we say $(a, b) \in \mathbb{R}^2 = \mathbb{R} \times \mathbb{R}$ if $a \in \mathbb{R}$ and $b \in \mathbb{R}$. The "$\times$" here is not multiplication, but simply introducing an **ordered pair** (or n-tuple). We can also say, for example, that

$$\left(\sqrt{2}, 7\right) \in \mathbb{R} \times \mathbb{N}.$$

2.2.2 Definitions

Definition 2.3 (Operations on Sets)**.**

- *The* ***Complement*** *of an event A, denoted by A', is the set of all outcomes in Ω that are not contained in A.*
- *The* ***Intersection*** *of two events A and B, denoted by $A \cap B$ and read "A and B", is the event consisting of all outcomes that are in both A and B.*
- *The* ***Union*** *of two events A and B, denoted by $A \cup B$ and read "A or B", is the event consisting of all outcomes that are either in A or in B or in both events – htat is, all outcomes in at least one of the events.*
- *A and B are said to be* ***disjoint*** *or* ***mutually exclusive*** *events when they have no outcomes in common, i.e., $A \cap B = \emptyset$.*

Venn diagrams are often used to visualize these operations. See Fig 1.1 in your text. One must be clear here that your "universe" is the sample space of your experiment (of interest) – it particularly affects the **complement** of an event (or a set of events).

Example 2.2 (Counting coins)**.** Flip a coin three times and record the sequence of outcomes.

- Describe all possible outcomes, i.e., the sample space.

  Solution.

  $$\Omega = \{HHH, HHT, HTH, HTT, THH, THT, TTH, TTT\}$$

- Describe A, the event of observing at least two heads in a row.

  Solution.

  $$A = \{HHT, THH, HHH\}$$

- What's the complement of this event?

  Solution.

  $$A' = \{HTH, HTT, THT, TTH, TTT\}$$

- Describe the event $B =$ "The first flip is a head".

  Solution.

  $$B = \{HHH, HHT, HTH, HTT\}$$

- What's the intersection of A and B, and how do you describe it in English?

  Solution.

  $$A \cap B = \{HHT, HHH\}$$

  The description is somewhat an afterthought: the event where we have at least two heads, and the first two are heads.

Example 2.3. Consider the set of positive integers $\mathbb{N}$.

- If $\Omega = \mathbb{Z}$, that is, all integers, then $\mathbb{N}'$ is $\{0\} \cup \mathbb{Z}^-$, i.e., 0 and all negative integers.
- If $\Omega = \mathbb{R}$, then the complement includes all negative real numbers, 0, and all positive real numbers that are not integers.

Remark 2.1. *You may think of Ω as your "universe", which certainly endows your perspective on the same event. If your life experience is as dense as the real numbers, then you can see the opposite of an event in many more ways than if your experience contains merely single encounters like the sparse positive integers.*

3 Lecture III: The Axioms of Probability

In the previous class, we introduced the fundamental set operations. These are important because, in probability theory, events are modeled as sets. In this class, we turn to the axioms of probability, which provide the foundation for assigning probabilities to events and reasoning about them.

3.1 Fundamentals and Objective Frequency

Suppose we have an experiment with sample space Ω. It is now our job to quantify the event from this sample space by assigning it to a real number $P(A)$, called **the probability of event** A. We keep things general, without getting into detailed examples first.

To guarantee that probability assignments align with our intuitive understanding, they are required to satisfy the following axioms of probability.

Axioms.

1. *(nonnegativity) For any event A, $P(A) \geq 0$.*
2. *(unit measure) $P(\Omega) = 1$.*
3. *(additivity) If $A_1, A_2, A_3, \ldots$ is an infinite collection of disjoint events, i.e., $A_i \cap A_j = \emptyset$ for $i \neq j$, then*

$$P(A_1 \cup A_2 \cup A_3 \cup \cdots) = \sum_{i=1}^{\infty} P(A_i). \tag{3.1}$$

Proposition 3.1.

$P(\emptyset) = 0$. *This suggests that the previous results also hold for a finite collection.*

Proof. Read the text on Page 8 of the Carlton textbook. □

Interpretations of Probability

Sometimes, we often associate probability of an event with the relative frequency at which the event occurs to the sample space. A typical example is that it rains 2 days in a week, and we then would "predict" that probability of raining is 2/7 for the next day (which we know obviously depends on many things).

So, what is this relative frequency? Consider an experiment which you perform independently and identically, and let A be an event consisting of a fixed set of

outcomes of the experiments, e.g., seeing two heads after tossing a fair coin three times. You conduct this experiment many times, say n times, and then record the number of times that the desired event has occurred, say $n(A)$. Then, naturally you would associate

$$P(A) = \frac{n(A)}{n}.$$

This number, as n grows, could possibly approach a fixed number (recall the notion of convergence from Calculus III). Does every single experiment yield such a convergence property? Estimates like such gives rise to the *objective interpretation of probability*, though it is not mathematically well-grounded until we reach the *Law of Large Numbers* later in the course. Therefore, one must not trust short-term (small n) estimates, and ought to develop the skills to quantify "how large" is large enough. We will get to the heart of this business in the future.

However, based on this objective interpretation, we can still make valid assumptions when assigning probabilities. A head or a tail, for a fair coin, should be equally likely, and are thus of probability 0.5. Each face of a fair die has 1/6 probabilities of being landed on.

3.2 Probability based on Set Operations

Theorem 3.1 (Set Theoretic Rules)**.**

1. *(Complementation) For any event A, $P(A) = 1 - P(A')$.*
2. *For any event A, $P(A) \leq 1$.*

Proof. We know $1 = P(A) + P(A') \geq P(A)$, since $P(A') \geq 0$ by Axiom 1. □

Theorem 3.2 (Inclusion/Exclusion)**.** *For any event A and B,*

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

Proof. A picture speaks loudly, but we must put it to use *with* the Axioms we gave. At the moment, the only rule that helps us work with multiple events is the disjoint event (set) axiom (Axiom 3 per Equation (3.1)). The picture can help us identify these disjoint sets.

First, $A \cup B = A \cup (B \cap A')$, and this union is disjoint, evidently from the picture. Thus,

$$P(A \cup B) = P(A) + P(B \cap A'). \tag{3.2}$$

Great, now we need to express either $B \cap A'$ as a disjoint union, or part of a disjoint union that forms a larger set. We observe that

$$B = (B \cap A') \cup (B \cap A)$$

which implies, by Axiom 2, again, that

$$P(B) = P((B \cap A') \cup (B \cap A)) = P(B \cap A') + P(B \cap A)$$

and thus

$$P(B \cap A') = P(B) - P(A \cap B).$$

Substituting this into Equation (3.2) yields the desired result. □

Example 3.1 (Pen usage)**.**
60% of the students in my class use red pens, 50% use black ones, and 40% use both. Find

- the proportion of students that only use red pens.

  Solution. Let R be the event of (observing a student) using the red pen, and B for black pen.

  $$P(R \cap B') = P(R) - P(R \cap B) = 0.6 - 0.4 = 0.2$$

- the proportion of students that use at least one of the red or black pens.

  Solution. This is a classic inclusion-exclusion problem. Let A and B be the event of finding a student using red and black pens, respectively. Then, by the Inclusion-Exclusion principle, we have

  $$\begin{aligned} P(R \cup B) &= P(R) + P(B) - P(R \cap B) \\ &= 0.6 + 0.5 - 0.4 \\ &= 0.7 \end{aligned}$$

- the proportion of students that use neither pens.

  Solution. By De Morgan's law (verification is in your HW)

  $$P(neither) = P(R' \cap B') = 1 - P(R \cup B) = 1 - 0.7 = 0.3.$$

  Remark. *Intuitively, this should make sense. The event of neither happening should be complemented by the event of either happening.*

3.3 Probability based on Equally Likely Outcomes

Suppose each simple event E_i in an experiment with N possible outcomes has equal probability p. Then, since simple events are mutually exclusive,

$$1 = \sum_{i=1}^{N} P(E_i) = \sum_{i=1}^{N} p = pN \implies p = \frac{1}{N}.$$

For compound event A, let $N(A)$ denote the number of outcomes contained in A. Then

$$P(A) = \sum_{E_i \in A} P(E_i) = \sum_{E_i \in A} \frac{1}{N} = \frac{N(A)}{N}.$$

Example 3.2 (Probability of rolling to a sum of 7)**.** Consider rolling two dice separately. What is the probability of finding the sum to be 7?

Solution. We have a total of $N = 36$ outcomes, each of which is equally likely should both dice are fair. Thus, observing any particular simple event has probability $P(E_i) = 1/36$. Let A be the event where the we observe the sum of two numbers is 7.

$$P(A) = \frac{N(A)}{N} = \frac{6}{36} = \frac{1}{6}.$$

Can we compute the probability in general when asked for any permissible sum from rolling two dice separately? Yes, in fact, we form the incredible pyramid of probability. See the sketch in class.

4 Lecture IV: Basic Counting Principles

When the outcome of each experiment is equally likely, then the task of computing probabilities reduces to counting. Experiments with this property, such as rolling a fair six-sided die or a fair two-sided coin, are quite easy to analyze due to the small size of their sample spaces. However, when the number of possible outcomes becomes big, such as a fair 38-slotted Roulette machine, counting becomes computational prohibitive, and we must resort to more intelligent ways of counting. Furthermore, the individual outcomes also may not be equally likely, for which we must provide a set of rules of computation.

Let's count with organization. First, we state the **Principle of Counting**:

If we have two experiments, the first with n possible outcomes and the second with m possible outcomes, then there are altogether $n \times m$ possible outcomes.

We can generalize to more than two experiments via the **Generalized Principle of Counting**:

If we have r experiments, the first with n_1 possible outcomes, the second with n_2, and so on (up to the rth with n_r possible outcomes), then the total number of possible outcomes is

$$n_1 \cdot n_2 \cdot \cdots \cdot n_r = \prod_{j=1}^{r} n_j.$$

The product sign $\prod$ is just a useful abbreviation for large products – as the sum symbol Σ is for sums.

Example 4.1 (License Plates)**.** Consider a 7-digit license plates where the first three slots have to be filled with letters and the four others with numbers.

- How many possible license plates are there?

  Solution. By generalized principle of counting, we have

  $$26 \cdot 26 \cdot 26 \cdot 10 \cdot 10 \cdot 10 \cdot 10 = 26^3 \cdot 10^4 = 175,760,000.$$

- What if we only allow every letter and every number to appear at most once on the plate?

  Solution. With some restriction, we have

  $$26 \cdot 25 \cdot 24 \cdot 10 \cdot 9 \cdot 8 \cdot 7 = 78,624,000.$$

4.1 Permutations

Example 4.2 (Rearrangement of letters)**.** How many ways are there to arrange the letters a, b and c?

Solution. Brute-force listing yields 6 possible cases. However, what if then the question is all 26 letters? Can we list them all?

Generalized counting helps here:

The first slot has 3 possibilities. Whatever letter you end up using first, it won't be used for the second slot, which then must have just 2 possibilities. The last slot then is fixed. We have in total

$$3 \cdot 2 \cdot 1 = 6.$$

In general, if we have n distinct objects, then we have

$$n \cdot (n-1) \cdot (n-2) \cdot \cdots \cdot 2 \cdot 1 = \prod_{j=1}^{n} j = n!$$

This product is so useful that we give it a name, "factorial" – it will be used quite often next.

Example 4.3 (Scores)**.** We have m guys and n girls in this probability class. Assuming that everyone's score at the end of the term is distinct.

- How many different rankings are possible?

  Solution.

  $$(n+m)!$$

- If guys and girls are ranked separated, how many different rankings are possible?

  Solution. Separately, we have $m!$ and $n!$ possible rankings for each group, respectively. Then by the Principle of Counting, the number of permutations is

  $$n! \cdot m!.$$

Example 4.4 (With Repeated Letters I)**.** How many words can be formed with the letters from the following word "**ERR**"?

Solution. Here, we have a slight issue. R is repeated twice, which means, exchanging one R with another will not produce a new word.

Suppose that the R's are distinctly labeled, say, we have the word $E_1R_1R_2$. Since they are fictitiously distinct, we have a total of 3! ways of arranging them. In fact, we can list them

$$\begin{array}{ll} ER_1R_2, & ER_2R_1 \\ R_1ER_2, & R_2ER_1 \end{array}$$

$$R_1R_2E, \quad R_2R_1E.$$

In reality, the R's are not distinct. From the list, one can observe that each row is equivalent, which means, the true number of words formed by **ERR** is 3.

There must be a better mathematical way of counting this. Consider the following argument: **given a fixed configuration** of the non-repeated letter, in this case, just the single letter E, how many ways can you arrange the labeled R's? 2!. This means you overcounted twice as much, and in order to get the true number of words, you perform

$$\frac{3!}{2!} = 3.$$

Remark 4.1. *If a letter appears more than once, then whenever you fix the positions of all the other letters, you can still swap those identical copies among themselves without creating a new word. For example, with two R's there are always* 2! *ways to shuffle them, but they all look the same once the labels are removed. To avoid counting these duplicates, we divide by* 2!.

Let's take this idea from the remark and go to a slightly more complicated arrangement problem.

Example 4.5 (With Repeated Letters II)**.** How many words can be formed with the letters from the following word "**PEPPER**"?

Solution. Notice that the letter **P** is repeated thrice, and E twice. This means that exchanging two P's will NOT yield a new word.

Let's pretend that all letters are distinct – that is, we give each letter a label

$$\mathbf{P}_1\mathbf{E}_1\mathbf{P}_2\mathbf{P}_3\mathbf{E}_2\mathbf{R}_1$$

and try to find the total number of arrangement without looking at the letters. There are 6! possibilities.

To be able to "rule out" the repeated **P**'s, just think about how many possible ways you can arrange 3 **P**'s – clearly, 3! possibilities, for **a given configuration of other non-P letters**.

Do the same for the double **R**, which has 2! possibilities, for **a given configuration of other non-P letters**.

By the argument laid out in Remark 4.1, we find that the number of words formed by "**PEPPER**" is

$$\frac{6!}{3!2!} = 60.$$

Remark 4.2. *Generalized Principle of Counting also applies here:*

*Let the number of words formed by "**PEPPER**" be x, that is, the number of words **without labeling**. Then,*

$$\underbrace{x}_{\text{unlabelled words}} \cdot \underbrace{3!}_{\text{labelled permutation of } \boldsymbol{P}} \cdot \underbrace{2!}_{\text{labelled permutation of } \boldsymbol{E}} = \underbrace{6!}_{\text{labelled permutations}}$$

which then implies that $x = \frac{6!}{3!2!} = 60$.

Theorem 4.1 (Multiset arrangement)**.** *In general, if you have n items, each item with multiplicity n_1, n_2, $\ldots n_r$ where $\sum_{i=1}^{r} n_i = n$, then the number of (unlabelled) arrangement is given by*

$$\frac{n!}{n_1!n_2!\ldots n_r!}$$

5 Lecture V: From Combinations to Conditional Probability

We continue down the same thread from last class on combinations, and then segue to a general formulation of sample spaces where outcomes are equally likely, and lastly formulate the groundwork for conditional probability.

5.1 Combinations

Example 5.1 (3-person group from 5 people)**.** Among five distinct students $ABCDE$, how many 3-person groups can you form?

Note first that the group ABC and CBA consist of the same people, so these groups are the same. This means order doesn't matter here, and our counting needs to rid of the effect of ordering.

Suppose first that we do care about ordering, then the first person has 5 possibilities, the second 4 and the third 3, which yields 60 total possible groups – which contain overcounted groups like the example we just gave.

In order to undo the overcounting, we need to divide the total possibilities by the number of ways one can arrange three items in a group (recall again that ABC and CBA are precisely the same group). There are 3! ways to do so.

Altogether, the number of distinct 3-person groups is

$$\frac{5 \cdot 4 \cdot 3}{3!} = 10.$$

Theorem 5.1 (General Combinations)**.** *The number of ways to choose k objects out of n is given by*

$$\begin{aligned} &\frac{n \cdot (n-1) \cdot (n-2) \cdot \cdots \cdot (n-k+1)}{k!} \\ =&\frac{n \cdot (n-1) \cdot (n-2) \cdot \cdots \cdot (n-k+1) \cdot (n-k) \cdot \cdots \cdot 2 \cdot 1}{k! \, (n-k) \cdot \cdots \cdot 2 \cdot 1} \\ =&\frac{n!}{k! \, (n-k)!} \end{aligned}$$

We often shorten this expression by denoting

$$\binom{n}{k} = \frac{n!}{k! \, (n-k)!}$$

where $\binom{n}{k}$ is known as the binomial coefficient, read as "n choose k".

Example 5.2 (World Series Revisited)**.** If Team A wins a 7-game series in exactly 5 games, Game 5 must be won by Team A, so there is no other choices here. However, in order to figure out the total number of possible series where this event happens, all we need to do is to find the number of ways to place the only win Team B has in the first four games. This is precisely a combination problem, where the answer is 4 choose 1, or $\binom{4}{1} = 4$.

Another approach is to find the number of ways to put 3 Team A victories in the first 4 games, which yields $\binom{4}{3} = 4$ as well.

5.2 The Laplacian Sample Space

In many experiments, we have the cases that every outcome is equally likely (e.g., rolling a fair die or flipping a fair coin). In the case of a sample of positive integers, $\Omega = \{1, 2, \ldots, N\}$, this means

$$P(\{1\}) = P(\{2\}) = \ldots = P(\{N\}) = \frac{1}{N},$$

where the last equality followed from Axiom 2 and 3 (why?).

In a *frequentist* interpretation, denoting by $|E|$ as the number of elements in the event $E \subset \Omega$, we have

$$P(E) = \frac{|E|}{|\Omega|} = \frac{\text{Number of elements in } E}{\text{Number of elements in } \Omega} \sim \frac{\text{“Number of successful outcomes”}}{\text{“Number of all outcomes”}}. \tag{5.1}$$

Definition 5.1 (Laplacian Sample Space)**.** *Sample spaces with equally likely outcomes are called Laplacian sample space*

Example 5.3 (Rolling a die twice and getting a sum of 7)**.** We did this before. Counting the number of outcomes that sum to 7, we find 6 of them (successes). There are a total of 36 outcomes. More precisely, the set of successful outcomes can be written as

$$E = \{(1,6), (2,5), (3,4), (4,3), (5,2), (6,1)\}$$

and

$$P(E) = \frac{|E|}{|\Omega|} = \frac{6}{36} = \frac{1}{6}.$$

Example 5.4 (Using Permutations and Combinations)**.**
An urn contains eleven balls, six of which are white and five black. We draw three balls. What is the probability that one of them is white and two are black?

Solution. Two ways of thinking about this:

1. Consider three slots of ordered draws. Total number of possible outcomes is

$$11 \cdot 10 \cdot 9 = 990$$

since we draw without putting the ball back.

Then we consider the number of successful combinations using the desired colors (as requested by the problem).

- WBB: $6 \cdot 5 \cdot 4 = 120$.
- BWB: $5 \cdot 6 \cdot 4 = 120$.
- BBW: $5 \cdot 4 \cdot 6 = 120$.

Great, and these events cannot happen simultaneously, meaning that they are mutually exclusive. Thus,

$$P(E) = \frac{120 + 120 + 120}{990} = \frac{4}{11} \approx 36.16\%.$$

2. Use combinations. Unrestricted number of combinations of 3 items from 11 items is $\binom{11}{3}$. Moreover, we have $\binom{6}{1}$ possibilities to draw 1 out of 6 white balls, and $\binom{5}{2}$ possibilities to draw 2 out of 5 black balls. Therefore,

$$P(E) = \frac{\text{Number of elements in } E}{\text{Number of elements in } \Omega} = \frac{\binom{6}{1} \cdot \binom{5}{2}}{\binom{11}{3}} = \frac{60}{165} = \frac{4}{11} \approx 36.16\%.$$

Example 5.5 (Classical Birthday Problem)**.**
How likely is it that (at least) 2 of N students in class have the same birthday (without considering leap years)?

Solution. Same approach, different numbers. Denote the event of interest as E. Every student has 365 possible birthday (outcomes), and therefore, the total number of outcomes is 365^N.

Sometimes, when a question is asking for "at least (small number) of items", you may consider the complement of this event, which is "at most (small number) of items". In this case, the complement of the event of interest, E^c, is that all N students have distinct birthdays. This is to say that we choose distinct N dates from 365 days, which yields $(365 \cdot 364 \cdot \cdots \cdot (365 - N + 1))$.

Altogether

$$P(E) = 1 - P(E^c) = 1 - \frac{365 \cdot 364 \cdot \ldots \cdot (365 - N + 1)}{365^N}.$$

This is an increasing function of N, since the probability of the complement (the term being subtracted) should decrease as the number of students grow, i.e., it should be increasingly difficult for us to find students of distinct birthday. For a class of our size $N = 35$, $P(E) \approx 81.4\%$.

What if I tell you now that one student has birthday on January 1st? Does the above probability change? Maybe you think this piece of information is insignificant since it's just one out of 365 days. What if I tell you the birthdays of two students? What fundamentally is happening to the probability space here?

All experiments we have done so far are ones that you have no information about, other than the procedures to carry them out, such as (the physical act of) tossing a coin or rolling a die. Your denominator is always the **unrestricted** sample space. However, as information flows in (provided by your further observation or external factors), your underlying sample space may change, because some previously possible outcomes are no longer possible. It is thus important to consider and study how additional information affects the sample space, since its size governs the probability of the event of interest (per Equation (5.1)).

6 Lecture VI: Conditional Probability

Recall that all of our examples, including the ones with Laplacian sample spaces, use Ω as the universe, i.e., we are always comparing the *relative frequency* of an event of interest to the **total number of outcomes**, which is $|\Omega|$.

Let's consider a familiar experiment, but with a different question in mind. Roll a die twice, and list the outcomes that sum up to 6. Denote this event by

$$E = \{(1,5), (2,4), (3,3), (4,2), (5,1)\}.$$

Furthermore, list the outcomes such that the first roll is a 2, denoted by

$$F = \{(2,1), (2,2), (2,3), (2,4), (2,5), (2,6)\}.$$

Very clearly, by the Laplacian sample space property, we have

$$P(E) = \frac{5}{36}, \quad P(F) = \frac{6}{36}.$$

Suppose, during an experiment, we pose the first question about event E, and then go on to observe the first roll, and see a 2 (event F), what is **now** the probability that the sum is 6? Do you expect the probability to go up or down, knowing that we have gained information (as opposed to no information at all)?

Let's just look at the two events. Knowing that the first roll is 2 which draws us to look **at** E **and** F **together**, which outcome(s) yield a sum of 6, with what probability? Just one, i.e., $(2,4)$, which has a probability of $\frac{1}{36}$ of occurring, unrestricted. But knowing that the first roll is 2 informs us that the total number of outcomes has been restricted to the ones with 2 first. This means

$$P(\text{sum to 6, given first roll is 2}) = \frac{P(\{(2,4)\})}{P(F)} = \frac{\frac{1}{36}}{\frac{1}{6}} = \frac{6}{36} > \frac{5}{36}.$$

We note that this probability is bigger than the unrestricted 5/36, which means our chances to get a sum of 6 have improved as we learned the outcome of the first roll.

The crucial lesson from the above example is that if we know already something about the experiment, then we have only to consider the cases where this holds true. We call this the conditional probability of E given F and denote it by

$$P(E \mid F)$$

Remark 6.1. *Conditional probability is not merely a formal construct. In practice, no random event comes to us entirely without context: additional observations almost always shape how we assess uncertainty. Scientists, for example, invest great*

effort in collecting and interpreting information so that experimental outcomes can be judged with confidence and quantified guarantees. Conditional probability provides the mathematical framework for incorporating such information. It allows us to refine the probability of an event E once relevant background knowledge is taken into account. In this way, it stands as one of the most fundamental and realistic tools for analyzing chance.

6.1 Conditional Probability

Definition 6.1 (Conditional Probability)**.**
The conditional probability of E given F, with $P(F) > 0$, is defined as

$$P(E \mid F) = \frac{P(E \cap F)}{P(F)}. \tag{6.1}$$

Remark 6.2. *Note that the numerator is the probability of an intersection. Recall to the previous dice roll example, where the numerator probability is computed by looking at the events E and F together, which calls for* ***intersection****, naturally.*

One immediate consequence of the formula is that if $F = \Omega$, i.e., the provided information is the sample space, then the formula reduces trivially. In other words,

$$P(E \mid \Omega) = \frac{P(E \cap \Omega)}{P(\Omega)} = P(E),$$

or, the probability of E given "something happens" with no specificity, is precisely the probability of E.

Example 6.1 (Red Pen Black Pen revisited)**.** 60% of the students in my class use red pens, 50% use black ones, and 40% use both. Given that a student uses red pens, what is the probability that the student also uses black pens? What about the other way around?

Solution. Let R and B represent the event that a student uses the red pen, and the black pen, respectively. The first question is asking for $P(B \mid R)$. We proceed by definition

$$P(B \mid R) = \frac{P(B \cap R)}{P(R)} = \frac{0.4}{0.6} = \frac{2}{3}.$$

The second question, now, approached similarly, is asking for

$$P(R \mid B) = \frac{P(B \cap R)}{P(R)} = \frac{0.4}{0.5} = \frac{4}{5}.$$

Remark 6.3. *Notice that the numerator did not change. This will help us relate $P(B)$ and $P(R)$ in a useful formula.*

Equation (6.1) can be rearrange as a product, as the intersection of two events is something we are also interested in (in applications). It is quite a handy formula.

Theorem 6.1 (The Multiplication Rule)**.**

$$P(A \cap B) = P(A \mid B)P(B)$$

Example 6.2 (Computer or Chemistry class?)**.** Bev can either take a course in computers or in chemistry. If Bev takes the computer course, then she will receive an A grade with probability 1/2; if she takes the chemistry course then she will receive an A grade with probability 1/3 Bev decides to base her decision on the flip of a fair coin. What is the probability that Bev will get an A in chemistry?

Solution. Let C be the event of Bev taking chemistry, and let A be the event of Bev scoring an A in any class. Then, the event of interest is $A \cap C$.

$$P(A \cap C) = P(C)P(A \mid C) = \frac{1}{2} \cdot \frac{1}{3} = \frac{1}{6}$$

Note that the information about the computer class is absolutely irrelevant. $P(C) = 1/2$ came strictly from the fair coin flip.

Example 6.3 (Urn problem revisited)**.** We have 11 balls in an urn, 6 white and 5 black. What's the probability of drawing two blacks in two draws without replacement?

Solution. Let B_i be the event of drawing a black ball at the ith draw, for $i = 1, 2, 3$. Then, we are interested in the following probability

$$P(B_1 \cap B_2) = P(B_2 \mid B_1)P(B_1) = \frac{4}{10} \cdot \frac{5}{11} = \frac{2}{11}.$$

One can verify using combinations:

$$P(\text{two blacks in two draws}) = \frac{\binom{5}{2}}{\binom{11}{2}} = \frac{2}{11},$$

or counting principles

$$P(\text{two blacks in two draws}) = \frac{5 \cdot 4}{11 \cdot 10} = \frac{2}{11}.$$

Note that the formulaic procedure of the counting principle and the conditional probability argument look identical.

Example 6.4 (Hat checking)**.** An old tradition at gatherings is to hat check at the entrance of venue. Mix-ups do happen when people retrieve the hats at the end of the event. Three men toss their hats to the busboy, who is unfortunately overwhelmed and didn't record whose hat is whose. What's the probability of that none get their own hat back when they check out?

Solution. Let E_j be the event of the jth person getting his hat back. We are interested in $P(E_1^{\mathrm{c}} \cap E_3^{\mathrm{c}} \cap E_3^{\mathrm{c}})$. By De Morgan's law,

$$P(E_1^{\mathrm{c}} \cap E_3^{\mathrm{c}} \cap E_3^{\mathrm{c}}) = P((E_1 \cup E_2 \cup E_3)^{\mathrm{c}}) = 1 - P(E_1 \cup E_2 \cup E_3).$$

By the inclusion-exclusion formula you proved in Discussion 1, the probability of the union of three sets requires the probability of individual events, their pairwise intersections, and lastly the joint intersection.

First, note that $P(E_j) = 1/3$ since it's just the single person trying to select his own hat out of 3 in the pile. Meanwhile, for $i \neq j$, we use conditional probability

$$P(E_i \cap E_j) = P(E_i \mid E_j)P(E_j) = \frac{1}{2} \cdot \frac{1}{3}$$

where the reason for $P(E_i \mid E_j) = 1/2$ is that given someone who has selected correctly his hat, the next person will then have an improved chance of $1/2$ to select his hat from now a reduced pile of 2 hats.

Lastly, to study the joint intersection, we use the conditional probability formula again by treating $E_2 \cap E_3$ as a single event,

$$P(E_1 \cap E_2 \cap E_3) = P(E_1 \mid E_2 \cap E_3)P(E_2 \cap E_3) = 1 \cdot \frac{1}{6}$$

where the first probability is just describing the chance that the other two person got their own hats, which means the last person is guaranteed his own hat.

Altogether,

$$\begin{aligned} P\left(E_1 \cup E_2 \cup E_3\right) &= P\left(E_1\right) + P\left(E_2\right) + P\left(E_3\right) - P\left(E_1 \cap E_2\right) - P\left(E_1 \cap E_3\right) - P\left(E_2 \cap E_3\right) \\ &\quad + P\left(E_1 \cap E_2 \cap E_3\right) \\ &= \frac{1}{3} + \frac{1}{3} + \frac{1}{3} - \frac{1}{6} - \frac{1}{6} - \frac{1}{6} + \frac{1}{6} \\ &= 1 - \frac{1}{2} + \frac{1}{6} \\ &= \frac{2}{3} \end{aligned}$$

and thus

$$P\left(E_1^{\mathrm{c}} \cap E_2^{\mathrm{c}} \cap E_2^{\mathrm{c}}\right) = 1 - P\left(E_1 \cup E_2 \cup E_3\right) = 1 - \frac{2}{3} = \frac{1}{3}.$$

Example 6.5 (Extra Credit HW)**.** Generalize the above problem to N hats.

6.2 Law of Total Probability

The multiplication rule can also be visualized in a tree. See Fig 1.11 in your textbook on Page 33. We here use an example in another context that helps you see the chain of reasoning via the multiplication rule.

One important implication of conditional probability and its formula is that it allows a different perspective to calculate $P(E)$ for any event E. Consider a collection of mutually exclusive events A_i, $i = 1, 2, \ldots, N$ such that $\bigcup_{i=1}^{N} A_i = \Omega$. Then by Axiom 3,

$$P(E) = P(E \cap A_1) + P(E \cap A_2) + \cdots + P(E \cap A_N) = \sum_{i=1}^{N} P(E \cap A_i)$$

Based on this result, we use conditional probability to further express the probability of intersections, and arrive at the following Theorem.

Theorem 6.2 (Law of Total Probability)**.** *Let $\{A_i\}_{i=1}^{N}$ be a collection of mutually exclusive events such that $\bigcup_{i=1}^{N} A_i = \Omega$ (exhaustive property). Then*

$$P(E) = \sum_{i=1}^{N} P(E \cap A_i) = \sum_{i=1}^{N} P(E \mid A_i) P(A_i) \tag{6.2}$$

In other words, if you want to know the probability of something happening, you can break it down according to different "scenarios" that cover all possibilities. Then you add up the probabilities of the event happening in each scenario, weighted by how likely that scenario is.

7 Lecture VII: Bayes' Theorem and Its Applications

7.1 Baby Bayes' Theorem

The Law of Total Probability, for two events A and B, can help us illustrate their probability based on mutual observations in yet another important theorem below.

Theorem 7.1 (Baby Bayes' Theorem)**.** *For events A and B with nonzero probabilities, we have*

$$P(A \mid B) \overset{\dagger}{=} \frac{P(B \mid A)\,P(A)}{P(B)} \overset{*}{=} \frac{P(B \mid A)\,P(A)}{P(B \mid A)\,P(A) + P(B \mid A^{\mathrm{c}})\,P(A^{\mathrm{c}})}.$$

In general, for a mutually exclusive collection of events A_i, for $i = 1, 2, \ldots, N$, we have

$$P(A_i \mid B) = \frac{P(B \mid A_i)\,P(A_i)}{\sum_{i=1}^{N} P(B \mid A_i)\,P(A_i)}. \tag{7.1}$$

Proof. To prove the equality †, we invoke the definition of conditional probability

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)},$$
$$P(B \mid A) = \frac{P(B \cap A)}{P(A)},$$

Since intersections are symmetric, the numerators are the same probabilities, and therefore, we can solve for them and equate the rest. More precisely,

$$P(A \mid B)P(B) = P(B \mid A)P(A) \implies P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)}.$$

To prove the equality $*$, simply use the Law of Total Probability for the denominator B with two mutually exclusive events A and A'.

We leave the proof for the general case (Equation (7.1)) as an exercise. □

We demonstrate the power of this law in a realistic scenario involving medical tests.

Example 7.1 (False positives)**.** A medical test is designed to detect a certain disease. The disease affects 2% of the population. The test is not perfect:

- If a person has the disease, the test returns positive 95% of the time.

- If a person does not have the disease, the test returns negative 97% of the time.

1. If a randomly chosen person tests positive, what is the probability that they actually have the disease? What's the probability of a false positive?

2. If a randomly chosen person tests negative, what is the probability that they are disease-free? What's the probability of a false negative?

Solution. Let D denote the event "person has the disease" and $\overline{D}$ its complement. Let $+$ denote a positive test and $-$ a negative test. Given:

$$P(D) = 0.02, \qquad P(\overline{D}) = 0.98,$$

$$P(+ \mid D) = 0.95, \qquad P(- \mid \overline{D}) = 0.97.$$

1. Probability the person has the disease given a positive test (true positive). By definition of conditional probability and the Law of Total Probability,

$$P(D \mid +) = \frac{P(+ \mid D)\, P(D)}{P(+)} = \frac{P(+ \mid D)\, P(D)}{P(+ \mid D)P(D) + P(+ \mid \overline{D})P(\overline{D})}. \tag{7.2}$$

Compute numerator and denominator explicitly:

$$P(+ \mid D)P(D) = 0.95 \times 0.02 = 0.019,$$

$$\boxed{P(+ \mid \overline{D})}P(\overline{D}) = 0.03 \times 0.98 = 0.0294,$$

where we use the complementation rule for conditional probability for the boxed probability (why? See footnote).

So

$$P(+) = 0.019 + 0.0294 = 0.0484.$$

Thus

$$P(D \mid +) = \frac{0.019}{0.0484} = \frac{95}{242} \approx 39.26\%.$$

The definition of a false positive is the event that the person is disease-free AND the test is positive, which means

$$P(\overline{D} \cap +) = P(+ \mid \overline{D})P(\overline{D}) = 0.0294 = 2.94\%,$$

a calculation we have done above.

2. Probability the person is disease-free given a negative test (true negative). We want $P(\overline{D} \mid -)$. Use complement relations. First compute [1]

$$P(- \mid \overline{D}) \overset{\dagger}{=} 1 - P(+ \mid \overline{D}) = 1 - 0.03 = 0.97,$$

$$P(- \mid D) \overset{\dagger}{=} 1 - P(+ \mid D) = 1 - 0.95 = 0.05.$$

Then

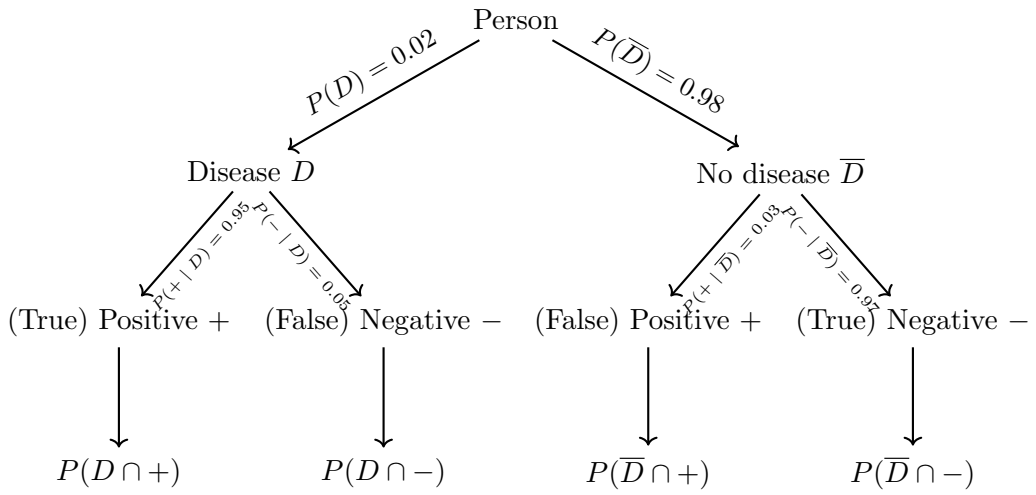
$$P(\overline{D} \mid -) = \frac{P(- \mid \overline{D})P(\overline{D})}{P(-)} = \frac{0.97 \times 0.98}{0.97 \times 0.98 + 0.05 \times 0.02} \approx 99.8949\%. \tag{7.3}$$

The definition of a false negative is the event that the person has the disease AND the test is negative, which means

$$P(D \cap -) = P(- \mid D)P(D) = 0.97 \cdot 0.02 = 0.0194 = 1.94\%,$$

a calculation we have done above.

Remark 7.1. *Although the test has high sensitivity (95%) and a fairly low false positive rate (3%), the disease's low prevalence (2%) makes the positive predictive value moderate (about 39%). Conversely, a negative test is extremely reassuring here (about 99.9% chance of being disease-free).*



[1] †

Question 7.1. Why can we just use the complementation rule here, i.e., subtract from 1? HW2 – prove that conditional probability is also a probability.

7.2 Generalized Bayes' Theorem

We saw just now a simple application of the Law of Total Probability, rather than conditioning on a large collection of mutually disjoint events $\{B_i\}_{i=1}^N$, is to just consider B and B', as exercised in the last example. So, a baby version of the Law of Total Probability can be written as

$$P(A) = P(A \mid B)P(B) + P(A \mid B')P(B').$$

Noticing from Equations (7.2) and (7.3), we have invoked exactly this baby version of the Law of Total Probability to compute the probability in the denominator. In fact, such a nice formula for the conditional probability $P(A_i \mid B)$ formulates the groundwork for a much bigger result, the Bayes' Theorem, which, in turn, gives rise to the meat-and-potato subject in statistical inference, known as Bayesian Inference. It generalizes from two disjoint events B and B' to a collection of disjoint events.

Theorem 7.2 (Bayes' Theorem)**.** *Let $A_1, \ldots, A_k$ be a collection of mutually disjoint, with $P(A_i) > 0$ for $i = 1, \ldots, k$ and $\sum P(A_i) = 1$ (exhaustive). Then for* ***any other event*** *B for which $P(B) > 0$, we have*

$$P\left(A_i \mid B\right) = \frac{P\left(A_i \cap B\right)}{P\left(B\right)} = \frac{P\left(B \mid A_i\right) P\left(A_i\right)}{\sum_{i=1}^{k} P\left(B \mid A_i\right) P\left(A_i\right)}. \tag{7.4}$$

This formula facilitates an analysis of experiments that have more than two outcomes (last example has two experiments, each with two outcomes), which means, we can grow that tree with as many branches as possible, and this formula can still tell us the conditional probability of a particular outcome given a fixed observation B. In other words, you do not need to fear a complicated problem, as long as you can distinguish the experiments and identify their possible outcomes, and leave the rest to this power formula that solves it all.

8 Lecture VIII: Independence

8.1 Definition and Example for Independence of Two Events

Definition 8.1 (Independence)**.** *Two events are said to be* ***independent*** *if* $P(A \mid B) = P(A)$*, and are otherwise* ***dependent.***

Remark 8.1. *This definition makes perfect sense because knowing* B *does not change the probability of* A*.*

Proposition 8.1. A *and* B *are* ***independent*** *if and only if*

$$P(A \cap B) = P(A)P(B).$$

Proof. By the Multiplication Rule, $P(A \cap B) = P(A \mid B) \cdot P(B) = P(A) \cdot P(B)$ if and only if $P(A \mid B) = P(A)$. □

Example 8.1 (Two Fair Dice (from Ross, Probability Model))**.** Suppose we toss two fair dice. Let E_1 denote the event that the sum of the dice is 6, and F denote the event that the first die equals 4. Then

$$P(E_1 \cap F) = P(\{4, 2\}) = \frac{1}{36}$$

while

$$P(E_1) \cdot P(F) = \frac{5}{36} \cdot \frac{1}{6} = \frac{5}{216}$$

and hence E_1 and F are not independent.

Intuitively, the reason for this is clear for if we are interested in the possibility of throwing a six (with two dice), then we will be quite happy if the first die lands four (or any of the numbers 1, 2, 3, 4, 5) because then we still have a possibility of getting a total of six. On the other hand, if the first die landed six, then we would be unhappy as we would no longer have a chance of getting a total of six. In other words, our chance of getting a total of six depends on the outcome of the first die and hence E_1 and F cannot be independent.

On the other hand, consider E_2 as the event that the sum of the dice equals 7. Is E_2 independent of F. We check

$$P(E_2 \cap F) = P(\{4, 3\}) = \frac{1}{36}$$

while

$$P(E_2) \cdot P(F) = \frac{1}{6}\frac{1}{6} = \frac{1}{36}.$$

What happened here? What would be an intuitive argument why the event that the sum of the dice equals seven is **independent** of the outcome on the first die? HW3.

8.2 Independence of More than Two Events

The notion of independence can be extended to more than two events. To distinguish the case with two and more than two, we refer to the former as **pairwise independence**. More precisely, a collection of events $A_1, \ldots, A_n$ is called **pairwise independence** if and only if for all distinct pairs of indices i, j, we have

$$P(A_i \cap A_j) = P(A_i)P(A_j).$$

Now, we are ready to introduce independence of more than two events.

Definition 8.2 (Mutual Independence)**.** *Events $A_1, \ldots, A_n$ are **mutually independent** if for every k ($k = 2, 3, \ldots, n$) and every subset of indices $1 \leq i_1 \leq i_2 \leq \ldots \leq i_k \leq n$,*

$$P(A_{i_1} \cap A_{i_2} \cap \ldots \cap A_{i_k}) = P(A_{i_1})P(A_{i_2}) \ldots P(A_{i_k}),$$

or in shorter notation (same meaning)

$$P\left(\bigcap_{j=1}^{k} A_{i_j}\right) = \prod_{j=1}^{k} P(A_{i_j})$$

In other words, the events are mutually independent if the probability of the intersection of any subset of the n events is equal to the product of the individual probabilities.

Example 8.2 (Pairwise Independent but not Mutually Independent)**.** Consider the experiment of drawing from an urn with four balls, labeled $\{1, 2, 3, 4\}$. Let $E = \{1 \text{ or } 2\}$, $F = \{1 \text{ or } 3\}$ and $G = \{1 \text{ or } 4\}$. Assuming that all four outcomes are equally likely, analyze the independence structure of the following events.

Solution. First, are these events pairwise independent? Yes, since each draw is independent from each other.

$$P(E \cap F) = P(E)\,P(F) = \frac{1}{4}$$
$$P(E \cap G) = P(E)\,P(G) = \frac{1}{4}$$
$$P(F \cap G) = P(F)\,P(G) = \frac{1}{4}$$

However,

$$P(E \cap F \cap G) = P(\{1\}) = \frac{1}{4} \neq P(E)P(F)P(G).$$

8.3 A First Look of Bernoulli Trials

A Bernoulli trial is a random experiment with exactly two possible outcomes, conventionally called "success" and "failure", where the probability of success remains constant for every repetition.

Example 8.3 (Roll a die)**.** What's the probability that we get at least one 6 in five rolls?

Solution. Call the event of interest E. The Bernoulli trial here is simply, for each roll, we get a 6 (success) or not (failure).

$$P(E) = 1 - P(E') = 1 - \left(\frac{5}{6}\right)^5 = 1 - \frac{3125}{7776} \approx 59.8\%,$$

where we use the complement event E' that we get no 6 in five rolls.

Read Example 1.39 (page 46) in the Carlton Textbook.

Example 8.4 (Electric Circuit in Parallel)**.** Circuits in parallel have one advantage over those in series – if one switch breaks down, the whole system still operates. Consider an electric network that has 20 parallel switches. The probability that one of them is turned on is 20%, independent of the status of all other switches. What's the probability that the whole system is working?

Solution. Let S_n denote the event that the nth switch is turned on, and F be the event that the whole system is working.

The key observation here is that the system fails to work if ALL switches are turned off. Therefore,

$$\begin{aligned} P(F) &= 1 - P(F') \\ &= 1 - P\left(\bigcap_{i=1}^{20} S_n'\right) \\ &= 1 - \prod_{i=1}^{20} P\left(S_n'\right) \\ &= 1 - (0.8)^{20} \\ &\approx 89.26\%. \end{aligned}$$

Part II
Discrete Random Variables

9 Lecture IX: Discrete Random Variables and Their Distributions

So far, we have defined probability through the three Axioms. We have learned how to count the number of outcomes of an experiment – both simple and complex – in systematic ways. We have also dissected the sample space using the Law of Total Probability and introduced conditional probability to model the act of observation in everyday situations, supported by some mathematical groundwork. Finally, we began with an intuitive definition of independence and then formalized it using conditional probability to capture that intuition.

However, in most experiments, we rarely describe outcomes with full sentences. Instead, we often use symbols or shorthand to avoid long-winded descriptions. Moreover, we are usually less interested in the outcome itself than in **some numerical quantity** associated with the outcome. For this reason, it is convenient to introduce the following concept.

9.1 Random Variable

Definition 9.1 (Random Variable)**.** *A function $X : \Omega \to \mathbb{R}$ is called a* ***random variable****.*

Remark 9.1. *It is a common misconception that a variable is by default, the input of some function. Here, the random variable is a* ***dependent variable****, i.e., the y in $y = f(x)$. It takes some outcome from the sample space as input, and spits out a real number.*

Example 9.1 (Coin tosses)**.** We are tossing three fair coins. Denote by Y the number of heads. Thus Y can achieve the values 0, 1, 2, and 3. Our Laplacian sample space is

$$\Omega = \{(T,T,T),(H,T,T),(T,H,T),(T,T,H),(H,H,T),(T,H,H),(H,T,H),(H,H,H)\}.$$

For each outcome, Y counts the number of heads in it, and returns a real number (an integer in this case). For example, $Y(\{T,T,T\}) = 0$, while $Y(\{H,H,T\}) = 2$.

We can then frame the following questions into a statement about the random variable Y:

1. What's the probability that three heads are flipped?

$$P(Y=3) = P(\{(T,T,T)\}) = \frac{1}{8}$$

2. What's the probability that we have at least two heads?

$$\begin{aligned} P(Y \geq 2) &= P(Y=2) + P(Y=3) \\ &= P(\{(H,T,T),(T,H,T),(T,T,H)\}) + P(\{(T,T,T)\}) \\ &= \frac{3}{8} + \frac{1}{8} \\ &= \frac{1}{2} \end{aligned}$$

Important Remark 9.1. *We have used (and will use) the shorthand*

$$P(Y=j) = P(\{Y=j\}) = P(\{\omega : Y(\omega) = j\})$$

where the last quantity reads

the probability of the ***events*** *such that the random variable Y equals* 3.

Note that it is much easier to work with Y, rather than listing all outcomes.

Naturally, the random variable must obey the basic probability axioms, that is, the sum of probabilities of simple events must be 1, i.e.,

$$\sum_{j=0}^{3} P(Y=j) \overset{\text{Axiom 3}}{=} P\left(\bigcup_{j=0}^{3} \{Y=j\}\right) = P(\Omega) \overset{\text{Axiom 2}}{=} 1$$

9.2 Discrete Random Variables

It seems the quantity $P(Y=j)$ is rather important. Before we give it a proper name, we first define discrete random variables.

Definition 9.2 (Discrete Random Variables)**.** *A random variable X is called discrete, if it takes at most countably many values x_1, x_2, …, x_j, …. Its* ***probability mass function (pmf*** *in short) is defined by*

$$p(x) = p_X(x)^2 = P(X=x)$$

[2]The subscript X is to distinguish between random variables if more than one are analyzed.

that satisfies

$$\begin{aligned} p(x_j) &\geq 0, \\ p(x) &= 0, \ \textit{for all } x \neq x_j, \\ \sum_{k=1}^{\infty} p(x_k) = \sum_{k=1}^{\infty} P(X = x_k) &= P\left(\bigcup_{k=1}^{\infty} \{X = x_k\}\right) = P(\Omega) = 1. \end{aligned}$$

Remark 9.2. *Why is this called "discrete"? Because the possible values X can take is countable, namely, you can always separate two values with finitely many other values.*

Remark. *We will learn about the* ***continuous*** *analog of this definition after the midterm.*

Example 9.2 (PMF for Number of Heads)**.** Consider the same setting in the previous example. The random variable X counts the number of heads for tossing a fair coin 3 times. We enumerate its possible values, i.e., $\{0, 1, 2, 3\}$ on the x-axis, and their corresponding probabilities on the y-axis. See Figure 1.

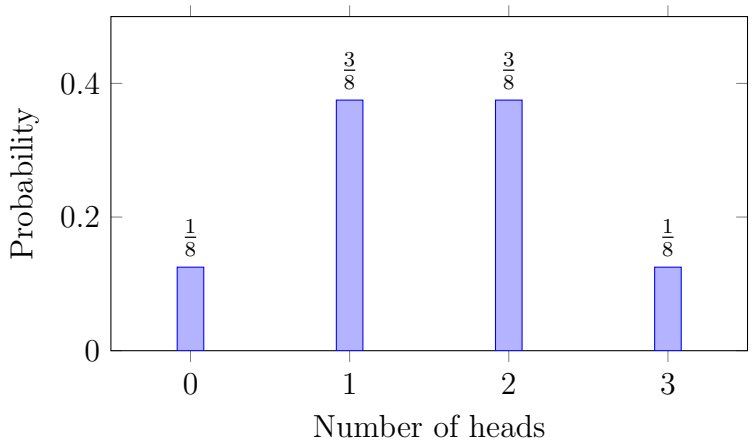


Figure 1: Probability Mass Function of the Random Variable X that counts the number of heads for tossing a coin three times.

Example 9.3 (Sum of Two Fair Six-sided Dice Roll)**.** We can revisit the random variable Y that computes the sum of two fair six-sided dice roll, see Figure 2.

9.3 Cumulative Distribution Function

While the pmf is particularly useful to pinpoint probabilities that a discrete random variable takes on an exact value, it falls short to probabilities such as $P(X \leq 10)$,

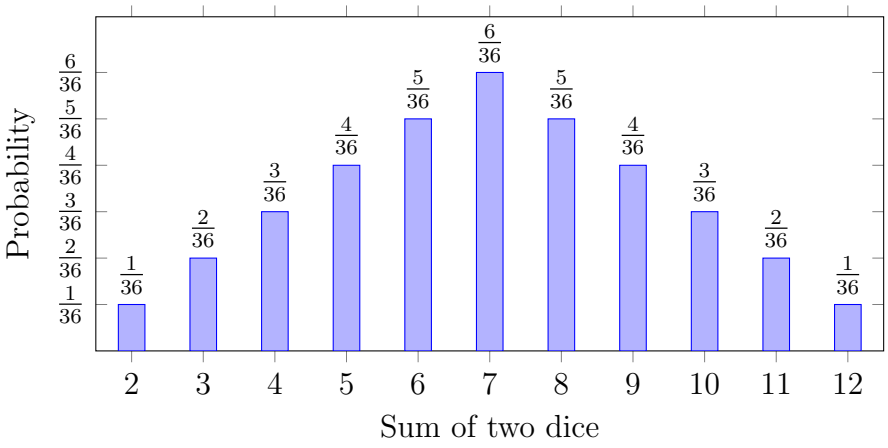


Figure 2: Probability Mass Function of the Random Variable X that computes the sum of two fair six-sided dice.

because this probability requires that you consider all possible values for X from 3 and up, which amounts to a long calculation such as $P(X = 0) + P(X = 1) + P(X = 2) + \cdots + P(X = 10)$.

Is there a systematic way to compute probabilities such as $P(X \leq x)$, based on a given pmf?

Definition 9.3 (Cumulative Distribution Function (CDF))**.**

The ***cumulative distribution function*** *(cdf in short) of a discrete random variable X with pmf $p(x)$ is defined by every number x by*

$$F(x) = P(X \leq x) = \sum_{y:y \leq x} p(y) \tag{9.1}$$

For any number x, $F(x)$ is the probability that the observed value of X will be ***at most*** *x.*

Remark. *We sometimes write $F_X(x)$, with a subscript, in case there are other random variables in the context.*

Remark. *A more detailed way to look at $F(x)$ is*

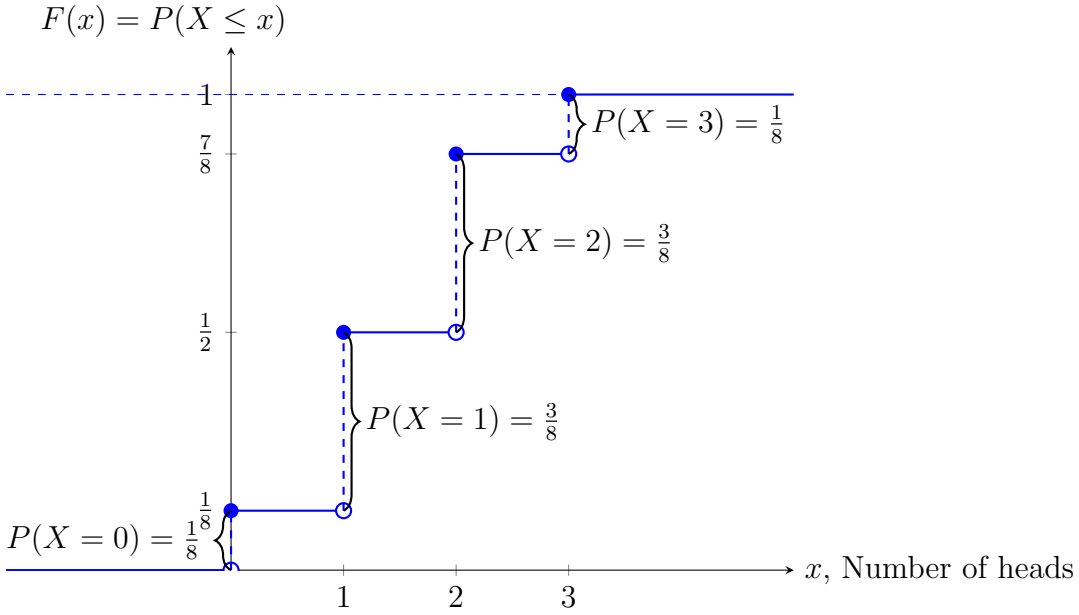
$$F(x) = \sum_{y:y \leq x} p(y) = P\left[\bigcup_{y \leq x} \{X = y\}\right] = P(X \leq x)$$

Example 9.4 (CDF for Number of Heads)**.** Based on the pmf in Example 9.2 and its graph Figure 1, we can compute the CDF for the random variable X that counts

the number of heads for tossing a coin three times.

$$F(x) = \begin{cases} 0, & \text{if } x < 0; \\ \frac{1}{8}, & \text{if } 0 \le x < 1; \\ \frac{1}{2}, & \text{if } 1 \le x < 2; \\ \frac{7}{8}, & \text{if } 2 \le x < 3; \\ 1, & \text{if } x \ge 3. \end{cases}$$

The CDF looks like the following



Important Remark 9.2. *You must note that the CDF is a piecewise function for this discrete random variable, and is in fact only continuous from the left. You can observe that the function only takes value at the* ***left*** *endpoints, and jumps to the next step. This gives rise to the following proposition about the CDF.*

Proposition 9.1 (Properties of the CDF)**.** *If F is a cdf, then it holds that*

1. *F is non-decreasing on $\mathbb{R}$.*
2. *F is right-continuous, i.e. $\lim_{x \to y^+} = F(y)$ for all $y \in \mathbb{R}$.*
3. *$\lim_{x \to -\infty} F(x) = 0$ and $\lim_{x \to \infty} F(x) = 1$.*

Remark 9.3 (Sketch of proof for the non-decreasing property.)**.** *Axiom 1, i.e., the probability of any events is non-negative. Why? Think about it.*

Remark 9.4 (Optional sketch of proof for right-continuity)**.** *The CDF is the probability of the event $\{X \leq x\}$. If you nudge x just a tiny bit to the right, you are not excluding any outcomes you had before, you're only possibly including new ones. That's why the function always settles at the right-hand endpoint.*

Example 9.5 (CDF of the Sum of Two Dice)**.** Let X denote the sum of the outcomes when two fair six-sided dice are rolled. We provide a systematic way to generate the CDF of this random variable.

1. PMF. The PMF is already sketched in Figure 2.

2. CDF values. The cumulative distribution function $F(x) = P(X \leq x)$ is obtained by summing the PMF up to each point. For example,

$$F(2) = \tfrac{1}{36}, \quad F(3) = \tfrac{3}{36}, \quad F(4) = \tfrac{6}{36}, \quad F(5) = \tfrac{10}{36}, \quad \ldots, \quad F(12) = 1.$$

3. Sketching the CDF. To sketch $F(x)$, it suffices to compute the nodal values $F(2), F(3), \ldots, F(12)$, since the rest of the graph consists of flat horizontal segments between these values. The CDF is right-continuous: each jump is drawn with a hollow dot at the left end and a solid dot at the right end (for a fixed x value).

Let's use this figure to compute/verify a few things.

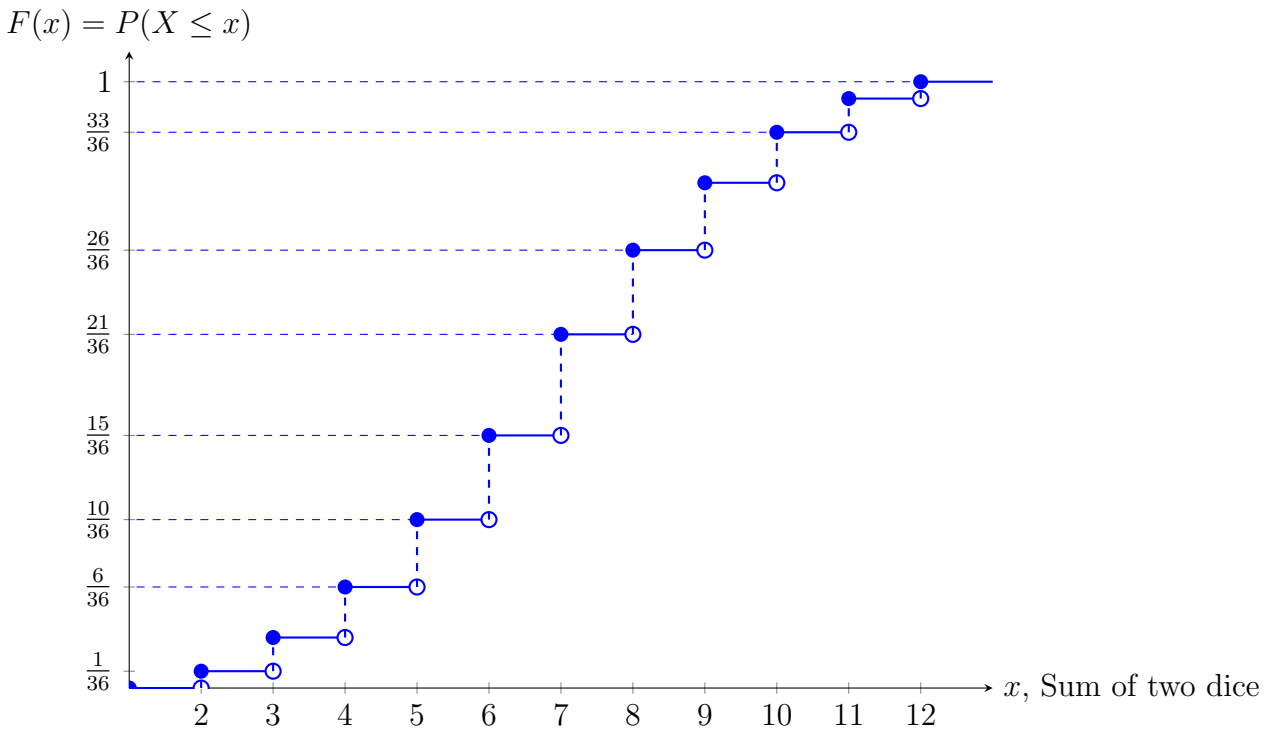
We know the probability of obtaining a 7 as a sum is 1/6. How can we tell from the graph of CDF? From the last example about the number of heads in three tosses, we learned that the size of the jump is precisely the probability of the event that leads to the jump!

$$P\left(X = 7\right) = F\left(X = 7\right) - F\left(X = 6\right) = \frac{21}{36} - \frac{15}{36} = \frac{1}{6}.$$

But this is not the strength of the CDF. What if now the question is to find the probability that the sum is between 5 and 8? We actually use Axiom 3, which remains useful up to this point.

$$\begin{aligned} P\left(5 \leq X \leq 8\right) &\overset{\text{Axiom 3}}{=} P\left(X \leq 8\right) - P\left(X < 5\right) \\ &\overset{\dagger}{=} P\left(X \leq 8\right) - P\left(X \leq 4\right) \\ &\overset{\text{Definition of CDF}}{=} F\left(8\right) - F\left(4\right) \\ &= \frac{26}{36} - \frac{6}{36} \\ &= \frac{5}{9} \end{aligned}$$

where the step at † is because of the fact that X has a range of integers.



From the sketching experience and a few example application of the graph of CDF, we seem to learn a few things about how to compute probabilities using the CDF (as opposed to the PMF). Here is a collection of rules to use the CDF.

Proposition 9.2. *For any two numbers a and b with $a \leq b$,*

$$P(a \leq x \leq b) = F(b) - F\left(a^{-}\right)$$

where a^{-} represents the largest possible X value that is strictly less than a. In particular, if the only possible values are integers and if a and b are integers, then

$$P(a \leq x \leq b) = P(X = a, \text{ or } a+1, \text{ or } \ldots, \text{ or } b) = F(b) - F(a-1) \tag{9.2}$$

In particular,

$$P(X = a) = F(a) - F(a-1)$$

by taking $a = b$ in Equation (9.2).

Question 9.1. When was the last time that you saw a quantity over a range $[a, b]$ is expressed as a difference $F(b) - F(a)$? Do you remember? Keep it a secret.

10 Lecture X: Expectation and Variance of A Random Variable

Now that we have explored the fundamental probabilistic building block of a random variable via its pmf and cdf, we can begin to ask further questions such as "what is the average roll of a die?", or something more complicated, such as "On average in how many years does the Boston Red Sox get into the playoffs?".

One deals with the notion of averages every day. The bus may not be exactly on time, so one quantifies it with an average wait time, so that we may or may not have a reason to yell at the system. The exam average score is one assessment of how the class performed. The amount of chicken fingers the canteen prepares is based on an average estimate of the number of students, which certainly changes from day to day, at different seasons of the year.

Once we learn or estimate an average, we **expect** that the next occurrence would be close to the average. But occurrences of events are not necessarily of equal probability, unless we have a Laplacian sample space. So, how do we quantify average more fairly, so that the probability of individual event is captured in the process?

10.1 Expected Value

Definition 10.1 (Expected Value/Expectation/Mean Value)**.**

*Let X be a discrete r.v. with set of possible values D and pmf $p(x)$. The **expectation**, or **expected value** or **mean value** of X, denoted by $E(X)$, or μ_X, or just μ, is*

$$E(X) = \mu_X = \mu = \sum_{x \in D} x \cdot p(x).$$

Example 10.1 (Average of a Die Roll)**.** If one asks you, what is the average of the integers $\{1, 2, 3, 4, 5, 6\}$, you would compute

$$\text{average} = \frac{1+2+3+4+5+6}{6} = 3.5$$

and we would be correct. But, this is just adding things up and divided by the number of them.

In the experiment of rolling a fair die where we denote X as the value of the roll, the expected value of one roll is also 3.5, but its calculation is more precisely,

$$\begin{aligned} E(X) &= \sum_{x \in \{1,2,3,4,5,6\}} x \cdot p(x) \\ &= 1 \cdot \frac{1}{6} + 2 \cdot \frac{1}{6} + 3 \cdot \frac{1}{6} + 4 \cdot \frac{1}{6} + 5 \cdot \frac{1}{6} + 6 \cdot \frac{1}{6} \end{aligned}$$

$$
\begin{aligned}
&= \frac{1+2+3+4+5+6}{6} \\
&= 3.5
\end{aligned}
$$

which marks a crucial difference in the procedure.

In fact, this procedure can help calculate the expected value of a skewed die, say, with pmf

$$p(1) = \frac{1}{4}, \quad p(2) = \frac{1}{4}, \quad p(3) = \frac{1}{6}, \quad p(4) = \frac{1}{6}, \quad p(5) = \frac{1}{12}, \quad p(6) = \frac{1}{12}$$

In this case, the expected value should be smaller than 3.5 because more probabilities are assigned to smaller values. Indeed,

$$
\begin{aligned}
E(X) &= \sum_{x \in \{1,2,3,4,5,6\}} x \cdot p(x) \\
&= 1 \cdot \frac{1}{4} + 2 \cdot \frac{1}{4} + 3 \cdot \frac{1}{6} + 4 \cdot \frac{1}{6} + 5 \cdot \frac{1}{12} + 6 \cdot \frac{1}{12} \\
&= \frac{1+2}{4} + \frac{3+4}{6} + \frac{5+6}{12} \\
&= \frac{3}{4} + \frac{7}{6} + \frac{11}{12} \\
&= \frac{34}{12} \\
&\approx 2.83
\end{aligned}
$$

Important Remark 10.1. *Even though we use the word "average" and "expected value" interchangeably in this course, it is important to note the difference. Average typically involves a simple procedure that finds the sum of numbers and divides it by the total count, assuming that each outcome is equally likely, while the expected value is computed by* ***taking the probability distribution into account****, which is not necessarily spread evenly for each outcome in the sample space.*

10.2 Expected Value of A Function of a Random Variable

Often we will be interested in the expected value of some function $h(X)$ rather than X itself. What would be the formula then?

Example 10.2 (Drawing spheres with random radius)**.** Suppose an urn has three spheres with radius $a < b < c$, respectively. What's the expected volume from one draw?

Solution. The function dependence here is $V(R) = \frac{4}{3}\pi R^3$ where R is a discrete random variable that takes value a, b or c.

One can approach the problem by brute force by listing the volume of each sphere

$$\left\{\frac{4}{3}\pi a^3, \frac{4}{3}\pi b^3, \frac{4}{3}\pi c^3\right\}.$$

Since drawing each sphere is equally likely, the expected volume from one draw is

$$E[V(R)] = \frac{4}{3}\pi a^3 \cdot \frac{1}{3} + \frac{4}{3}\pi b^3 \cdot \frac{1}{3} + \frac{4}{3}\pi c^3 \cdot \frac{1}{3} = \frac{4}{9}\pi\left(a^3 + b^3 + c^3\right).$$

Notice that this is precisely the same thing as

$$E[V(R)] = \sum_{r\in\{a,b,c\}} V(r) \cdot p(r)$$

where the pmf $p(r) = \frac{1}{3}$ for $r \in \{a, b, c\}$.

Proposition 10.1 (Expected Value of a Function of a Discrete Random Variable)**.** *If the discrete r.v. X has a set of possible values D and pmf $p(x)$, then the expected value of any function $h(X)$, denoted by $E[h(X)]$, is computed by*

$$E[h(X)] = \sum_{r\in D} h(x) \cdot p(x).$$

This is sometimes referred to as the ***Law of the Unconscious Statistician****.*

10.3 Linearity of Expectation

One should view taking expectation as a weighted sum of a bunch of numbers, where the weights are the probabilities. Since it is a sum for discrete random variables, it enjoys properties of a sum, such as additivity and the constant multiple.

Proposition 10.2 (Linearity of Expectation)**.** *For any functions $f(X)$ and $g(X)$, and any constants a_1, a_2 and b,*

$$E[a_1 f(X) + a_2 g(X) + b] = a_1 E[f(X)] + b_1 E[g(X)] + b.$$

In particular, for any linear function $aX + b$,

$$E[aX + b] = aE[X] + b.$$

Proof. By definition of expected value of a discrete random variable. Read Page 87 of your text, or come up with a proof yourself. □

10.4 Variance of a Random Variable

The exam average is only one of the measures of students' performance. It is in fact not a very reliable measure. An average of 50 can arise from everyone scoring exactly 50 points, or from half the class scoring 0's while the other half scoring 100's. From the instructor's standpoint, it is rather hard to judge the performance based on the average alone.

This brings us an important point about the expected value of observations of an experiment because there can be infinitely many collections of observations that give you exactly the same expected value, yet their individual properties may be significantly different. We then bring in the idea of **Variance**, a measure of how spread out the observed data is.

Definition 10.2 (Variance)**.** *Let X have pmf $p(x)$ and expected value μ. Then the* ***variance*** *of X, denoted by $\mathrm{Var}(x)$ or σ_X^2, or just σ^2, is*

$$\mathrm{Var}(X) = \sum_{x \in D} \left[(x-\mu)^2 \cdot p(x) \right] = E\left[(X-\mu)^2 \right].$$

The ***standard deviation*** *(SD) of X, denoted by $\mathrm{SD}(X)$ or σ_X, or just σ, is*

$$\sigma_X = \sqrt{\mathrm{Var}(X)}.$$

Important Remark 10.2. *Variance of a random variable is an example of the expected value of a function of a random variable. This function has the particular form*

$$f(X) = (X-\mu)^2.$$

For each value x of X, $f(x)$ is calculating the square distance between the value x and the expected value μ_X. Altogether, variance is calculating the average square distance between the observed values of X and μ_X. This means variance is always non-negative.

Example 10.3 (Variance of a Die Roll)**.** We know already that the expected value of a die roll is $\mu = 3.5$. Now, we compute the variance by definition.

$$\begin{aligned}
\mathrm{Var}(X) &= \sum_{x \in D} \left[(x-\mu)^2 \cdot p(x) \right] \\
&= (1-3.5)^2 \frac{1}{6} + (2-3.5)^2 \frac{1}{6} + (3-3.5)^2 \frac{1}{6} + (4-3.5)^2 \frac{1}{6} + (5-3.5)^2 \frac{1}{6} + (6-3.5)^2 \frac{1}{6} \\
&= \frac{35}{12} \\
&\approx 2.91
\end{aligned}$$

Now, if this number does not translate quickly to a physical meaning, then the **standard deviation** will

$$\sigma_X = \sqrt{\mathrm{Var}\,(X)} = \sqrt{\frac{35}{12}} \approx 1.708.$$

In words, this means each roll is on average at a distance of 1.708 from the mean.

In practice, however, this explicit calculation is very costly because it involves multiple additions, subtractions and multiplications. We here present an alternative where only one subtraction is used.

Proposition 10.3 (Variance Shortcut Formula)**.**

$$\mathrm{Var}\,(X) = E\left(X^2\right) - \mu^2 \tag{10.1}$$

Proof. By definition,

$$\begin{aligned}
\mathrm{Var}\,(X) &= E\left[(X-\mu)^2\right] \\
&= E\left[X^2 - 2\mu X + \mu^2\right] \\
&= E\left[X^2\right] - 2\mu E\left[X\right] + \mu^2 \\
&= E\left[X^2\right] - 2\mu^2 + \mu^2 \\
&= E\left[X^2\right] - \mu^2
\end{aligned}$$

by linearity of expectation. □

Remark. *This formula says, compute first, $E\,(X^2)$ by definition of the expected value of a function, and then simply subtract mean squared to obtain the variance. This formula obviates the multiple subtractions you saw in the last example.*

We use the formula to compute the variance again,

$$\begin{aligned}
E\left(X^2\right) &= \sum_{x\in D} x^2 P\,(X=x) \\
&= \left[1^2\cdot\frac{1}{6} + 2^2\cdot\frac{1}{6} + 3^2\cdot\frac{1}{6} + 4^2\cdot\frac{1}{6} + 5^2\cdot\frac{1}{6} + 6^2\cdot\frac{1}{6}\right] \\
&= \frac{91}{6}
\end{aligned}$$

and then

$$\mathrm{Var}\,(X) = E\left[X^2\right] - \mu^2 = \frac{91}{6} - \left(\frac{21}{6}\right)^2 = \frac{546}{36} - \frac{441}{36} = \frac{35}{12},$$

as expected.

Since variance is an expected value of a function of a random variable, does it enjoy the linearity property as well?

Proposition 10.4.

$$\mathrm{Var}\,(aX+b) = a^2\sigma_X^2, \quad \mathrm{SD}\,(aX+b) = |a|\,\sigma_X.$$

In particular,

$$\sigma_{aX} = |a|\,\sigma_X, \quad \sigma_{X+b} = \sigma_X.$$

Proof. Left as an exercise. □

Remark. *The last result $\sigma_{X+b} = \sigma_X$ must be understood intuitively. Adding a constant b to a bunch of numbers (possible values of X) will NOT change the spread of the numbers.*

11 Lecture XI: Examples of Discrete Random Variables

In this lecture, we give examples of discrete random variables. We have in fact studied some of them already without knowing their setups.

11.1 Bernoulli Random Variable

Definition 11.1 (Bernoulli Random Variables)**.** *Any random variable whose only possible values are* 0 *and* 1 *is called a* ***Bernoulli random variable****.*

The coin toss is a classical example of a Bernoulli random variable that consists of only two outcomes. One may treat heads as success and tails as failure. This kind of random variable is particularly useful when the sample space is a dichotomy, and often end up being useful in constructing other more complicated random variables.

Bernoulli random variables are often associated with a parameter p to represent the probability of success, and of course, the probability of failure is $1 - p$. More precisely, the pmf of a Bernoulli random variable satisfies

$$P(X = 1) = p, \quad P(X = 0) = 1 - p.$$

Proposition 11.1 (Expected value of a Bernoulli r.v.)**.** *Let X be a Bernoulli random variable with parameter p. Then*

$$E[X] = \sum_{x \in \{0,1\}} x \cdot P(X = x) = 1 \cdot p + 0 \cdot (1 - p) = p.$$

11.2 Geometric Random Variable

Building on the Bernoulli random variables which indicates success or failure of one trial, we are also interested in **the number of trials to achieve one success**. For example, consider the number of tosses of a fair die to see a 6. Since it depends on the success probability, this random variable is also associated with a parameter p, like the Bernoulli random variable.

Definition 11.2 (Geometric Random Variables)**.** *A geometric random variable counts the number of trials required to achieve one success, with success probability p. Its pmf satisfies*

$$p(n) = P(X = n) = (1 - p)^{n-1} p.$$

We write $X \sim Geo(p)$ as a shorthand. Note that X is a nonnegative integer, and has an unbounded range.

Remark. *This pmf is intuitive in the following way: in order to achieve success for the first time after n trials, the first $n-1$ trials must all be failures. With each trial being independent from each other, the probability of failing $n-1$ times is precisely $(1-p)^{n-1}$. Then, after $n-1$ failures, the next trial is a success with probability p.*

The following example not only showcases the way to use this pmf, but also motivates the calculation of the expected value.

Example 11.1. On average, how many tosses of a die do we need to see an even value?

Solution. Define X as the random variable that counts the number of tosses of a fair die needed to see an even value. Then $X \sim Geo\left(\frac{1}{2}\right)$. Therefore, by definition,

$$\begin{aligned}\mathbb{E}[X] &= \sum_{x=1}^{\infty} xP(X=x) \\ &= \sum_{x=1}^{\infty} x\,(1-p)^{x-1}\,p\end{aligned}$$

Hmmm, this seems to be a difficult sum. Game over?

No, we employ a nice result about geometric series (exactly where the name Geometric random variable comes from).

$$\begin{aligned}\sum_{x=1}^{\infty} x\,(1-p)^{x-1}\,p &= p\sum_{x=1}^{\infty} x\,(1-p)^{x-1} \\ &= p\left[-\frac{d}{dp}\sum_{x=0}^{\infty}(1-p)^{x}\right] \\ &= -p\left[\frac{d}{dp}\frac{1}{1-(1-p)}\right] \\ &= -p\left[-\frac{1}{p^2}\right] \\ &= \frac{1}{p}\end{aligned}$$

Using that $p = 1/2$, we have

$$\mathbb{E}[X] = 2,$$

that is, on average, we expect to see an even value after tossing a fair die twice. Intuitively, this makes sense also because even numbers make up half of the sample space, while each outcome is equally likely.

11.3 Binomial Random Variable

While the geometric random variable counts the number of trials required to achieve one success, the **Binomial random variable** counts the number of successes in a fixed amount of trials.

Definition 11.3 (Binomial random variable)**.** *Given a binomial experiment consisting of n trials, the **binomial random variable** X associated with this experiment is defined as*

$$X = \text{the number of successes among the } n \text{ trials}$$

with success probability p.

A Binomial Random Variable is often written as $X \sim Bin(n, p)$ to indicate that the r.v. depends on two parameters.

Example 11.2.

1. The number of heads in tossing a fair coin 5 times.
2. The number of cracked eggs you find in a carton, assuming each wreckage is independent from each other.
3. The number of customers who use cash in a checkout line.
4. The number of spam emails out of a fixed number of incoming emails.
5. The number of invitees who eventually show up to your party.

Notice that we don't care about which trials achieve the successes, but only the number of the successes.

Theorem 11.1 (PMF of A Binomial Random Variable)**.** *Let X be a binomial random variable with parameters (n, p). Then its pmf satisfies*

$$P(X = k) = b(k; n, p) = \binom{n}{k} p^k (1-p)^{n-k}, \quad k = 0, 1, 2, \ldots, n.$$

The notation $b(k; n, p)$ is to help indicate that it depends on k as an independent variable, but n and p as parameters (usually given as fixed inputs).

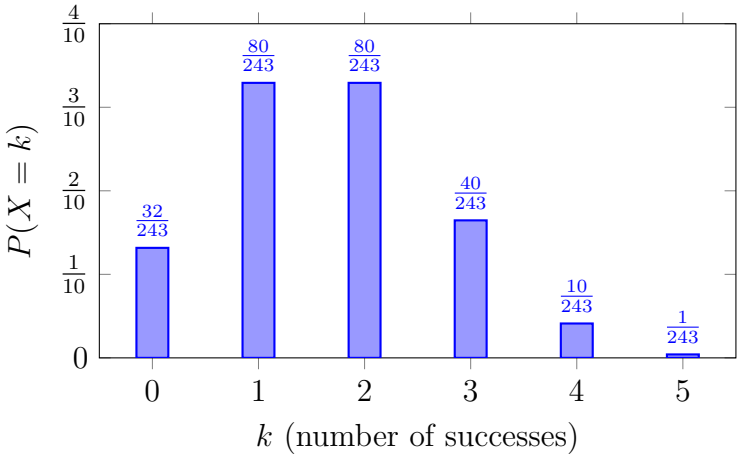
Question 11.1. Bonus HW: prove that

$$\sum_{k=0}^{n} b(k; n, p) = 1$$

that is, the function $b(k; n, p)$ is indeed a pmf.

Example 11.3. The probability of selecting a black ball from an urn is 1/3. Sketch the pmf of the random variable that counts the number of times a black ball is drawn with replacement in 5 draws.

Solution. The random variable is $X \sim Bin(5, 1/3)$.



The probability associated with each bar height are computed as follows:

$$P(X=0) = \binom{5}{0}\left(\frac{1}{3}\right)^{0}\left(1-\frac{1}{3}\right)^{5-0}$$
$$P(X=1) = \binom{5}{1}\left(\frac{1}{3}\right)^{1}\left(1-\frac{1}{3}\right)^{5-1}$$
$$P(X=2) = \binom{5}{2}\left(\frac{1}{3}\right)^{2}\left(1-\frac{1}{3}\right)^{5-2}$$
$$P(X=3) = \binom{5}{3}\left(\frac{1}{3}\right)^{3}\left(1-\frac{1}{3}\right)^{5-3}$$
$$P(X=4) = \binom{5}{4}\left(\frac{1}{3}\right)^{4}\left(1-\frac{1}{3}\right)^{5-4}$$
$$P(X=5) = \binom{5}{5}\left(\frac{1}{3}\right)^{5}\left(1-\frac{1}{3}\right)^{5-5}$$

Example 11.4. If we roll seven dice, how likely is it to see six 6's?

Solution. Let X be the number of 6's in 7 rolls. We identify that $X \sim Bin(7, 1/6)$. We are interested in

$$P(X=6) = \binom{7}{6}\left(\frac{1}{6}\right)^{6}\left(1-\frac{1}{6}\right)^{7-6} = 7\left(\frac{1}{6}\right)^{6}\left(\frac{5}{6}\right) = 7\left(\frac{1}{6}\right)^{6}\frac{5}{6} = \frac{35}{6^7} \approx 0.000125 = 0.0125\%$$

which is not very likely.

Example 11.5 (Playing Craps)**.** Consider the experiment of rolling two dice and observing the sum. After conducting the experiment 10 times, how likely is it to see a sum of 7 five times?

Solution. Let X be the number of 7's in 10 experiments. We identify that $X \sim Bin(10, 1/6)$. We are interested in

$$P(X = 5) = \binom{10}{5} \left(\frac{1}{6}\right)^5 \left(1 - \frac{1}{6}\right)^{10-5} = \binom{10}{5} \left(\frac{1}{6}\right)^5 \left(\frac{5}{6}\right)^5 \approx 1.3\%.$$

Proposition 11.2. *The mean of a binomial random variable $X \sim Bin(n, p)$ is np.*

Proof.

$$\begin{aligned}
E[X] &= \sum_{k=0}^{n} k \cdot P(X = k) \\
&= \sum_{k=0}^{n} k \cdot \binom{n}{k} p^k (1-p)^{n-k} \\
&= \sum_{k=0}^{n} k \cdot \frac{n!}{k!\,(n-k)!} p^k (1-p)^{n-k} \\
&= \sum_{k=0}^{n} \frac{n!}{(k-1)!\,(n-k)!} p^k (1-p)^{n-k} \\
&= \sum_{k=0}^{n} n \frac{(n-1)!}{(k-1)!\,(n-1-(k-1))!} p^k (1-p)^{n-k} \\
&= \sum_{k=0}^{n} np \binom{n-1}{k-1} p^{k-1} (1-p)^{n-k} \\
&= np \underbrace{\sum_{k=0}^{n} \binom{n-1}{k-1} p^{k-1} (1-p)^{n-k}}_{=1 \text{ since it sums the pmf of } Bin(n-1,p)} \\
&= np
\end{aligned}$$

□

Question 11.2. Bonus HW: using a similar technique as above, prove that

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

11.4 Poisson Random Variable

The number of bus arrivals in an hour, the number of customers arriving in a line in five minutes, the number of particles filtered by a sieve, are all examples of an arrival process, typically modeled by the Poisson Random Variable. It is often used as a good and simple approximation of a Binomial random variable when the number of trials n is large and the success probability p is small. These two parameters form a single parameter λ for the Poisson random variable, more precisely defined below.

Definition 11.4. *A random variable is said to have a Poisson distribution with parameter $\lambda > 0$ if the pmf of X is*

$$P(X = k) = p(k; \lambda) = \frac{e^{-\lambda}\lambda^k}{k!}, \quad k = 0, 1, 2, \ldots.$$

We verify that this function is indeed a pmf by computing

$$\sum_{k=0}^{\infty} P\left(X = k\right) = \sum_{k=0}^{\infty} \frac{e^{-\lambda}\lambda^k}{k!} = e^{-\lambda} \sum_{k=0}^{\infty} \frac{\lambda^k}{k!} = e^{-\lambda}e^{\lambda} = 1$$

where we used the Maclaurin series

$$e^x = \sum_{k=0}^{\infty} \frac{x^k}{k!}.$$

The following examples are taken from the textbook "Probability Models" by Sheldon Ross.

Example 11.6 (Typos)**.** Suppose that the number of typographical errors on a single page of this book has a Poisson distribution with parameter $\lambda = 1$. Calculate the probability that there is at least one error on this page.

Solution.

$$P(X \geq 1) = 1 - P(X = 0) = 1 - e^{-1} \approx 0.633.$$

Example 11.7 (Accidents)**.** If the number of accidents occurring on a highway each day is a Poisson random variable with parameter $\lambda = 3$, what is the probability that no accidents occur today?

Solution.

$$P(X \geq 0) = e^{-3} \approx 0.05.$$

Example 11.8 (Radioactive emissions)**.** Consider an experiment that consists of counting the number of α-particles given off in a one-second interval by one gram of radioactive material. If we know from past experience that, on the average, 3.2 such α-particles are given off, what is a good approximation to the probability that no more than two α-particles will appear?

Solution. If we think of the gram of radioactive material as consisting of a large number n of atoms each of which has probability $3.2/n$ of disintegrating and sending off an α-particle during the second considered, then we see that, to a very close approximation, the number of α-particles given off will be a Poisson random variable with parameter $\lambda = 3.2$. Hence the desired probability is

$$P(X \leq 2) = e^{-3.2} + 3.2e^{-3.2} + \frac{3.2^2}{2!}e^{-3.2} \approx 0.382.$$

Proposition 11.3. *Let X is a Poisson random variable with parameter λ.*

$$E(X) = \lambda, \quad \mathrm{Var}(X) = \lambda.$$

Proof.

$$\begin{aligned} E[X] = \sum_{k=0}^{\infty} kP(X = k) &= \sum_{k=0}^{\infty} k\frac{e^{-\lambda}\lambda^k}{k!} \\ &= e^{-\lambda}\sum_{k=1}^{\infty}\frac{\lambda^k}{(k-1)!} \\ &= \lambda e^{-\lambda}\sum_{k=1}^{\infty}\frac{\lambda^{k-1}}{(k-1)!} \\ &= \lambda e^{-\lambda}\sum_{k=0}^{\infty}\frac{\lambda^k}{k!} \\ &= \lambda e^{-\lambda}e^{\lambda} \\ &= \lambda \end{aligned}$$

Note that we changed the sum from $k = 0$ to $k = 1$ because the term $k = 0$ contributes nothing to the sum.

The variance calculation is left as an exercise in HW3, using a very similar technique. □

Part III
Continuous Random Variables

12 Lecture XII: Probability Density Functions

Think about two common experiments:

- Toss a (fair) coin five times and record the number of heads. The outcome set is discrete and small: the random variable takes values $0, 1, 2, 3, 4, 5$. We compute probabilities by summing point masses (pmf).
- Now imagine measuring the time (in seconds) it takes a driver to clear a congested intersection. The measurement could be 12.37 s, 12.3721 s, or 12.372154... s. There is no natural list of possible values — the variable is *continuous*.

Fundamental question to carry through today: *How do the probabilistic ideas we used for the discrete case change (if at all) when we move to continuous measurements?* The short answer: conceptually nothing mystical — we replace sums by integrals, and many familiar formulas are direct consequences of calculus. We will make this precise.

Plan for this lecture.

1. Definitions and intuition for Probability Density Functions (pdf) and Cumulative Distribution Functions (cdf);
2. Computing probabilities from the pdf and cdf;
3. Recovering the pdf from the cdf and the role of the Fundamental Theorem of Calculus;
4. Examples.

12.1 Probability distributions for continuous random variables

Definition 12.1. *A function $f : \mathbb{R} \to [0, \infty)$ is a* probability density function *(pdf) for a random variable X if*

$$\int_{-\infty}^{\infty} f(x)\,dx = 1,$$

and for every (measurable) interval A,

$$P(X \in A) = \int_A f(x)\,dx.$$

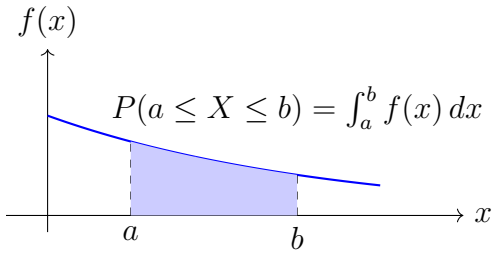
Intuitively, $f(x)$ is probability per unit length *around x: for a small window $[x, x + \Delta x]$,*

$$P(x \le X \le x + \Delta x) \approx f(x)\,\Delta x.$$

Important consequences (contrast with discrete).

- For a continuous X, single points carry zero probability: $P(X = c) = 0$ for any real c. This is why inclusion or exclusion of endpoints in interval probabilities does not matter:

$$P(a \le X \le b) = P(a < X < b) = P(a < X \le b) = P(a \le X < b).$$



- Computation of probabilities is **integration** rather than summation.

Concept / Operation	**Discrete X (pmf $p(x)$)**	**Continuous X (pdf $f(x)$)**
Interval probability	$P(a \le X \le b) = \sum_{x_i \in [a,b]} p(x_i)$	$P(a \le X \le b) = \int_a^b f(x)\,dx$
Singleton (point mass) probability	$P(X = c) = p(c)$ (can be > 0)	$P(X = c) = 0$
Normalization	$\sum_{\text{all } x_i} p(x_i) = 1$	$\int_{-\infty}^{\infty} f(x)\,dx = 1$

Takeaway 12.1. Moving from discrete to continuous random variables means replacing *finite sums of masses* by *integrals of densities*, and losing the notion of nonzero probability at a single point.

12.2 The Cumulative Distribution Function (cdf)

Definition 12.2 (CDF). *The cumulative distribution function F of a random variable X is*

$$F(x) = P(X \leq x).$$

For a continuous X with pdf f,

$$F(x) = \int_{-\infty}^{x} f(y)\, dy.$$

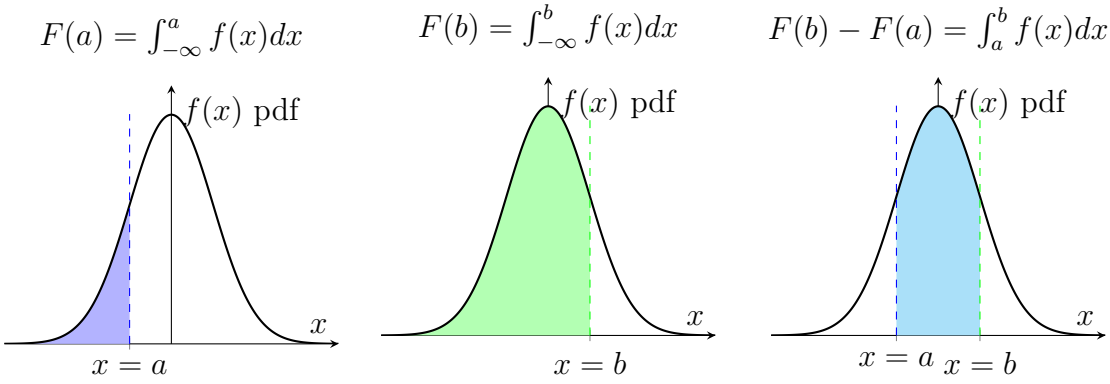
Hence F records the accumulated probability to the left of x.

Using the cdf to compute probabilities. A primary utility of F is that once F is known, many probabilities can be obtained by subtraction:

$$\boxed{\int_a^b f\,(x)\, dx = P(a \leq X \leq b) = F(b) - F(a)}$$

for any $a < b$. Graphically, the difference of the two cumulative areas equals the area under the density between a and b. This formula is the continuous analogue of the discrete identity where one uses differences of step values of the cdf to obtain block probabilities.

This result is also known as the **Connection between CDF and PDF via Integration**.



Remark (connection to discrete cdf). In Prop. 9.2, we have seen the discrete integer-valued case

$$P(a \leq X \leq b) = F(b) - F(a^-),$$

where $F(a^-) = \lim_{x \uparrow a} F(x)$ denotes the left limit of F at a. The left-limit appears because the discrete cdf has jumps located at mass points. (See your lecture notes for the discrete CDF discussion and properties of right-continuity.)

Concept / Operation	**Discrete X (cdf $F(x)$)**	**Continuous X (cdf $F(x)$)**
Definition	$F(a) = P(X \le a) = \sum_{x_i \le a} p(x_i)$	$F(a) = P(X \le a) = \int_{-\infty}^{a} f(x)\,dx$
Interval probability	$F(b) - F(a^-) = P(a \le X \le b)$ $= \sum_{x_i \in [a,b]} p(x_i)$	$F(b) - F(a) = P(a \le X \le b)$ $= \int_a^b f(x)\,dx$

where a^- represents the largest possible X value that is strictly less than a.

12.3 Recovering the pdf from the cdf; the Fundamental Theorem of Calculus

Theorem 12.1 (Connection between CDF and PDF via Differentiation)**.** *If F is differentiable at x then*

$$F'(x) = f(x).$$

This is the formal statement that the pdf is the derivative of the cdf wherever the derivative exists. The result is a direct consequence of the Fundamental Theorem of Calculus (FTC): the cdf is defined by an integral of f, so differentiating returns the integrand (almost everywhere).

Why this is not magic. You have seen the same pattern in pure calculus: if $G(x) = \int_a^x g(t)\,dt$ then $G'(x) = g(x)$. Here F (the cdf) plays the role of G and f (the pdf) plays the role of g. Probabilities for intervals become definite integrals, and formulas for manipulating probabilities become routine applications of the calculus toolkit (FTC, substitution, integration by parts when needed, etc.).

12.4 Worked examples

1. **Uniform on $[A, B]$.** If $X \sim \text{Unif}[A, B]$ then

$$f(x) = \frac{1}{B-A}\mathbf{1}_{[A,B]}(x), \qquad F(x) = \begin{cases} 0, & x < A, \\ \dfrac{x-A}{B-A}, & A \le x < B, \\ 1, & x \ge B. \end{cases}$$

Therefore $P(a \le X \le b) = \dfrac{b-a}{B-A}$ for $A \le a < b \le B$.

2. **A polynomial pdf on** $[0, 2]$**.** Suppose

$$f(x) = \frac{1}{8} + \frac{3}{8}x, \quad 0 \le x \le 2.$$

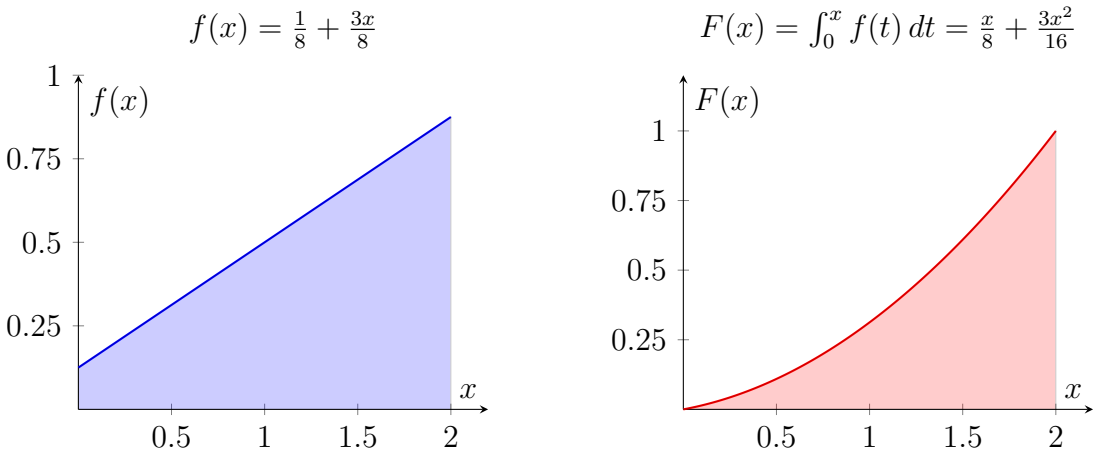
Then

$$F(x) = \int_0^x \left(\frac{1}{8} + \frac{3}{8}y\right) dy = \frac{x}{8} + \frac{3x^2}{16}, \quad 0 \le x \le 2,$$

and, for instance,

$$P(1 \le X \le 1.5) = F(1.5) - F(1) = \frac{19}{64} \approx 0.297.$$

Notice that $F(2) = 1$ which confirms that $f(x)$ is indeed a pdf. It's always a nice sanity check by evaluating the CDF at the largest possible value of X. See the sketch to visualize the area (a trapezoid with height 2, with width $1/8 + 7/8 = 1$).



Takeaway 12.2 (Main points from this lecture)**.**

- Moving from discrete to continuous replaces sums by integrals: nothing mystical, only calculus.
- The cdf is the bridge between probability and calculus: $F(x) = \int_{-\infty}^{x} f(x)$. Subtraction of cdf values gives interval probabilities; differentiation (FTC) recovers the density.
- Watch endpoints in discrete problems; in continuous problems they are irrelevant (probability zero).

Optional read: where exactly does the left-limit notation $F(a^-)$ come from?

In discrete distributions F jumps at mass points. If X is integer-valued and a is integer, then

$$P(a \le X \le b) = F(b) - F(a^-),$$

because $F(a^-)$ excludes the jump at a while $F(a)$ includes it. For continuous distributions the jump size is zero, so $F(a^-) = F(a)$ and the simpler formula $F(b) - F(a)$ is correct and clean. This is exactly where the discrete/continuous formulas align: the same subtraction idea, but replaced by differences of steps in the discrete case and differences of areas (integrals) in the continuous case.

12.5 Important Physical Connection: Probability Density as Mass Density

A helpful physical picture: imagine a thin wire or rod lying along the real line, with *total mass equal to 1.* Let $f(x)$ be its *mass density* (mass per unit length). Then

$$\text{Mass on } [a, b] = \int_a^b f(x)\, dx.$$

If we simply rename "mass" to "probability," the analogy is exact: $f(x)$ is the probability density and $P(a \le X \le b)$ is the "mass" of that region. This is why f must integrate to 1 — the total probability, like the total mass, must sum to 1.

This analogy is powerful: intuition about center of mass, balancing points, and distribution of weight all carry over to expectation, variance, and higher moments.

13 Lecture XIII: Expected Value, Variance and Moment Generating Functions

Motivating Questions

- In the last chapter, we defined the expected value for a **discrete** random variable as
$$E[X] = \sum_x x\, p(x).$$
How should this generalize if X is continuous?

- Recall from last class that probability is now represented by *areas under curves* rather than sums of point masses.

- Physically: if $x_1, x_2, \ldots, x_n$ are "masses" with local density $\rho(x_i)$, the balancing point (the fulcrum) is
$$\frac{\sum_{i=1}^n x_i \rho(x_i) \Delta x}{\sum_{i=1}^n \rho(x_i) \Delta x}.$$
What happens as $n \to \infty$ and $\Delta x \to 0$?

- If we know only the mean and variance, do we truly know **everything** about the random variable's behavior?

13.1 Expected Value for Continuous Random Variables

Definition 13.1. *If X is a continuous random variable with pdf $f(x)$,*
$$E[X] = \int_{-\infty}^{\infty} x f(x)\, dx.$$

- The expected value is now an **integral**, not a sum.

- The integral weights each infinitesimal contribution x by its "mass density" $f(x)$.

- **Mnemonic:** expectation = center of mass.

Proposition 13.1 (Linearity of Expectation)**.** *For any real numbers a, b and any continuous random variable X with pdf $f(x)$,*
$$E[aX + b] = a\, E[X] + b.$$

Proof.

$$E[aX+b] = \int_{-\infty}^{\infty} (ax+b)f(x)\,dx = a\int_{-\infty}^{\infty} xf(x)\,dx + b\int_{-\infty}^{\infty} f(x)\,dx = aE[X]+b,$$

□

Example 13.1 (Uniform Distribution)**.** Let $X \sim \text{Unif}\,[a,b]$.

$$f(x) = \frac{1}{b-a}, \quad a \le x \le b.$$

Then

$$E[X] = \int_a^b x\frac{1}{b-a}\,dx = \frac{b+a}{2},$$

which is just the midpoint, as expected for a uniform bar of constant density.

13.2 The Law of the Unconscious Statistician (LOTUS)

Theorem 13.1. *If $g(X)$ is any function of X,*

$$E[g(X)] = \int_{-\infty}^{\infty} g(x)f(x)\,dx.$$

This is the ***continuous analog*** *of*

$$E[g(X)] = \sum_x g(x)p(x).$$

Important Remark 13.1. *You do* not *need to find the distribution of $Y = g(X)$ first—just integrate.*

Example 13.2. If $X \sim \text{Unif}[a,b]$,

$$E[X^2] = \int_a^b x^2\frac{1}{b-a}dx = \frac{b^3-a^3}{3(b-a)} = \frac{b^2+ab+a^2}{3}.$$

13.3 Variance and Standard Deviation

Definition 13.2. *Variance is defined as*

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

where the expected values here are understood as integrals.

Example 13.3. For $X \sim \text{Unif}[a,b]$,

$$\text{Var}(X) = E[X^2] - (E[X])^2 = \frac{b^2+ab+a^2}{3} - \frac{(b+a)^2}{4} = \frac{(b-a)^2}{12}.$$

Summary Table: Discrete vs. Continuous

Concept	**Discrete (pmf $p(x)$)**	**Continuous (pdf $f(x)$)**
Normalization	$\sum_x p(x) = 1$	$\int_{-\infty}^{\infty} f(x)\,dx = 1$
Interval Probability	$\sum_{x\in[a,b]} p(x)$	$\int_a^b f(x)\,dx$
Expectation	$E[X] = \sum_x x\,p(x)$	$E[X] = \int_{-\infty}^{\infty} x f(x)\,dx$
Function of X	$E[g(X)] = \sum g(x)\,p(x)$	$E[g(X)] = \int_{-\infty}^{\infty} g(x)f(x)\,dx$
Variance	$\mathrm{Var}(X) = E[X^2] - (E[X])^2$	Same formula, with integrals
Linearity	$E[aX+b] = aE[X] + b$	Same result

Takeaway 13.1. The glue between discrete and continuous r.v. is **calculus**: sums become integrals.

Why Go Beyond Mean and Variance?

Motivating Question: If two random variables have the same mean and variance, do they necessarily behave the same?

Example 13.4.

$$\begin{aligned} \text{Dataset A:} &\quad \{1,1,4\} \\ \text{Dataset B:} &\quad \{0,3,3\} \end{aligned}$$

Compute the mean:

$$\mu_A = \frac{1+1+4}{3} = 2, \qquad \mu_B = \frac{0+3+3}{3} = 2.$$

So both datasets have mean 3.

Compute the variance:

$$\sigma_A^2 = \frac{1}{3}\big[(1-2)^2 + (1-2)^2 + (4-2)^2\big] = \frac{6}{3} = 2.$$

$$\sigma_B^2 = \frac{1}{3}\big[(0-2)^2 + (3-2)^2 + (3-2)^2\big] = \frac{6}{3} = 2.$$

Thus, both datasets have the same mean *and* variance, but clearly they don't have the same data.

Mean and variance alone do not capture the full distributional behavior. This motivates the study of higher-order moments.

13.4 Moment Generating Functions (MGFs)

Definition 13.3. *The moment generating function (mgf) of X is*

$$M_X(t) = E[e^{tX}] = \int_{-\infty}^{\infty} e^{tx} f(x)dx,$$

defined for t near 0 where the integral converges.

13.4.1 Why does the MGF generate moments?

Key Knowledge – Maclaurin Series for e^{tX} near $t = 0$:

$$e^{tX} = 1 + tX + \frac{t^2X^2}{2!} + \frac{t^3X^3}{3!} + \dots$$

Proposition 13.2 (Moment Generation)**.** *Let $M_X(t)$ be the moment generating function of the r.v. X. Then*

$$M_X^{(n)}(0) = E[X^n]$$

i.e., the nth derivative of the mgf evaluated at $t = 0$ yields the nth moment of X.

Proof. Taking expectations term by term:

$$M_X(t) = E[e^{tX}] = 1 + tE[X] + \frac{t^2E[X^2]}{2!} + \frac{t^3E[X^3]}{3!} + \cdots = \sum_{n=0}^{\infty} \frac{t^n}{n!}E[X^n]$$

Differentiate term-by-term with respect to t:

$$M_X'(t) = \sum_{n=1}^{\infty} \frac{nt^{n-1}}{n!}E[X^n] = \sum_{n=1}^{\infty} \frac{t^{n-1}}{(n-1)!}E[X^n].$$

Evaluating at $t = 0$ leaves only the $n = 1$ term:

$$M_X'(0) = E[X].$$

Similarly,

$$M_X''(t) = \sum_{n=2}^{\infty} \frac{t^{n-2}}{(n-2)!} E[X^n] \quad \Rightarrow \quad M_X''(0) = E[X^2],$$

and more generally $M_X^{(r)}(0) = E[X^r]$ for any positive integer r.

Thus:

$$M_X^{(1)}(0) = E[X], \qquad M_X^{(2)}(0) = E[X^2], \qquad M_X^{(n)}(0) = E[X^n].$$

□

Important Remark 13.2. *One function $M_X(t)$ encodes all moments of X, up to arbitrary order.*

13.4.2 Important Example: MGF of a Uniform Random Variable

Example 13.5 (Moment Generating Function of a Uniform Random Variable)**.**

Let $X \sim \text{Unif}[a, b]$. Then the pdf is

$$f(x) = \frac{1}{b-a}, \qquad a \le x \le b.$$

Compute the mgf:

$$M_X(t) = E[e^{tX}] = \int_a^b e^{tx} \frac{1}{b-a}\, dx.$$

Integrate:

$$M_X(t) = \frac{1}{b-a} \cdot \frac{e^{tb} - e^{ta}}{t}, \qquad t \neq 0.$$

At $t = 0$, use the limit

$$M_X(0) = \lim_{t \to 0} \frac{e^{tb} - e^{ta}}{t(b-a)} = 1,$$

as expected for an mgf.

Differentiate to find moments:

First derivative:

$$M_X'(t) = \frac{d}{dt} \left[\frac{e^{tb} - e^{ta}}{t(b-a)} \right].$$

Use the quotient rule (or differentiate numerator carefully):

$$M_X'(t) = \frac{(be^{tb} - ae^{ta})t - (e^{tb} - e^{ta})}{t^2(b-a)}.$$

Evaluate at $t = 0$ using L'Hôpital's rule (HW 4):

$$M_X'(0) = \frac{a+b}{2} = E[X].$$

Second derivative:
Differentiate $M_X'(t)$ again (algebra is a bit tedious but straightforward):

$$M_X''(t) = \frac{d}{dt}\left[\frac{(be^{tb} - ae^{ta})t - (e^{tb} - e^{ta})}{t^2(b-a)}\right].$$

At $t = 0$,

$$M_X''(0) = E[X^2] = \frac{a^2 + ab + b^2}{3}.$$

Finally, compute the variance:

$$\mathrm{Var}(X) = E[X^2] - (E[X])^2 = \frac{a^2 + ab + b^2}{3} - \left(\frac{a+b}{2}\right)^2 = \frac{(b-a)^2}{12},$$

in agreement with our earlier formula.

Takeaway 13.2.

- Continuous expectation is an integral: think "center of mass."
- LOTUS allows computing $E[g(X)]$ directly.
- Variance formula is identical to the discrete case.
- Mean and variance are not enough; we need higher moments.
- MGFs provide a single calculus-based object to generate all moments.

14 Lecture XIV: Gaussian Random Variable (The Normal Distribution)

Motivating questions

We introduced the Uniform distribution as our first continuous distribution. Here is a second and extraordinarily important one: the *normal* (Gaussian) distribution. Why devote a section to it? Because many measurement errors and aggregated quantities are (approximately) normal, and sums/averages of independent variables tend to be normal (Central Limit Theorem).

The most important technique **demanded** to understand the most important distribution is your integration techniques. We first define the pdf of a normal random variable, and then demonstrate that it is a valid pdf, and has the statistical properties we assigned it to have.

14.1 Definition and PDF

Definition. A random variable X is *normal* (or Gaussian) with mean $\mu \in \mathbb{R}$ and standard deviation $\sigma > 0$ if its pdf is

$$f(x;\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} \exp\Big(-\frac{(x-\mu)^2}{2\sigma^2}\Big), \qquad -\infty < x < \infty.$$

We write $X \sim N(\mu, \sigma)$.

Remark. *Note that a normal random variable has $\mathbb{R}$ as its range (and thus its pdf has $\mathbb{R}$ as its domain), unlike the uniform distribution that is always supported on a closed, finite interval.*

14.2 Standard normal and standardization

The *standard normal* is $Z \sim N(0,1)$ with pdf $f_Z(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}$ and cdf $\Phi(z) = P(Z \le z) = \int_{-\infty}^{z} f_Z(y)\,dy$. There is no elementary closed form for Φ, so we use tables or software (`pnorm` in R, `normcdf` in MATLAB).

To convert $X \sim N(\mu,\sigma)$ to standard normal: $Z = (X-\mu)/\sigma \sim N(0,1)$, hence

$$P(a \le X \le b) = \Phi\Big(\frac{b-\mu}{\sigma}\Big) - \Phi\Big(\frac{a-\mu}{\sigma}\Big).$$

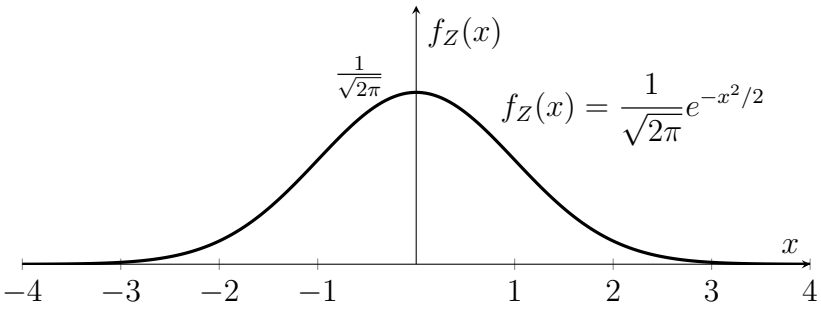


Figure 3: The Probability Density Function of a **standard** normal random variable, with $\mu = 0$ and $\sigma = 1$.

14.3 A Calculus Review for an Important Fact and Technique in Gaussian Integration

Here is a technical result (often referred to as a lemma, not as substantial as a theorem but important enough). The result itself is not surprising, but it is the technique employed in the calculations that are worth the note.

Lemma 14.1 (standard normal density integrates to 1)**.**

$$\int_{-\infty}^{\infty} \frac{1}{\sigma\sqrt{2\pi}} \Big(- \frac{(x-\mu)^2}{2\sigma^2} \Big)\, dx = 1.$$

Proof. **Step 0: Reduction to standard normal.**

The density of a general normal random variable $X \sim N(\mu, \sigma^2)$ is

$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}}\, e^{-(x-\mu)^2/(2\sigma^2)}.$$

By the change of variable

$$z = \frac{x-\mu}{\sigma} \qquad \Longrightarrow \qquad dx = \sigma\, dz,$$

the integral of f_X over the real line becomes

$$\int_{-\infty}^{\infty} f_X(x)\, dx = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-z^2/2}\, dz.$$

Thus it suffices to show that the standard normal density integrates to 1.

Step 1: Consider the square of the integral. Let

$$I = \int_{-\infty}^{\infty} e^{-x^2/2}\, dx.$$

Then

$$I^2 = \left(\int_{-\infty}^{\infty} e^{-x^2/2}\,dx\right)\left(\int_{-\infty}^{\infty} e^{-y^2/2}\,dy\right) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{-(x^2+y^2)/2}\,dx\,dy.$$

Step 2: Switch to polar coordinates. Let

$$x = r\cos\theta, \quad y = r\sin\theta, \quad dx\,dy = r\,dr\,d\theta.$$

Then

$$I^2 = \int_0^{2\pi}\int_0^{\infty} e^{-r^2/2} r\,dr\,d\theta.$$

Step 3: Evaluate the radial integral. Using the substitution $u = r^2/2$, $du = r\,dr$, we get

$$\int_0^{\infty} e^{-r^2/2} r\,dr = \int_0^{\infty} e^{-u}\,du = 1.$$

Step 4: Evaluate the angular integral. We have

$$I^2 = \int_0^{2\pi} 1\,d\theta = 2\pi \quad \Longrightarrow \quad I = \sqrt{2\pi}.$$

Step 5: Normalize. Therefore,

$$\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-x^2/2}\,dx = \frac{1}{\sqrt{2\pi}} \cdot \sqrt{2\pi} = 1.$$

□

We now have verified that the normal pdf is indeed a pdf. Can we further verify that the expected value and variance are μ and σ^2 respectively? They are mere integration exercises.

If you are here for Problem 10 on HW4, please do the problem yourself first without looking at the proof. Develop your own thoughts first, then compare them with mine.

Lemma 14.2 (Mean of a normal distribution)**.** *If $X \sim N(\mu, \sigma^2)$ then $\mathbb{E}[X] = \mu$.*

Proof. By definition,

$$\mathbb{E}[X] = \int_{-\infty}^{\infty} x f_X(x)\,dx = \int_{-\infty}^{\infty} x \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) dx.$$

Perform the change of variable $z = \dfrac{x-\mu}{\sigma}$. Then $x = \mu + \sigma z$ and $dx = \sigma\, dz$. The integral becomes

$$\mathbb{E}[X] = \int_{-\infty}^{\infty} (\mu + \sigma z) \frac{1}{\sqrt{2\pi}} e^{-z^2/2}\, dz = \mu \underbrace{\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-z^2/2}\, dz}_{=1} + \sigma \underbrace{\int_{-\infty}^{\infty} \frac{z}{\sqrt{2\pi}} e^{-z^2/2}\, dz}_{=0}.$$

The first integral equals 1 by the standard normal normalization. The second integral vanishes because the integrand $ze^{-z^2/2}$ is an odd function integrated over the symmetric interval $(-\infty, \infty)$. Hence $\mathbb{E}[X] = \mu$. □

If you are here for Problem 10 on HW4, please do the problem yourself first without looking at the proof. Develop your own thoughts first, then compare them with mine.

Lemma 14.3 (Variance of a normal distribution)**.** *If* $X \sim N(\mu, \sigma^2)$ *then* $\mathrm{Var}(X) = \sigma^2$.

Proof. By definition $\mathrm{Var}(X) = \mathbb{E}[(X-\mu)^2]$. Using the density and the same substitution $z = (x-\mu)/\sigma$ with $dx = \sigma\, dz$, we obtain

$$\mathrm{Var}(X) = \int_{-\infty}^{\infty} (x-\mu)^2 \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/(2\sigma^2)}\, dx = \int_{-\infty}^{\infty} \sigma^2 z^2 \frac{1}{\sqrt{2\pi}} e^{-z^2/2}\, dz = \sigma^2 \underbrace{\int_{-\infty}^{\infty} \frac{z^2}{\sqrt{2\pi}} e^{-z^2/2}\, dz}_{=1?}.$$

It remains to show the normalizing second moment equals 1:

$$\int_{-\infty}^{\infty} z^2 \frac{1}{\sqrt{2\pi}} e^{-z^2/2}\, dz = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} z^2 e^{-z^2/2}\, dz.$$

Set $J = \int_{-\infty}^{\infty} z^2 e^{-z^2/2}\, dz$. Evaluate J by integration by parts: write

$$J = \int_{-\infty}^{\infty} z \cdot \left(ze^{-z^2/2}\right) dz.$$

Take $u = z$ and $dv = ze^{-z^2/2}\, dz$. Then $du = dz$ and $v = -e^{-z^2/2}$ (since $\frac{d}{dz}(-e^{-z^2/2}) = ze^{-z^2/2}$). Thus

$$J = \left[-ze^{-z^2/2}\right]_{-\infty}^{\infty} + \int_{-\infty}^{\infty} e^{-z^2/2}\, dz.$$

The boundary term vanishes because $ze^{-z^2/2} \to 0$ as $|z| \to \infty$. Hence

$$J = \int_{-\infty}^{\infty} e^{-z^2/2}\, dz = \sqrt{2\pi},$$

where the last equality is the standard Gaussian integral. Therefore

$$\int_{-\infty}^{\infty} \frac{z^2}{\sqrt{2\pi}} e^{-z^2/2}\, dz = \frac{J}{\sqrt{2\pi}} = \frac{\sqrt{2\pi}}{\sqrt{2\pi}} = 1,$$

and $\mathrm{Var}(X) = \sigma^2$. □

14.4 MGF of the normal distribution

The moment-generating function of a random variable X is $M_X(t) = E[e^{tX}]$ (when it exists near $t = 0$). For the standard normal Z we compute

$$M_Z(t) = \int_{-\infty}^{\infty} e^{tz} \frac{1}{\sqrt{2\pi}} e^{-z^2/2}\, dz = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{1}{2}(z^2 - 2tz)}\, dz.$$

Complete the square $z^2 - 2tz = (z - t)^2 - t^2$ to obtain

$$M_Z(t) = e^{t^2/2} \cdot \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-(z-t)^2/2}\, dz = e^{t^2/2}.$$

For a general $X \sim N(\mu, \sigma)$, write $X = \mu + \sigma Z$ to get

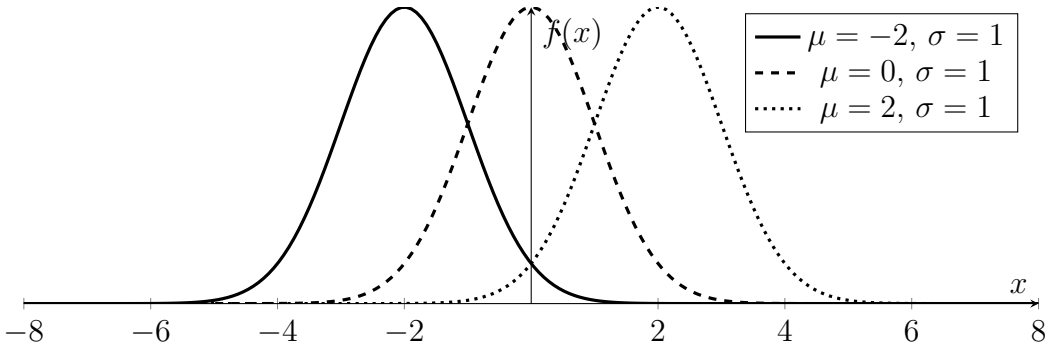
$$M_X(t) = e^{\mu t} M_Z(\sigma t) = \exp\left(\mu t + \frac{\sigma^2 t^2}{2}\right).$$

Differentiating at $t = 0$ recovers $E[X] = \mu$ and $\mathrm{Var}(X) = \sigma^2$.

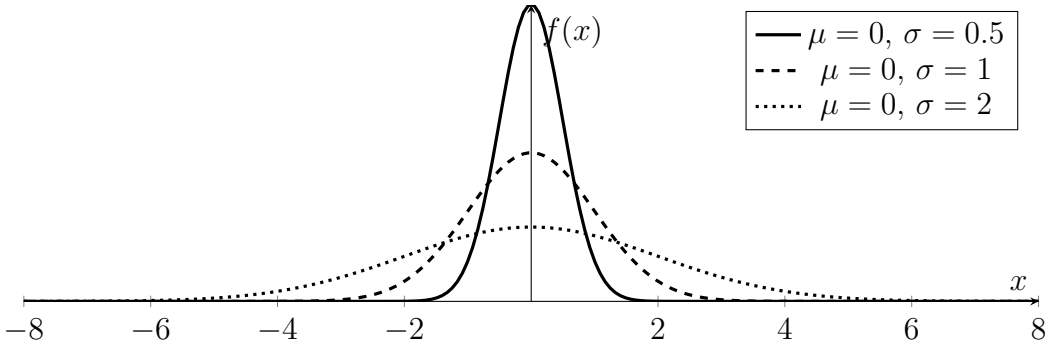
14.5 General normal: sketches

Below are a few sketches illustrating how the pdf changes when its mean μ and standard deviation σ varies.

Fixed variance, varying mean



Fixed mean, varying variance



14.6 Worked example (with picture)

Example 14.1. Let $X \sim N(100, 15)$. Find $P(X \leq 130)$.

Solution. Compute the standardized value

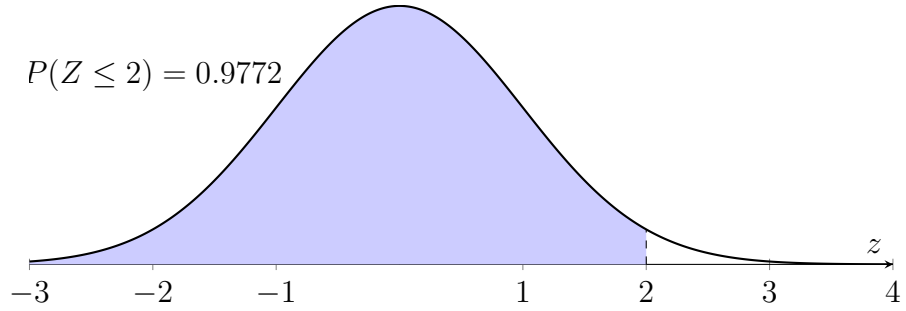
$$Z = \frac{X - 100}{15}, \qquad z = \frac{130 - 100}{15} = \frac{30}{15} = 2.$$

Thus

$$P(X \leq 130) = P\Big(\frac{X - 100}{15} \leq 2\Big) = P(Z \leq 2) = \Phi(2) \approx 0.9772.$$

This final number can be obtained from the MATLAB command normcdf(100,15,2).

The shaded area under the pdf below illustrates this probability.



Takeaway 14.1. The normal distribution is central both as a model for many empirical quantities and as the limiting law for sums/averages. We will rely heavily on the normal in later inference and in the statement/proof of the Central Limit Theorem.

15 Lecture XV: The Exponential Families

15.1 Motivation and context

We just finished the Normal distribution, whose pdf is symmetric about the mean and is a prototypical *bell curve*. Many natural measurements are approximately symmetric, but not all. Some important scientific quantities are *asymmetric* (skewed):

- Waiting times (time between arrivals at a service desk), reaction times, and lifetimes of some components are nonnegative and often skewed to the right.
- Rainfall duration, inter-arrival times in queues, and lengths of hospital stays are typical applied examples.

Reminder: these are *still* probability densities and cumulative distributions — **the underlying probabilistic principles (cdf, pdf, expectation, conditional probability, mgf) do not change**; only the formulae do. The exponential and gamma families are canonical models for positive-valued, skewed data, and they also connect tightly with the Poisson process and sums of exponentials.

15.2 Exponential distribution

The exponential distribution models positive-valued waiting times (as one example). It makes no sense to have a negative waiting time, so the support is $x > 0$.

15.2.1 PDF and CDF

Definition 15.1 (PDF for an Exponential Random Variable)**.** *For $\lambda > 0$, the exponential pdf is*

$$f(x;\lambda) = \begin{cases} \lambda e^{-\lambda x}, & x > 0, \\ 0, & \textit{otherwise.} \end{cases}$$

We say $X \sim \mathrm{Exp}\,(\lambda)$ when the random variable has the above pdf.

Notation 15.1. *$f(x;\lambda)$ is understood as a function of x (as its independent variable) with λ as a parameter. More precisely,*

1. *x, independent variable (argument):*
   x is the input to the function. For a pdf, this is the possible value that the random variable X might take; when you plot f, x is the horizontal axis.

2. *λ, parameter:*
   *λ is a **fixed** constant that determines the shape of the distribution/density function. Once **chosen** (according to the context), it doesn't vary within the function evaluation.*

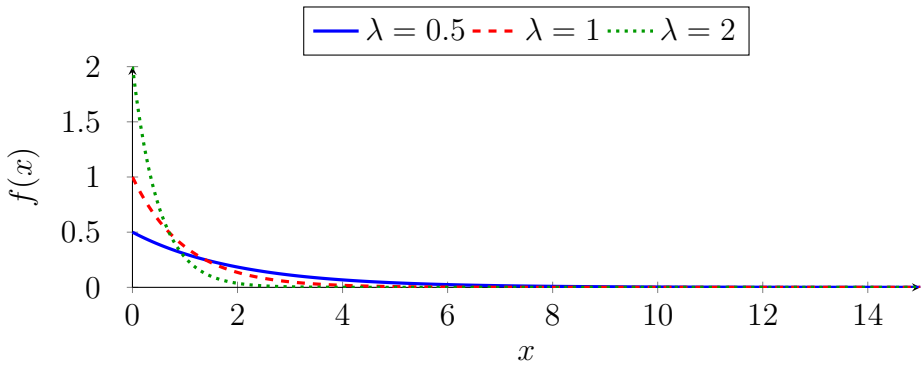
Proposition 15.1 (CDF for an Exponential Random Variable)**.** *Given $X \sim \text{Exp}(\lambda)$, its CDF is*

$$F(x) = 1 - e^{-\lambda x}, \quad x > 0.$$

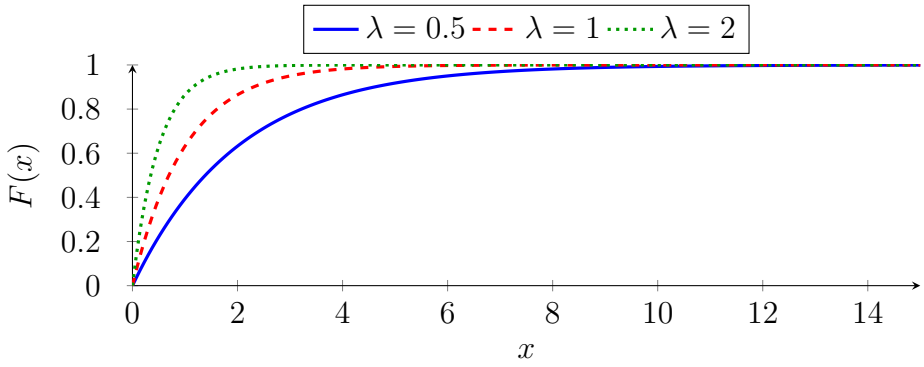
Proof. Integrating the pdf gives the cdf for $x > 0$:

$$F(x;\lambda) = P(X \leq x) = \int_0^x \lambda e^{-\lambda t}\, dt = 1 - e^{-\lambda x}.$$

□



(a) PDF of an exponential distribution with $\lambda = 1$.



(b) CDF of the same exponential distribution.

Figure 4: Exponential distribution with parameter $\lambda = 1$. (a) PDF; (b) CDF.

Proposition 15.2 (Expected value)**.** *Let* $X \sim \text{Exp}(\lambda)$ *with pdf*

$$f(x;\lambda) = \lambda e^{-\lambda x}, \qquad x > 0,\ \lambda > 0.$$

Then

$$\mathbb{E}[X] = \frac{1}{\lambda}.$$

Proof. By definition,

$$\mathbb{E}[X] = \int_0^\infty x\,\lambda e^{-\lambda x}\,dx.$$

Integration by parts: set $u = x$, $dv = \lambda e^{-\lambda x}\,dx$. Then $du = dx$ and $v = -e^{-\lambda x}$. Thus

$$\int_0^\infty x\lambda e^{-\lambda x}\,dx = -xe^{-\lambda x}\Big|_0^\infty + \int_0^\infty e^{-\lambda x}\,dx.$$

The boundary term vanishes, since $xe^{-\lambda x} \to 0$ as $x \to \infty$. Therefore,

$$\mathbb{E}[X] = \int_0^\infty e^{-\lambda x}\,dx = -\tfrac{1}{\lambda}e^{-\lambda x}\Big|_0^\infty = \frac{1}{\lambda}.$$

□

Important Remark 15.1 (Interpretation of the expected value of an exponential random variable)**.** *Suppose the waiting time between bus arrivals is an exponential random variable* $X \sim \text{Exp}\,(\lambda)$*. The expected value is thus the average waiting time.*

λ, in this context, means the number of buses that arrives per time unit. The larger it is, the more buses arrive within this time unit. So, λ is a ***rate*** *constant, in units* #/time*.*

The expected waiting time thus is intuitively ***inverse*** *of λ because the higher the rate constant, the more frequent you will see bus arrivals, and thus the shorter you need to wait to see one arrival, on average.*

Next, we consider the moment generating function of the exponential random variable.

Proposition 15.3 (Moment Generating Function)**.** *For* $X \sim \text{Exp}(\lambda)$*, the moment generating function exists for* $t < \lambda$ *and is*

$$M_X(t) = \mathbb{E}[e^{tX}] = \frac{\lambda}{\lambda - t}, \qquad t < \lambda.$$

Proof. By definition,

$$M_X(t) = \int_0^\infty e^{tx}\, \lambda e^{-\lambda x}\, dx = \lambda \int_0^\infty e^{-(\lambda - t)x}\, dx.$$

The integral converges when $\lambda - t > 0$ (an important fact based on improper integrals). For $t < \lambda$,

$$M_X(t) = \lambda \cdot \frac{1}{\lambda - t} = \frac{\lambda}{\lambda - t}.$$

Sanity check for the expected value: differentiating gives

$$M_X'(t) = \frac{\lambda}{(\lambda - t)^2}, \quad \Rightarrow \quad M_X'(0) = \frac{1}{\lambda},$$

which matches Prop. 15.2. □

Remark. *Is the restriction $t < \lambda$ intuitive for the mgf of an exponential random variable? Remember that the mgf is always positive since it is the exponential of some function. So, the mgf, after we went through the calculation, only makes when positive, achieved precisely by the condition $t < \lambda$.*

15.2.2 Example calculation

Example 15.1. Suppose the mean response time is 5 seconds (so $E(X) = 1/\lambda = 5$ and $\lambda = 0.2$). The probability the response time is at most 10 seconds is

$$P(X \leq 10) = 1 - e^{-0.2 \cdot 10} = 1 - e^{-2} \approx 0.865.$$

15.2.3 Memoryless property (primer on conditional probability)

A striking property is that the exponential distribution is *memoryless*:

$$P(X > t + s \mid X > t) = P(X > s) \qquad (s, t \geq 0).$$

Proof: by definition of conditional probability,

$$P(X > t + s \mid X > t) = \frac{P(X > t + s)}{P(X > t)} = \frac{e^{-\lambda(t+s)}}{e^{-\lambda t}} = e^{-\lambda s} = P(X > s).$$

Important Remark 15.2 (Interpretation of Memorylessness)**.** *if a component has already survived time t, the distribution of remaining life is the same as the original distribution. This gives an elementary motivation for* ***Markovian jump processes*** *and constant hazard rates.*

15.3 Gamma distribution

15.3.1 Where it arises

The gamma family generalizes the exponential and appears as:

- the distribution of sums of independent exponential random variables (Erlang when the shape parameter is integer);
- a flexible model for skewed lifetimes and waiting-times in reliability, biology, and queuing;
- a conjugate prior family for Poisson-like likelihoods in Bayesian inference.

The exponential and gamma families provide flexible models for nonnegative, skewed data while keeping analytical tractability (closed-form pdf, mgf, and relations to Poisson processes).

Definition 15.2 (PDF of a Gamma Random Variable)**.** *A continuous random variable X is said to have a* ***gamma distribution*** *if the pdf of X is*

$$f(x;\alpha,\beta)=\begin{cases}\frac{1}{C(\alpha,\beta)}x^{\alpha-1}e^{-\frac{x}{\beta}}, & x>0,\\ 0, & \textit{otherwise}\end{cases}$$

where the parameters α and β are both positive. When $\beta=1$, X is said to have a ***standard gamma distribution****, and its pdf may be denoted as $f(x;\alpha)$.*

The proposition explicitly computes the normalization constant so that $f(x;\alpha,\beta)$ is indeed a pdf.

Proposition 15.4 (The Normalization Constant $C(\alpha,\beta)$)**.**

$$C(\alpha,\beta)=\beta^{\alpha}\Gamma(\alpha),$$

where $\Gamma(\alpha)$ is the gamma function

$$\Gamma(\alpha)=\int_0^{\infty}x^{\alpha-1}e^{-x}dx.$$

Proof. The pdf must integrate to 1, so

$$1=\int_0^{\infty}f(x;\alpha,\beta)\,dx=\int_0^{\infty}\frac{1}{C(\alpha,\beta)}x^{\alpha-1}e^{-\frac{x}{\beta}}dx$$
$$\implies C(\alpha,\beta)=\int_0^{\infty}x^{\alpha-1}e^{-\frac{x}{\beta}}dx$$

With a change of variable $y = \frac{x}{\beta}$, we see that

$$\begin{aligned} C(\alpha, \beta) &= \int_0^\infty x^{\alpha-1} e^{-\frac{x}{\beta}} dx \\ &= \int_0^\infty (\beta y)^{\alpha-1} e^{-y} \beta \, dy \\ &= \beta^{\alpha-1} \beta \int_0^\infty y^{\alpha-1} e^{-y} \, dy \\ &= \beta^\alpha \Gamma(\alpha) \end{aligned}$$

□

Remark 15.1. *If $\alpha = 1$ and $\beta = \frac{1}{\lambda}$, then we retrieve the exponential distribution.*

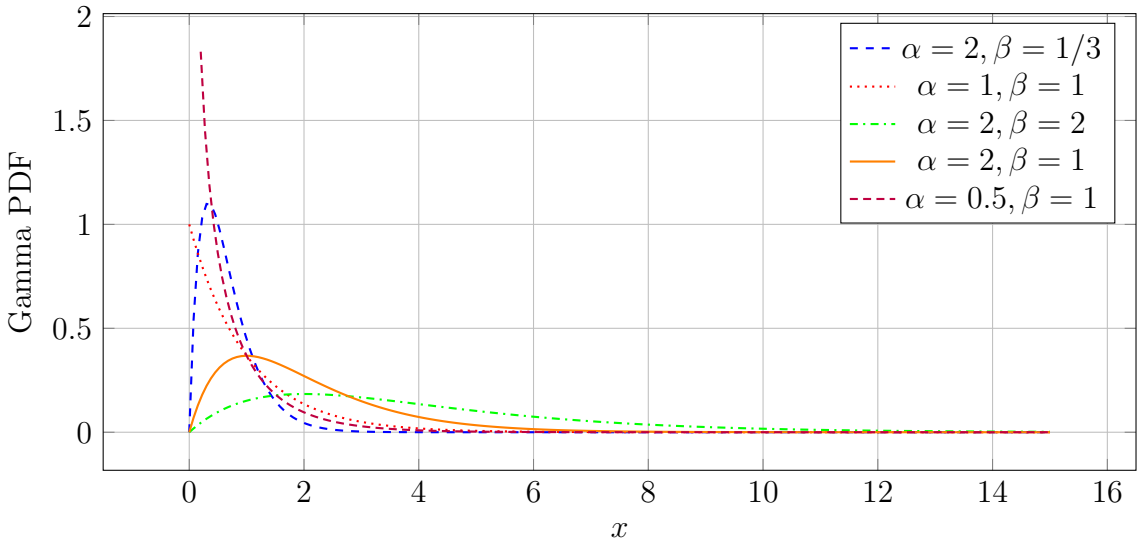


Figure 5: PDF of gamma random variables with different pairs of parameters. α is the **shape parameter** as it significantly alters the monotonicity of the pdf (see $\alpha = 1$ vs $\alpha = 2$). β is the **scale parameter** because its value only changes how stretched/compressed the shape is (fixed by α).

Remark 15.2 (This is so ugly. Why are we doing this?)**.** *Think of α and β as degrees of freedom, which makes the gamma distribution extremely* ***versatile*** *in terms of modeling real world data. You can fit it to data from extremely spiked distribution (say $\alpha < 0.5$), to bell curve-like one ($\alpha > 2$). Furthermore, it makes the exponential distribution (and many others) a special case.*

15.3.2 Computing Probabilities from Gamma Distribution

To compute probabilities using the pdf or cdf of the gamma distribution often requires numerical integration techniques, since the Gamma function itself is defined as an integral that does not have a closed-form anti-derivative. The following table lists the built-in command to compute probabilities using the cdf of gamma distribution.

	Gamma	**Exponential**
Function:	cdf	cdf
Notation:	$G(x/\beta;\alpha)$	$F(x;\lambda) = 1 - e^{-\lambda x}$
Matlab:	`gamcdf(x, alpha, beta)`	`expcdf(x, 1/lambda)`
R:	`pgamma(x, alpha, 1/beta)`	`pexp(x, lambda)`

Table 1: Summary of code for gamma and exponential distributions

Remark. *Nothing fundamental has changed. You are still using the cdf of a distribution, with the only difference from before that the cdf itself has changed.*

16 Lecture XVI: Transformation of Random Variables

16.1 Motivation

We know the **LOTUS formula** (Law of the Unconscious Statistician) allows us to compute expectations of functions of random variables:

$$E[g(X)] = \int g(x) f_X(x)\, dx.$$

So why do we need to go further and find the actual distribution (pdf/cdf) of a transformed variable $Y = g(X)$?

- In practice we often want more than just mean or variance: we may need quantiles, probabilities of events, or to simulate the distribution of Y.
- Example: If X is income, $Y = \log(X)$ is often used for regression and visualization (stabilizes variance, reduces skew). We need the *distribution* of Y, not just its mean.
- Example: In reliability engineering, if X is a lifetime random variable, $Y = X^2$ might represent energy or stress to failure. To compute $P(Y \leq y)$ directly, we need the pdf of Y.
- Example: In wireless communication, if X is a Gaussian noise amplitude, then $Y = X^2$ (signal power) follows a chi-square/gamma distribution. The distribution of Y is essential to calculate error probabilities.

Thus: understanding the pdf/cdf of transformed variables is not just about computing moments, but about full probabilistic modeling.

16.2 Definition and connection to calculus

Suppose X has pdf $f_X(x)$ and we define $Y = g(X)$ for some monotone, differentiable function g. Then, define $h = g^{-1}$, i.e., the inverse of g. We have the transformation formula

$$f_Y(y) = f_X\big(h(y)\big) \left| \frac{dh}{dy}(y) \right|.$$

This is exactly the **change of variables** formula from Calculus II integrals:

$$\int f(x)\, dx = \int f(h(y)) \left| \frac{dh}{dy} \right| dy.$$

Question 16.1. Why the absolute value sign?

Answer. Think of dx as a little piece of length. Lengths are never negative. When we change variables, the derivative tells us how much the length is stretched or shrunk, but we always keep the absolute value — because we're measuring size, not orientation.

Question 16.2. More importantly: Does h always exist? What if it doesn't? Is it then game over? We will get to this later.

16.3 Primer: Finding an Inverse Function

Recall: an inverse function $h = g^{-1}$ satisfies

$$g(h(y)) = y \quad \text{and} \quad h(g(x)) = x.$$

Procedures to find an inverse.

1. Start with $y = g(x)$.
2. Solve this equation for x in terms of y.
3. Rename the solved expression as $h(y)$.

Example 16.1. If $g(x) = 3x + 2$, then

$$y = 3x + 2 \quad \Longrightarrow \quad x = \frac{y-2}{3}.$$

So $h(y) = (y-2)/3$.

Example 16.2. If $g(x) = e^x$, then

$$y = e^x \quad \Longrightarrow \quad x = \ln y, \quad y > 0.$$

So $h(y) = \ln y$.

When an inverse does not exist. A function must be *one-to-one* (injective) on its domain for an inverse to exist. If two different x values map to the same y, then $h(y)$ would be ambiguous.

Definition 16.1 ((One-to-one / Injective))**.** *A function g is called* one-to-one *(or* injective*) on a domain D if it never takes the same value twice.*

Formally: for all $x_1, x_2 \in D$,

$$g(x_1) = g(x_2) \quad \Longrightarrow \quad x_1 = x_2.$$

Equivalently (the contrapositive statement): if $x_1 \neq x_2$, then $g(x_1) \neq g(x_2)$.

Theorem 16.1 (Continuous Injective Functions are Monotone)**.** *A continuous function is injective if and only if it is monotone.*

Bad example (non-invertible). If $g(x) = x^2$ on the whole real line, then

$$y = x^2 \quad \Longrightarrow \quad x = \pm\sqrt{y}.$$

There are two possible values of x for each $y > 0$, so g is not one-to-one and a single inverse does not exist. Moreover, the Theorem also applies here: since $y = x^2$ is not monotone on the real line, it cannot be injective, and thus no inverse exists on the real line.

Remedy. Restrict the domain so that g is monotone. For example, on $[0, \infty)$ the function $g(x) = x^2$ has inverse $h(y) = \sqrt{y}$. On $(-\infty, 0]$, the inverse is $h(y) = -\sqrt{y}$.

16.4 Derivative of an Inverse Function

Suppose g is differentiable and strictly monotone on its domain, so that it has an inverse $h = g^{-1}$. If $y = g(x)$, then by the chain rule

$$h'(y) = \frac{1}{g'(h(y))}.$$

Example 1. If $g(x) = e^x$, then $h(y) = \ln y$. Since $g'(x) = e^x = y$, we get

$$h'(y) = \frac{1}{y}.$$

Example 2. If $g(x) = x^3$, then $h(y) = \sqrt[3]{y}$. Since $g'(x) = 3x^2$, and $x = \sqrt[3]{y}$, we obtain

$$h'(y) = \frac{1}{3(\sqrt[3]{y})^2} = \frac{1}{3y^{2/3}}.$$

Connection to probability. This formula is exactly what shows up in the transformation of random variables:

$$f_Y(y) = f_X(h(y)) \cdot \big|h'(y)\big|.$$

So the mysterious "absolute value of the derivative of the inverse" in the transformation theorem is nothing more than the derivative-of-inverse formula from Calculus I.

16.5 Transformation of Random Variables: Examples

Example 16.3 (Logarithmic transformation)**.** Let X be uniformly distributed on $(1, e)$. Define $Y = \log(X)$.

Then $X = e^Y$ with derivative $dX/dY = e^Y$. The pdf of Y is

$$f_Y(y) = f_X(e^y) \cdot e^y.$$

Since $f_X(x) = 1/(e-1)$ for $1 < x < e$, we get

$$f_Y(y) = \frac{e^y}{e-1}, \qquad 0 < y < 1.$$

So Y is supported on $(0, 1)$ with linear density in e^y.

Example 16.4 (Square root transformation)**.** Let $X \sim \text{Exp}(1)$ with pdf $f_X(x) = e^{-x}, x > 0$. Define $Y = \sqrt{X}$. Then $X = Y^2$, so $dx/dy = 2y$. Thus

$$f_Y(y) = f_X(y^2) \cdot 2y = e^{-y^2} \cdot 2y, \qquad y > 0.$$

This is the well-known *Rayleigh distribution* used in communications.

Example 16.5 (Reciprocal transformation)**.** Suppose $X \sim \text{Exp}(\lambda)$. Define $Y = 1/X$. Then $X = 1/Y$, $dx/dy = -1/y^2$. So

$$f_Y(y) = f_X(1/y) \cdot \frac{1}{y^2} = \lambda e^{-\lambda/y} \cdot \frac{1}{y^2}, \qquad y > 0.$$

This distribution arises in areas like extreme value theory and finance (rates of return).

16.6 When the function is not one-to-one

What if $Y = g(X)$ and g is not one-to-one? The formula above can fail if we apply it blindly.

Example 16.6 (Counterexample:)**.** Suppose $Y = X^2$ with $X \sim U(-1, 1)$.

Naively: $X = \sqrt{Y}$, $dX/dY = 1/(2\sqrt{Y})$. Then

$$f_Y(y) = f_X(\sqrt{y}) \cdot \frac{1}{2\sqrt{y}} = \frac{1}{2} \cdot \frac{1}{2\sqrt{y}},$$

which suggests a pdf that does not integrate to 1 (and also misses mass).

The problem: $g(x) = x^2$ is not one-to-one on $(-1, 1)$.

Remedy

Decompose the domain into monotone pieces:

$$(-1,1) = (-1,0) \cup (0,1).$$

On each, x^2 is monotone. So we consider both branches $x = \pm\sqrt{y}$. Then

$$f_Y(y) = f_X(\sqrt{y}) \cdot \frac{1}{2\sqrt{y}} + f_X(-\sqrt{y}) \cdot \frac{1}{2\sqrt{y}},$$

for $0 < y < 1$. Since $f_X(x) = 1/2$ for $x \in (-1,1)$, we obtain

$$f_Y(y) = \frac{1}{2\sqrt{y}}, \qquad 0 < y < 1,$$

which indeed integrates to 1. This is the correct distribution of X^2 when X is uniform on $(-1,1)$.

On the existence of an inverse function

- **Necessary and sufficient condition:** g must be *one-to-one* (injective) and continuous on the support of X. Then g is strictly monotone, so g^{-1} exists and is continuous on the image (the range of g).
- If g is not one-to-one, then $h = g^{-1}$ does not exist globally. In this case, we must decompose the domain of X into intervals where g is monotone. On each piece an inverse exists, and we sum contributions from all preimages[3] of y.

This explains why, in the "bad example" $Y = X^2$ with $X \sim U(-1,1)$, we had to split $(-1,1)$ into two monotone intervals $(-1,0)$ and $(0,1)$ and account for both preimages $\pm\sqrt{y}$.

Takeaway 16.1. Transformations of random variables allow us to model derived quantities, not just compute expectations. The change of variable formula is the tool to move from f_X to f_Y. Care must be taken when the function is not one-to-one: we must account for all preimages of y under g. This prepares us for the higher-dimensional Jacobian case next week.

Looking ahead, in higher dimensions this transformation formula generalizes to involve the Jacobian determinant (of the transformation), which we will cover next week. So, get ready for some Calculus IV review!

[3]In general, if $g : A \to B$ is a function and $y \in B$, the *preimage* of y under g is the set $g^{-1}(\{y\}) = \{x \in A : g(x) = y\}$. So when we say "both original x–values map to the same y," we mean that the preimage of y contains more than one element.

Part IV
Joint Probability Distributions

17 Lecture XVII: Joint Discrete Random Variables, Their Joint PMFs and Independence

17.1 Motivation

Up until now, we have studied random variables one at a time. But many real-world problems demand that we consider *multiple random variables together*.

Example 17.1 (Dice)**.** If we roll two dice, the outcome is not just a single number, but a pair (X, Y) where X is the outcome of die #1 and Y is the outcome of die #2. Questions like

$$P(X + Y = 7)$$

cannot be answered by considering X or Y alone. However, we **made** it work by brute-forcing the sample space.

Example 17.2 (Reliability: Joint Lifetimes)**.** Two machine components have random lifetimes X and Y.

- If the system is in series, it fails when the first component fails: $T = \min(X, Y)$.
- If the system is in parallel, it fails when both components fail: $T = \max(X, Y)$.

In either case, knowing only the individual distribution of X or Y is not enough. We need the **joint distribution** to study the lifetime of the system.

Example 17.3 (BMI)**.** Let X denote a person's height and Y their weight.

- Neither height nor weight alone captures the whole picture.
- Together, (X, Y) allows us to define derived variables such as

$$\text{BMI} = \frac{Y}{X^2}.$$

- Correlation between height and weight is also meaningful in medical and nutritional studies.

Example 17.4 (Stock Prices)**.** Suppose X is the return on stock A and Y is the return on stock B.

- Portfolio risk is driven by the variance of $aX + bY$, where a, b are weights.
- This variance involves the covariance $\mathrm{Cov}(X, Y)$, which only makes sense once the **joint distribution** is defined.
- Modeling X and Y separately is insufficient to capture co-movement risk (for example, the influence of one stock on the other and thereby on your financial portfolio).

Example 17.5 (Other Scientific Practices)**.**

- In engineering, signal and noise are modeled as random variables, and system performance depends on their combined behavior.
- **Meteorology (Rainfall and Temperature).** X = daily rainfall, Y = daily temperature. Agriculture depends on the *joint* behavior: crops need both sufficient rain and mild temperatures. For example, probabilities such as
$$P(X > 20 \text{ mm},\ Y < 30℃)$$
cannot be answered from marginals alone.
- **Neuroscience (Spike Timing and Amplitude).** X = inter-spike interval, Y = spike amplitude. Studying each marginal separately misses important dependencies: stronger spikes might follow longer pauses. The joint distribution is essential to capture this relationship.
- **Ecology (Predator–Prey Populations).** X = prey population, Y = predator population. System stability and ecological events depend on correlations between X and Y. Questions such as
$$P(X > 500,\ Y < 50)$$
are meaningful only in the joint setting.

A single-variable framework fails when the phenomena naturally involve *pairs* or even larger collections of random quantities. We need to describe their **joint behavior**.

17.2 Joint PMF (discrete case)

If X, Y are discrete, the **joint probability mass function (pmf)** is
$$p_{X,Y}(x, y) = P(X = x, Y = y).$$

- $p_{X,Y}(x, y) \geq 0$ for all (x, y).
- $\sum_x \sum_y p_{X,Y}(x, y) = 1$.

17.2.1 Examples with constant PMF

Example 17.6 (Dice Problem)**.** Let X and Y be the outcomes of rolling two fair six-sided dice. The joint PMF is

$$p_{X,Y}(x,y) = P(X = x, Y = y) = \frac{1}{36}, \quad x, y \in \{1,2,3,4,5,6\}.$$

1. Compute $P(X + Y = 7)$.

   Solution.

   $$P(X + Y = 7) = \sum_{x=1}^{6} P(X = x, Y = 7 - x) = \frac{6}{36} = \frac{1}{6}.$$

   Remark. *We are lucky here because we only have rolled it twice. Imagine how ugly the sum becomes if we roll it three times?*

2. Compute $P(X + Y > 8)$.

   Solution. Using the joint PMF,

   $$P(X + Y > 8) = \sum_{\substack{x=1,\dots,6 \\ y=1,\dots,6 \\ x+y>8}} P(X = x, Y = y) = \sum_{\substack{x=1,\dots,6 \\ y=1,\dots,6 \\ x+y>8}} \frac{1}{36}.$$

   Count the number of pairs (x, y) such that $x + y > 8$:

   $$(3,6), (4,5), (4,6), (5,4), (5,5), (5,6), (6,3), (6,4), (6,5), (6,6)$$

   There are 10 such outcomes. Therefore,

   $$P(X + Y > 8) = \frac{10}{36} = \frac{5}{18}.$$

   Remark. *We are again super lucky here because we only have rolled it twice, and we are capable of brute-forcing the sample space, with a constant pmf. Imagine how ugly the sum becomes if we roll it three times?*

Example 17.7 (Urn problem)**.** An urn contains 3 red balls and 2 blue balls. Two balls are drawn without replacement. Let X = number of red balls drawn, Y = number of blue balls drawn. The joint PMF is

$$p_{X,Y}(x,y) = P(X = x, Y = y) = \begin{cases} \frac{6}{10}, & (x,y) = (1,1), \\ \frac{3}{10}, & (x,y) = (2,0), \\ \frac{1}{10}, & (x,y) = (0,2), \\ 0, & \text{otherwise.} \end{cases}$$

Question 17.1. I am delegating to you to understand how each probability is determined. We have learned it this term.

17.2.2 Examples with variable PMF

Example 17.8 (Weighted dice-like random variables)**.** Suppose X and Y take values in $\{1, 2, 3\}$, but not all outcomes are equally likely. Let the joint PMF be defined by:

$$p_{X,Y}(x, y) = \frac{x \cdot y}{36}, \quad x, y \in \{1, 2, 3\}.$$

Compute $P(X + Y \geq 4)$, and sketch the joint PMF as a table.

Solution.

$$\begin{aligned} P(X + Y \geq 4) &= \sum_{\substack{x,y \in \{1,2,3\} \\ x+y \geq 4}} p_{X,Y}(x, y) \\ &= \frac{2 \cdot 2 + 2 \cdot 3 + 3 \cdot 1 + 1 \cdot 3 + 3 \cdot 2 + 3 \cdot 3}{36} \\ &= \frac{4 + 6 + 3 + 3 + 6 + 9}{36} \\ &= \frac{31}{36}. \end{aligned}$$

Joint PMF Table: ($X + Y \geq 4$ are **bolded**)

$y = 3$	$\mathbf{\frac{3}{36}}$	$\mathbf{\frac{6}{36}}$	$\mathbf{\frac{9}{36}}$
$y = 2$	$\frac{2}{36}$	$\mathbf{\frac{4}{36}}$	$\mathbf{\frac{6}{36}}$
$y = 1$	$\frac{1}{36}$	$\frac{2}{36}$	$\mathbf{\frac{3}{36}}$
	$x = 1$	$x = 2$	$x = 3$

- Bottom row labels the x-values increasing left to right.
- Left column labels the y-values decreasing top to bottom.
- Each cell contains $p_{X,Y}(x, y)$, corresponding to the probability of that outcome.

17.3 Marginal PMFs

We have now seen joint PMFs with both constant and non-constant probabilities. But often, our interest lies in only one of the two random variables, even if they were defined jointly. For instance, in the non-uniform example we just studied, suppose

we only care about the distribution of X. How do we "extract" it from the joint distribution?

This motivates the idea of a *marginal distribution*. The name comes from a probability table: if you sum along rows or columns and record the totals along the margins, you obtain the marginal probabilities.

Definition 17.1 (Marginal PMF)**.** *If (X, Y) has joint PMF $p_{X,Y}(x, y)$, then the* marginal PMF *of X is*

$$p_X(x) = \sum_y p_{X,Y}(x, y),$$

and the marginal PMF *of Y is*

$$p_Y(y) = \sum_x p_{X,Y}(x, y).$$

Example 17.9 (Weighted dice-like random variables cont.)**.** Recall our joint PMF

$$p_{X,Y}(x, y) = \frac{x \cdot y}{36}, \quad x, y \in \{1, 2, 3\}.$$

The marginal PMF of X is

$$p_X(x) = \sum_{y=1}^{3} \frac{x \cdot y}{36} = \frac{x(1+2+3)}{36} = \frac{6x}{36},$$

so

$$p_X(1) = \tfrac{1}{6}, \quad p_X(2) = \tfrac{1}{3}, \quad p_X(3) = \tfrac{1}{2}.$$

Similarly, for Y,

$$p_Y(y) = \sum_{x=1}^{3} \frac{x \cdot y}{36} = \frac{y(1+2+3)}{36} = \frac{6y}{36},$$

so

$$p_Y(1) = \tfrac{1}{6}, \quad p_Y(2) = \tfrac{1}{3}, \quad p_Y(3) = \tfrac{1}{2}.$$

Thus, although the joint distribution was two-dimensional, we now know the standalone distribution of X and of Y.

Important Remark 17.1 (Why marginals?)**.** *Marginal distributions allow us to focus on a single variable while still working from the joint distribution (which is typical in practice). They are the projections of the two-dimensional distribution onto one axis. They are crucial because:*

- *They describe what you see if you observe only one variable.*
- *They form the basis for checking* independence*: does $p_{X,Y}(x, y)$ factor into $p_X(x)p_Y(y)$?*
- *They are the building blocks for expectations and variances of each variable, which we will study next week.*

17.4 Independence

We now return to an important structural property: independence, now from the perspective of joint random variables.

Independence of Events **Recall** that for events A and B, we say they are independent if

$$P(A \mid B) = P(A).$$

In words: knowing that B occurred does not change the probability of A. Equivalently, we have derived the multiplication rule to help us identify independence based on the conditional probability definition,

$$P(A \cap B) = P(A)P(B).$$

Definition 17.2 (Independence of Random Variables)**.** *For discrete random variables X, Y, we say X and Y are independent if for all values x, y,*

$$P(X = x \text{ and } Y = y) = P(X = x)P(Y = y).$$

That is, the joint probability of (X, Y) splits into the product of the marginals.

This condition is exactly the discrete version of the multiplication rule. If the joint PMF can be factored into a product of the marginals, then neither variable carries any information about the other.

Example 17.10 (Weighted dice-like random variables cont.)**.** Recall the joint PMF

$$p_{X,Y}(x, y) = \frac{x \cdot y}{36}, \quad x, y \in \{1, 2, 3\}.$$

Earlier we found

$$p_X(x) = \frac{x}{6}, \qquad p_Y(y) = \frac{y}{6}.$$

Notice that

$$p_X(x)p_Y(y) = \frac{x}{6} \cdot \frac{y}{6} = \frac{xy}{36} = p_{X,Y}(x, y).$$

So indeed, X and Y are independent.

Example 17.11 (Non-example: **Dependence**)**.** Roll two fair dice D_1, D_2. Let

$$X = D_1, \qquad Y = D_1 + D_2.$$

Then (X, Y) takes values with joint PMF

$$p_{X,Y}(x, y) = P(D_1 = x,\ D_1 + D_2 = y).$$

For example, $p_{X,Y}(3, 8) = P(D_1 = 3, D_2 = 5) = \frac{1}{36}$.

Since all 36 dice pairs are equally likely,

$$p_{X,Y}(x, y) = \begin{cases} \frac{1}{36}, & \text{if } x = 1, \ldots, 6,\ y = x + 1, \ldots, x + 6, \\ 0, & \text{otherwise,} \end{cases}$$

where you must note that the range of y depends on x. But is this dependence probabilistic? We need to check the marginals.

The marginal distributions are

$$p_X(x) = \tfrac{1}{6}, \quad x = 1, \ldots, 6,$$

and

$$p_Y(y) = \frac{\#\{(d_1, d_2) : d_1 + d_2 = y\}}{36}, \quad y = 2, \ldots, 12.$$

If X and Y were independent, we would have

$$p_{X,Y}(3, 8) \stackrel{?}{=} p_X(3)\, p_Y(8).$$

But in fact

$$p_{X,Y}(3, 8) = \tfrac{1}{36}, \quad p_X(3)\, p_Y(8) = \tfrac{1}{6} \cdot \tfrac{5}{36} = \tfrac{5}{216}.$$

These are not equal. Thus X and Y are dependent: knowing X restricts the range of possible values of Y, **as expected**.

Takeaway 17.1 (Why Independence Matters)**.** Independence is a powerful simplification. If two random variables are independent, then

- The joint distribution is entirely determined by the marginals.
- Many probability calculations (expectations, variances) simplify dramatically under independence (next week).
- In practice, independence assumptions are often the first model we try.

17.5 Joint CDF

If (X, Y) are discrete random variables, their joint cumulative distribution function (CDF) is defined by

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

In the discrete case, this can be expressed in terms of the joint pmf $p_{X,Y}(u, v)$ as

$$F_{X,Y}(x, y) = \sum_{u \leq x} \sum_{v \leq y} p_{X,Y}(u, v).$$

Example 17.12. Suppose X and Y are independent and each uniformly distributed on $\{1, 2\}$. Then the joint pmf is

$$p_{X,Y}(x, y) = \begin{cases} \frac{1}{4}, & x, y \in \{1, 2\}, \\ 0, & \text{otherwise.} \end{cases}$$

- At $(x, y) = (1, 1)$:

$$F_{X,Y}(1, 1) = \sum_{u \leq 1} \sum_{v \leq 1} p_{X,Y}(u, v) = p_{X,Y}(1, 1) = \tfrac{1}{4}.$$

- At $(x, y) = (1, 2)$:

$$F_{X,Y}(1, 2) = \sum_{u \leq 1} \sum_{v \leq 2} p_{X,Y}(u, v) = p_{X,Y}(1, 1) + p_{X,Y}(1, 2) = \tfrac{1}{4} + \tfrac{1}{4} = \tfrac{1}{2}.$$

- At $(x, y) = (2, 1)$:

$$F_{X,Y}(2, 1) = \sum_{u \leq 2} \sum_{v \leq 1} p_{X,Y}(u, v) = p_{X,Y}(1, 1) + p_{X,Y}(2, 1) = \tfrac{1}{4} + \tfrac{1}{4} = \tfrac{1}{2}.$$

- At $(x, y) = (2, 2)$:

$$F_{X,Y}(2, 2) = \sum_{u \leq 2} \sum_{v \leq 2} p_{X,Y}(u, v) = 1.$$

This illustrates how the double summation "accumulates" probabilities in the lower-left rectangle of the support.

18 Lecture XVIII: Joint Continuous Random Variables

So far, we have worked with **discrete** joint random variables, where probabilities are described by a joint pmf. We saw how to compute marginals, define independence, and express the joint CDF.

The natural next step is to ask: what changes when our random variables are **continuous**?

The answer is: almost nothing in principle.

- The joint pmf is replaced by a *joint pdf*.
- Sums over discrete supports become integrals over continuous domains.
- Independence is still defined in terms of events, but turns out to be equivalent to factorization of the joint pdf or cdf, just as before.

This is why we studied the discrete case carefully: it gave us intuition about joint structure and independence. With that experience, the continuous case will feel like a direct analogy, and we can move more quickly through the basics, focusing on practice with calculations (with warm reminders of your integration skills).

18.1 Joint PDF (continuous case)

If X, Y are continuous, the **joint probability density function (pdf)** $f_{X,Y}(x, y)$ is defined such that

$$P\big((X, Y) \in A\big) = \iint_A f_{X,Y}(x, y)\, dx\, dy,$$

for any region $A \subseteq \mathbb{R}^2$.

Important Remark 18.1 (Properties)**.**

- $f_{X,Y}(x, y) \geq 0$ *for all* (x, y).
- $\iint_{\mathbb{R}^2} f_{X,Y}(x, y)\, dx\, dy = 1$.

18.2 Joint CDF

Let X and Y be random variables. The **joint cumulative distribution function (CDF)** is defined by

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

Important Remark 18.2 (Properties)**.**

- $F_{X,Y}(x, y)$ *is nondecreasing in both* x *and* y*.*
- $\lim_{x,y\to-\infty} F_{X,Y}(x, y) = 0$*.*
- $\lim_{x,y\to\infty} F_{X,Y}(x, y) = 1$*.*

18.3 Marginal Densities

Just as in the discrete case, where we sum out one variable to obtain the marginal pmf, here, we **integrate** out one variable to obtain the **marginal densities** of X and Y.

- Recall that if (X, Y) is **discrete** with pmf $p_{X,Y}$:

$$p_X(x) = \sum_y p_{X,Y}(x, y), \quad p_Y(y) = \sum_x p_{X,Y}(x, y).$$

- If (X, Y) is **continuous** with pdf $f_{X,Y}$:

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x, y)\, dy, \quad f_Y(y) = \int_{-\infty}^{\infty} f_{X,Y}(x, y)\, dx.$$

18.3.1 Independence

Proposition 18.1 (Independence via PDFs)**.** *In a similar setting to joint discrete random variables, we say that two continuous random variables* X *and* Y *are independent if for all values* x, y*,*

$$f_{X,Y}(x, y) = f_X(x) f_Y(y).$$

That is, the joint probability density of (X, Y) *splits into the product of the marginal densities.*

Equivalently, we can also get infer independence from the CDFs.

Proposition 18.2. X *and* Y *are independent if and only if*

$$F_{X,Y}(x, y) = F_X(x)\, F_Y(y), \quad \textit{for all } x, y.$$

18.4 Easier Examples

Example 18.1 (Independent Uniform Distributions)**.** Suppose X and Y are independent and each uniformly distributed on $(0, 1)$.

- The joint pdf is

$$f_{X,Y}(x, y) = \begin{cases} 1, & 0 < x < 1,\ 0 < y < 1, \\ 0, & \text{otherwise.} \end{cases}$$

- The marginal pdfs are $f_X(x) = 1$ for $0 < x < 1$, $f_Y(y) = 1$ for $0 < y < 1$.
- Independence is clear since $f_{X,Y}(x, y) = f_X(x) f_Y(y)$.
- Probability calculation:

$$P(X < \tfrac{1}{2}, Y < \tfrac{1}{3}) = \iint_{[0,1/2]\times[0,1/3]} 1\, dx\, dy = \tfrac{1}{6}.$$

- Joint CDF:

$$F_{X,Y}(x, y) = \begin{cases} xy, & 0 < x < 1,\ 0 < y < 1, \\ \text{extend by 0 or 1}, & \text{outside.} \end{cases}$$

Example 18.2 (Joint Uniform on the Unit Square (not independent))**.** Suppose (X, Y) is uniformly distributed on the triangle

$$T = \{(x, y) : 0 < x < 1,\ 0 < y < 1,\ x + y < 1\}.$$

- Area of T is $1/2$, so joint pdf is

$$f_{X,Y}(x, y) = \begin{cases} 2, & (x, y) \in T, \\ 0, & \text{otherwise.} \end{cases}$$

- Marginal of X:

$$f_X(x) = \int_0^{1-x} 2\, dy = 2(1 - x), \quad 0 < x < 1.$$

- Marginal of Y:

$$f_Y(y) = \int_0^{1-y} 2\, dx = 2(1 - y), \quad 0 < y < 1.$$

- Independence check: $f_X(x)f_Y(y) = 4(1-x)(1-y)$, not equal to 2 on T. So X, Y are not independent.
- Example probability:

$$P(X < 1/2, Y < 1/2) = \iint_{[0,1/2]^2 \cap T} 2\,dx\,dy.$$

Region of integration is the square $(0, 1/2) \times (0, 1/2)$ (since $x + y < 1$ holds automatically). Thus probability is

$$2 \cdot \left(\tfrac{1}{2}\right)\left(\tfrac{1}{2}\right) = \tfrac{1}{2}.$$

which is precisely the area formed by the intersection of the square $[0, 1/2]^2$ and the triangle T, as expected since this is a joint uniform random variable.

18.5 Harder Examples with Integration Techniques

In the following examples, we introduce the **most important technique** in joint continuous random variables: integration over regions with variable bounds.

Example 18.3 (Independent Exponential Random Variables)**.** Suppose X is an exponential random variable that models the waiting time for a bus, with rate $\lambda = 1$. What is the probability that the second bus arrives in 3 minutes?

Solution. Wait, single variable? I thought we are working with joint random variables! But the question is asking for it: it is the SECOND bus, not the first. Therefore, we need to define Y that has the same distribution as X **independently**, and the probability of interest is

$$P(X + Y < 3).$$

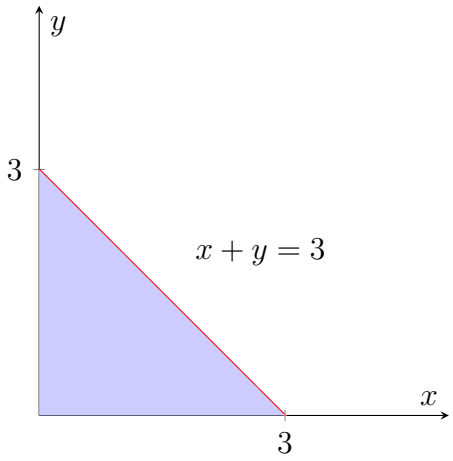
Since X and Y are independent, the joint pdf is

$$f_{X,Y}(x, y) = e^{-(x+y)}, \quad x, y > 0.$$

So

$$P(X + Y < 3) = \iint_{x+y<3,\ x,y<0} e^{-(x+y)}\,dx\,dy.$$

Method 1: Variable bounds of integration.
First, we need to sketch this region:



We integrate against y first, from 0 up to the curve $y = 3 - x$, and then x from 0 to 3.

$$\begin{aligned}
\iint_{\{x+y<3, \quad x,y>0\}} e^{-(x+y)} dA &= \int_0^3 \int_0^{3-x} e^{-(x+y)} dy dx \\
&= \int_0^3 e^{-x} \left(\int_0^{3-x} e^{-y} dy \right) dx \\
&= \int_0^3 e^{-x} \left[1 - e^{-(3-x)} \right] dx \\
&= \int_0^3 \left(e^{-x} - e^{-3} \right) dx \\
&= \left[-e^{-x} \right]_0^3 - 3e^{-3} \\
&= 1 - e^{-3} - 3e^{-3} \\
&= 1 - 4e^{-3} \approx 0.80085.
\end{aligned}$$

Method 2: Change of variable. You can try it now, but we will also learn it next week. Treat this one as a review of the integration technique.

Let $u = x + y$, $v = x$, $|J(u, v)| = 1$. Then

$$P(X + Y < 3) = \int_0^3 \int_0^u e^{-u} \, dv \, du = \int_0^3 u e^{-u} \, du.$$

Integration by parts? You do it.

Example 18.4 (Quality Control in Practice)**.** Suppose X and Y are independent uniform random variables on $[0, 1]$, representing the normalized defect sizes of two parts. A part is rejected if $X > 2Y$. What is the probability of rejection?

Solution. The joint pdf is

$$f_{X,Y}(x,y) = 1, \quad 0 < x < 1,\ 0 < y < 1.$$

The region is

$$R = \{(x,y) : 0 < x < 1,\ 0 < y < 1,\ x > 2y\}.$$

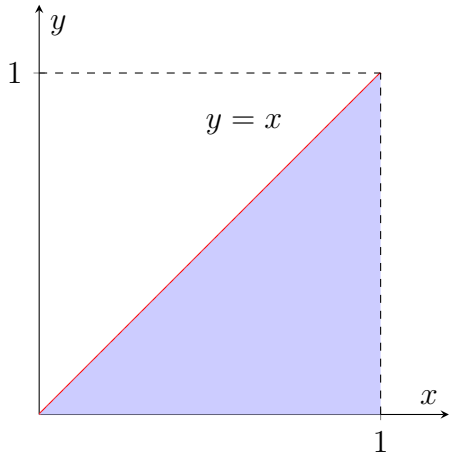
That is, $y < x/2$. Hence

$$P(X > 2Y) = \int_0^1 \int_0^{x/2} 1\, dy\, dx = \int_0^1 \frac{x}{2}\, dx = \frac{1}{4}.$$

Example 18.5 (Given Joint PDF, Check Independence)**.** Assume that the joint density of X and Y is given by

$$f_{X,Y}(x,y) = \begin{cases} 3x, & 0 \le x < 1, \quad 0 \le y < x \\ 0, & \text{otherwise} \end{cases}$$

Solution. We calculate the marginals and check if their product yields the joint density. The very first thing to do here is to sketch the region.



$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)\, dy = \begin{cases} \int_0^x 3x dy = 3x^2, & 0 \le x < 1 \\ 0, & \text{otherwise} \end{cases}$$

and

$$f_Y(y) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)\, dx = \begin{cases} \int_y^1 3x dx = \frac{3}{2}(1-y^2), & 0 \le y < 1 \\ 0, & \text{otherwise} \end{cases}$$

Take a **BIG NOTE** of the bounds of integration here, for each marginal.

Their product yields

$$f_X(x) \cdot f_Y(y) = \begin{cases} \frac{9}{2}x^2(1-y^2), & 0 \le x < 1, \quad 0 \le y < x \\ 0, & \text{otherwise} \end{cases}$$

which is obviouly not equal to the joint density. Therefore, X and Y are NOT independent.

18.6 Why Joint Distributions Matter

With joint distributions, we can:

- Compute probabilities of events involving multiple variables, such as $P(X+Y < 10)$.
- Define conditional distributions, e.g.

  $$f_{X|Y}(x|y) = \frac{f_{X,Y}(x,y)}{f_Y(y)},$$

  which will be crucial when we study dependence and Markov chains (next class and the final week).
- Study dependence structures: independence is characterized by

  $$f_{X,Y}(x,y) = f_X(x)f_Y(y) \quad \text{for all } x, y.$$

18.7 Looking Ahead

We introduced joint random variables and their distributions. This is the foundation for:

- Conditional probabilities and conditional expectation.
- Covariance and correlation.
- Transformations in multiple dimensions, which require Jacobians (next week).

18.8 Important Summary of Joint Random Variables

Concept	Discrete Case	Continuous Case
Joint law	pmf: $p_{X,Y}(x,y) = P(X=x, Y=y)$	pdf: $f_{X,Y}(x,y) \geq 0$, $\iint f_{X,Y} = 1$
Marginals	$p_X(x) = \sum_y p_{X,Y}(x,y)$ (sim. for Y)	$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)\,dy$ (sim. for Y)
Joint CDF	$F_{X,Y}(x,y) = \sum_{u\leq x}\sum_{v\leq y} p_{X,Y}(u,v)$	$F_{X,Y}(x,y) = \int_{-\infty}^{x}\int_{-\infty}^{y} f_{X,Y}(u,v)\,dv\,du$
Probabilities	$P((X,Y)\in A) = \sum_{(x,y)\in A} p_{X,Y}(x,y)$	$P((X,Y)\in A) = \iint_A f_{X,Y}(x,y)\,dx\,dy$
Independence	$p_{X,Y}(x,y) = p_X(x)p_Y(y)$	$f_{X,Y}(x,y) = f_X(x)f_Y(y)$

Takeaway 18.1. The moral: the structures are *parallel*. What was a sum in the discrete case becomes an integral in the continuous case, and all the ideas—marginals, CDFs, independence—carry over directly.

19 Lecture XIXa: Expected Value and Covariance in Multivariate Settings

Much of what we do in this lecture is a natural extension of the single random variable case. The expected value of a **function** of (X, Y) follows the same spirit as before: it is a weighted average, with the joint distribution providing the weights. In fact, many properties such as linearity carry over without surprise.

Where some care is required is in the notion of *covariance*. Just as variance measures the spread of a single random variable around its mean, covariance measures how two random variables vary *together*. The definition looks very similar to the single-variable case, but the interpretation is new and nuanced, making covariance one of **the most fundamentally important** statistical quantities.

If the concepts of expectation and variance for one random variable are clear, then the step to covariance and correlation will be quite straightforward.

19.1 Multivariate Expectation

Definition 19.1 (Expectation of a Function of Two Random Variables)**.** *Let (X, Y) be a pair of random variables and let $h : \mathbb{R}^2 \to \mathbb{R}$.*

- *If (X, Y) is discrete with joint pmf $p_{X,Y}(x, y)$, then*

$$\mathbb{E}[h(X, Y)] = \sum_x \sum_y h(x, y)\, p_{X,Y}(x, y).$$

- *If (X, Y) is continuous with joint pdf $f_{X,Y}(x, y)$, then*

$$\mathbb{E}[h(X, Y)] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} h(x, y)\, f_{X,Y}(x, y)\, dx\, dy.$$

Important Remark 19.1 (Big Red Flag)**.** *It makes no sense to compute something like $\mathbb{E}[(X, Y)]$. The expectation operator is applied to* functions *of the random variables. That is, we focus on expectations of the form $\mathbb{E}[h(X, Y)]$.*

19.2 Linearity of Expectation

Proposition 19.1 (Linearity)**.** *For random variables X, Y and constants a, b, and multivariable functions h_1 and h_2,*

$$\mathbb{E}[a_1 h_1(X, Y) + a_2 h_2(X, Y) + b] = a_1\, \mathbb{E}[h_1(X, Y)] + a_2\, \mathbb{E}[h_2(X, Y)] + b.$$

19.3 Independence and Expected Value

Theorem 19.1. *If X and Y are independent random variables, then*

$$\mathbb{E}[XY] = \mathbb{E}[X] \cdot \mathbb{E}[Y]. \tag{19.1}$$

Remark 19.1. *Note that the LHS of Equation (19.1) is given by the double integral involving the joint density*

$$\mathbb{E}[XY] = \iint_{\mathbb{R}^2} xy f_{X,Y}(x,y)\, dA$$

while each expected value on the RHS just uses the marginal densities of X and Y respectively.

19.4 Covariance

Definition 19.2 (Covariance)**.** *Let X, Y be random variables with finite mean.*

- *In the discrete case,*

$$\mathrm{Cov}(X,Y) = \sum_x \sum_y (x-\mu_X)(y-\mu_Y)\, p_{X,Y}(x,y).$$

- *In the continuous case,*

$$\mathrm{Cov}(X,Y) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (x-\mu_X)(y-\mu_Y)\, f_{X,Y}(x,y)\, dx\, dy.$$

Remark 19.2. *Notice that this is simply the expectation of the function $h(x,y) = (x-\mu_X)(y-\mu_Y)$. The mechanics of computing covariance are not difficult; the interpretation is the crucial part.*

Interpretation via Examples

The three diagrams in Fig. 4.4 (textbook, p. 258) illustrate:

- **Positive covariance:** when X increases, Y tends to increase.
- **Zero covariance:** no consistent linear relationship between X and Y.
- **Negative covariance:** when X increases, Y tends to decrease.

Proposition 19.2 (Covariance Shortcut Formula)**.** *Covariance can also be expressed as*

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

Example 19.1 (Computing Covariance from a Joint Density)**.** Let (X,Y) have joint pdf

$$f_{X,Y}(x,y) = \begin{cases} 2, & 0 < x < 1,\ 0 < y < 1-x, \\ 0, & \text{otherwise}, \end{cases}$$

i.e. uniform on the triangle $T = \{(x,y) : 0 < x < 1,\ 0 < y < 1-x\}$. ($\mathrm{Area}(T) = 1/2$, so the density is 2.)

We compute $\mathrm{Cov}(X,Y)$ following the standard steps.

1. Find the marginals.

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)\,dy = \int_0^{1-x} 2\,dy = 2(1-x), \qquad 0 < x < 1.$$

By symmetry,

$$f_Y(y) = \int_0^{1-y} 2\,dx = 2(1-y), \qquad 0 < y < 1.$$

2. Find the means μ_X, μ_Y.

$$\mu_X = \int_0^1 x f_X(x)\,dx = \int_0^1 x\cdot 2(1-x)\,dx = \int_0^1 (2x-2x^2)\,dx = \Big[x^2 - \tfrac{2}{3}x^3\Big]_0^1 = 1 - \tfrac{2}{3} = \tfrac{1}{3}.$$

By symmetry $\mu_Y = \frac{1}{3}$.

3. Compute $\mathbb{E}[XY]$.

$$\mathbb{E}[XY] = \iint_T xy \cdot 2\,dx\,dy = 2\int_{x=0}^1 \Big(\int_{y=0}^{1-x} xy\,dy\Big)dx.$$

Evaluate the inner integral:

$$\int_0^{1-x} xy\,dy = x\cdot\frac{(1-x)^2}{2}.$$

Thus

$$\mathbb{E}[XY] = 2\int_0^1 x\cdot\frac{(1-x)^2}{2}\,dx = \int_0^1 x(1-2x+x^2)\,dx = \int_0^1 (x-2x^2+x^3)\,dx.$$

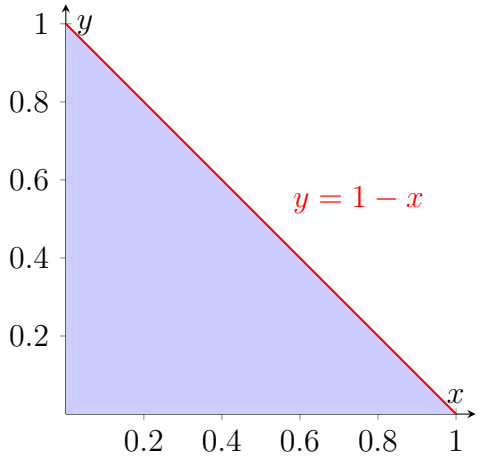
Compute the antiderivative:

$$\mathbb{E}[XY] = \Big[\tfrac{1}{2}x^2 - \tfrac{2}{3}x^3 + \tfrac{1}{4}x^4\Big]_0^1 = \tfrac{1}{2} - \tfrac{2}{3} + \tfrac{1}{4} = \frac{6-8+3}{12} = \frac{1}{12}.$$

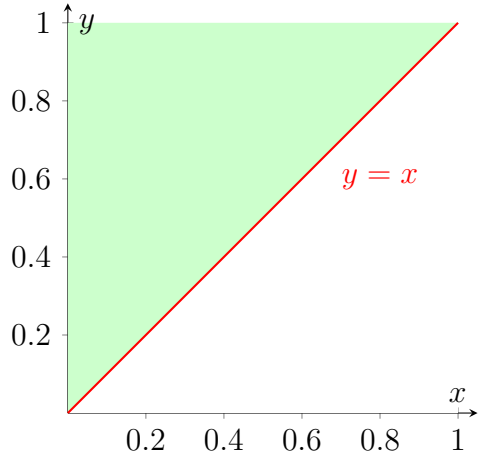
4. Use the shortcut formula for covariance.

$$\mathrm{Cov}(X,Y) = \mathbb{E}[XY] - \mu_X\mu_Y = \frac{1}{12} - \frac{1}{3}\cdot\frac{1}{3} = \frac{1}{12} - \frac{1}{9} = -\frac{1}{36}.$$

Interpretation. The negative covariance (and correlation $-\frac{1}{2}$) reflects the geometry of the support (see Example 19.1): when x is large (near 1) the allowed values of y are small (since $y < 1-x$), so X and Y tend to move in opposite directions (but not all the time). This produces negative linear association even though the marginals are identical.



(a) Support $T = \{(x,y) : 0 < x < 1,\ 0 < y < 1-x\}$ for the negative-covariance Example 19.1.



(b) Support $T' = \{(x,y) : 0 < x < 1,\ x < y < 1\}$ for the positive-covariance Example 19.2.

Example 19.2 (Computing Covariance from a Joint Density: Positive Covariance)**.** We here do a similar example where the density is slightly modified from the last one.

Let (X,Y) have joint pdf

$$f_{X,Y}(x,y) = \begin{cases} 2, & 0 < x < 1,\ x < y < 1, \\ 0, & \text{otherwise}, \end{cases}$$

i.e. uniform on the upper triangle $T' = \{(x,y) : 0 < x < 1,\ x < y < 1\}$ (area $= 1/2$ so density $= 2$).

1. Marginals.

$$f_X(x) = \int_{y=x}^{1} 2\,dy = 2(1-x), \qquad 0 < x < 1,$$

$$f_Y(y) = \int_{x=0}^{y} 2\,dx = 2y, \qquad 0 < y < 1.$$

2. Means.

$$\mu_X = \int_0^1 x \cdot 2(1-x)\,dx = \frac{1}{3}, \qquad \mu_Y = \int_0^1 y \cdot 2y\,dy = \frac{2}{3}.$$

3. Compute $\mathbb{E}[XY]$.

$$\mathbb{E}[XY] = \iint_{T'} xy \cdot 2\,dx\,dy = 2\int_{x=0}^{1}\left(\int_{y=x}^{1} xy\,dy\right)dx = 2\int_0^1 x \cdot \frac{1-x^2}{2}\,dx = \int_0^1 (x - x^3)\,dx = \frac{1}{4}.$$

4. Covariance.

$$\mathrm{Cov}(X,Y) = \mathbb{E}[XY] - \mu_X\mu_Y = \frac{1}{4} - \frac{1}{3}\cdot\frac{2}{3} = \frac{1}{36}.$$

Interpretation. The positive covariance reflects the geometry: larger values of x force y to be larger (since $y > x$), so X and Y tend to increase together. See Figure 6b for the sketch of the density.

Takeaway 19.1 (Steps to Compute Covariance from a Joint Density)**.** Suppose $f_{X,Y}(x,y)$ is a known joint pdf. To compute $\mathrm{Cov}(X,Y)$:

1. Find the marginal pdfs $f_X(x)$ and $f_Y(y)$.
2. Compute $\mu_X = \mathbb{E}[X]$ and $\mu_Y = \mathbb{E}[Y]$.
3. Compute $\mathbb{E}[XY] = \int\int xy\, f_{X,Y}(x,y)\,dx\,dy$.
4. Use the shortcut formula:
$$\mathrm{Cov}(X,Y) = \mathbb{E}[XY] - \mu_X\mu_Y.$$
5. Interpretation **must** be performed after computation.

19.5 Correlation

Definition 19.3 (Correlation)**.** *The correlation between X and Y is*

$$\rho_{X,Y} = \frac{\mathrm{Cov}(X,Y)}{\sigma_X \sigma_Y},$$

where σ_X, σ_Y are the standard deviations.

Example 19.3 (Last example cont.)**.** So, in the last example, we found the covariance. Here, to compute the correlation, we just need the individual variances from each variable.

Compute $\mathbb{E}[X^2]$:

$$\mathbb{E}[X^2] = \int_0^1 x^2 \cdot 2(1-x)\,dx = 2\int_0^1 (x^2 - x^3)\,dx = 2\left[\tfrac{1}{3} - \tfrac{1}{4}\right] = 2 \cdot \frac{1}{12} = \frac{1}{6}.$$

Hence

$$\mathrm{Var}(X) = \mathbb{E}[X^2] - \mu_X^2 = \frac{1}{6} - \frac{1}{9} = \frac{1}{18}.$$

By symmetry $\mathrm{Var}(Y) = \frac{1}{18}$.

Therefore the correlation is

$$\rho_{X,Y} = \frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{Var}(X)\mathrm{Var}(Y)}} = \frac{-1/36}{1/18} = -\tfrac{1}{2}.$$

Important Remark 19.2. *Correlation measures how well a linear relationship describes the dependence of Y on X; see Figure 7 in detail.*

- *Intuitively: If ρ is close to 1, X and Y move together in a strongly linear way. If ρ is close to -1, they move oppositely. If $\rho \approx 0$, their relationship is weakly linear or non-linear.*
- *Quantitatively: ρ is the normalized covariance, ensuring it always lies between -1 and 1. This standardization makes correlation scale-free and comparable across different problems.*

Corollary 19.1 (Independence implies zero covariance)**.** *If X and Y are independent, then $\mathrm{Cov}(X,Y) = 0$ and $\rho_{X,Y} = 0$.*

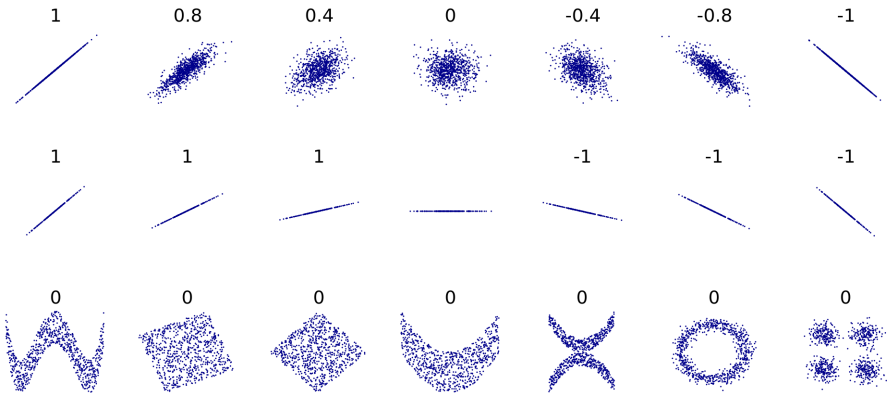


Figure 7: Various shape of xy-data with their correlation levels.

19.6 Zero Covariance Does NOT Imply Independence

Example 19.4 (A Non-example (discrete))**.** It is logically sound to note that Equation (19.1) (or zero covariance/correlation) is a **consequence** of independence, not the **cause** of it. In other words, if you happen to find the Equation (19.1) is satisfied, it does NOT mean the variables are independent. Here is a counterexample.

Let

$$X \in \{-1, 0, 1\}, \qquad P(X = -1) = \tfrac{1}{4},\ P(X = 0) = \tfrac{1}{2},\ P(X = 1) = \tfrac{1}{4},$$

and define $Y = X^2$ (so $Y \in \{0, 1\}$).

Compute expectations:

$$\mathbb{E}[X] = (-1)\tfrac{1}{4} + 0 \cdot \tfrac{1}{2} + 1 \cdot \tfrac{1}{4} = 0,$$

$$\mathbb{E}[XY] = \mathbb{E}[X^3] = (-1)^3\tfrac{1}{4} + 0 + 1^3\tfrac{1}{4} = 0.$$

Hence

$$\mathrm{Cov}(X, Y) = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y] = 0 - 0 \cdot \mathbb{E}[Y] = 0.$$

But X and Y are *not* independent. For example,

$$P(X = 1, Y = 1) = P(X = 1) = \tfrac{1}{4},$$

whereas

$$P(X = 1)\, P(Y = 1) = \tfrac{1}{4} \cdot \tfrac{1}{2} = \tfrac{1}{8},$$

so $P(X = 1, Y = 1) \neq P(X = 1)P(Y = 1)$. Thus dependence remains despite zero covariance.

Example 19.5 (A Non-example (continuous))**.** Let $X \sim N(0,1)$ be standard normal and set $Y = X^2$. Then

$$\mathbb{E}[X] = 0, \qquad \mathbb{E}[XY] = \mathbb{E}[X^3] = 0,$$

so $\mathrm{Cov}(X,Y) = 0$. However X and Y are not independent because Y determines $|X|$ (and hence gives information about X beyond its marginal law). For instance, knowing Y tells you the magnitude of X, so the joint law does not factorize.

Important Remark 19.3 (when zero covariance *does* imply independence)**.** *Zero covariance* $\mathrm{Cov}(X,Y) = 0$ does *imply independence in the special case when* (X,Y) *are jointly Gaussian: for bivariate normal vectors, uncorrelatedness* $\iff$ *independence. In general, however, uncorrelatedness is* ***strictly weaker*** *than independence.*

Lecture XIXb: Why Covariance/Correlation Informs about Linearity (and linearity only)

This partial lecture (with slides appended on Canvas) marks one of the most important intellectual questions about covariance. We saw in Figure 7 that if xy-data forms a line, then its correlation is either -1 or 1; meanwhile, when xy-data form something like a sine wave, a parabola, a hyperbola or a circle, correlation goes to 0 instead in all of these cases. Thus, we can surmise that correlation (and thus covariance)

1. Tells us how linearly dependent one variable is to another.
2. Fails to tell us how nonlinearly dependent one variable is to another.

In the following setup, we rediscover the definition of covariance (and thus correlation) from a much more practical and meaningful perspective, and thus helps us understand intuitively (and quantitatively) why the first point above makes sense.

Motivation. We often ask: if we try to predict a random variable Y from another variable X using a *linear* rule, what is the "best" slope? A natural optimality criterion is mean squared error: choose coefficients a, b to minimize

$$\mathbb{E}\big[(Y - aX - b)^2\big].$$

This is the population (not sample) least-squares problem. Solving it shows the covariance appears naturally in the numerator of the optimal slope — so covariance is exactly the quantity that drives the best linear fit.

Derivation (population least squares). Let $a, b \in \mathbb{R}$. Define the mean-square loss (just a multivariable function)

$$L(a, b) = \mathbb{E}\big[(Y - aX - b)^2\big].$$

We differentiate with respect to a and b and set derivatives to zero (these are convex quadratic problems so stationary point is the minimizer).

First, expand or compute partial derivatives directly using linearity of expectation.

Differentiate w.r.t. b:

$$\frac{\partial}{\partial b} L(a, b) = \mathbb{E}\big[2(Y - aX - b)(-1)\big] = -2\mathbb{E}[Y - aX - b].$$

Set to zero:

$$\mathbb{E}[Y - aX - b] = 0 \quad \Longrightarrow \quad b = \mathbb{E}[Y] - a\mathbb{E}[X] = \mu_Y - a\mu_X.$$

This is called **the centering result**: the best-fit line passes through the point of means (μ_X, μ_Y).

Differentiate w.r.t. a:

$$\frac{\partial}{\partial a} L(a,b) = \mathbb{E}\big[2(Y - aX - b)(-X)\big] = -2\mathbb{E}\big[X(Y - aX - b)\big].$$

Set to zero and substitute $b = \mu_Y - a\mu_X$:

$$\mathbb{E}\big[X(Y - aX - (\mu_Y - a\mu_X))\big] = 0.$$

Rearrange:

$$\mathbb{E}[XY] - a\mathbb{E}[X^2] - \mu_Y\mathbb{E}[X] + a\mu_X\mathbb{E}[X] = 0.$$

Using $\mathbb{E}[X] = \mu_X$ this simplifies to

$$\mathbb{E}[XY] - \mu_X\mu_Y - a(\mathbb{E}[X^2] - \mu_X^2) = 0.$$

Recognize covariance and variance:

$$\mathrm{Cov}(X,Y) - a\ \mathrm{Var}(X) = 0.$$

Thus the optimal slope is

$$\boxed{a^* \ = \ \frac{\mathrm{Cov}(X,Y)}{\mathrm{Var}(X)}}$$

and (recalling the formula for b) the intercept is

$$\boxed{b^* \ = \ \mu_Y - a^*\mu_X}.$$

Interpretation.

- The slope a^* is proportional to $\mathrm{Cov}(X,Y)$. If the covariance is zero, the best linear predictor (in mean-square) is a horizontal line $Y \approx \mu_Y$ — i.e. knowing X does not reduce mean-square error for a linear predictor.
- The variance $\mathrm{Var}(X)$ in the denominator rescales the slope: for a given covariance, a more variable X yields a smaller slope (because each unit change in X is "bigger").
- Equivalently, writing $a^* = \rho_{X,Y}\,(\sigma_Y/\sigma_X)$ (where ρ is the correlation and σ's are standard deviations) highlights that correlation measures the *direction and strength* of the linear association, while the factor σ_Y/σ_X converts it to slope units.

Why correlation measures "linearity" (brief quantitative argument). From the Cauchy–Schwarz inequality,

$$\big|\operatorname{Cov}(X,Y)\big| = \big|\mathbb{E}[(X-\mu_X)(Y-\mu_Y)]\big| \le \sqrt{\mathbb{E}[(X-\mu_X)^2]}\,\sqrt{\mathbb{E}[(Y-\mu_Y)^2]} = \sigma_X\sigma_Y.$$

Thus $|\rho_{X,Y}| \le 1$. Equality $|\rho| = 1$ holds iff $(X-\mu_X)$ and $(Y-\mu_Y)$ are linearly dependent almost surely; i.e. there exists constants α, β (not both zero) such that $\alpha(X-\mu_X)+\beta(Y-\mu_Y)=0$ with probability one — equivalently Y is an exact linear function of X (with probability one). This shows correlation equals ± 1 exactly in the case of perfect linear relationship, so $|\rho|$ quantifies how closely the pair follows a line.

Takeaway for students. Covariance is the raw quantity that determines the slope of the optimal linear predictor; correlation is its standardized form that lies in $[-1,1]$ and tells us how nearly the variables follow an exact linear relation. Thus covariance/-correlation are not arbitrary algebraic constructs — they are the quantities that arise when you ask the precise question: "Which linear predictor minimizes mean squared error?"

Quick exercise. Using the formula above, verify that when X and Y are independent (so $\operatorname{Cov}(X,Y)=0$), the best linear predictor of Y from X is the constant μ_Y.

20 Lecture XX: Conditional Distributions

We have seen how independence gives us a very clean factorization structure:

- For discrete random variables, $p_{X,Y}(x,y) = p_X(x)p_Y(y)$.
- For continuous random variables, $f_{X,Y}(x,y) = f_X(x)f_Y(y)$.

But in practice, variables are rarely independent. Still, it is useful to impose structure: *conditional distributions* provide exactly this. They describe how one variable is distributed given the value of another.

20.1 Definitions

Definition 20.1 (Conditional distributions)**.**

- *If (X,Y) is discrete with joint pmf $p_{X,Y}(x,y)$ and marginal pmf $p_X(x)$, then*

$$p_{Y|X}(y \mid x) = \frac{p_{X,Y}(x,y)}{p_X(x)}, \qquad p_X(x) > 0.$$

- *If (X,Y) is continuous with joint pdf $f_{X,Y}(x,y)$ and marginal pdf $f_X(x)$, then*

$$f_{Y|X}(y \mid x) = \frac{f_{X,Y}(x,y)}{f_X(x)}, \qquad f_X(x) > 0.$$

Example 20.1 (Discrete case with Joint PMF)**.** Let X take values in $\{1,2,3\}$ and Y take values in $\{A,B,C,D\}$ with the following joint pmf:

$p(x,y)$	$Y=A$	$Y=B$	$Y=C$	$Y=D$	$p_X(x)$
$X=1$	0.05	0.05	0.10	0.10	0.30
$X=2$	0.05	0.10	0.05	0.10	0.30
$X=3$	0.05	0.05	0.10	0.20	0.40
$p_Y(y)$	0.15	0.20	0.25	0.40	1

Solution. Here are some example calculations based on the table.

- $p_{Y|X}(B \mid 2) = \dfrac{p(2,B)}{p_X(2)} = \dfrac{0.10}{0.30} = \frac{1}{3}$.
- $p_{X|Y}(3 \mid D) = \dfrac{p(3,D)}{p_Y(D)} = \dfrac{0.20}{0.40} = 0.5$.

Note how the *law of total probability* is used to find marginals:

$$p_Y(D) = \sum_{x=1}^{3} p_{X,Y}(x, D) = 0.10 + 0.10 + 0.20 = 0.40.$$

Example 20.2 (Continuous case with Joint PDF)**.** Let (X, Y) have joint density

$$f_{X,Y}(x, y) = \begin{cases} 2, & 0 < x < 1,\ 0 < y < x, \\ 0, & \text{otherwise.} \end{cases}$$

Step 1: Find the marginal of X.

$$f_X(x) = \int_0^x 2\, dy = 2x, \qquad 0 < x < 1.$$

Step 2: Form the conditional density.

$$f_{Y|X}(y \mid x) = \frac{f_{X,Y}(x, y)}{f_X(x)} = \frac{2}{2x} = \frac{1}{x}, \qquad 0 < y < x.$$

So conditional on $X = x$, the random variable Y is uniform on $(0, x)$.

Important Remark 20.1. *Note that in the last result, it is important to remember that the conditional density depends on the condition, i.e., the value $X = x$. You should think of the conditional density as a function of the conditional value x (as opposed the first variable y).*

20.2 Independence revisited

Proposition 20.1. *X and Y are independent if and only if the conditional distribution of Y given X equals the marginal distribution of Y (and symmetrically for X). That is,*

$$f_{Y|X}(y \mid x) = f_Y(y) \quad \textit{for all } x \textit{ with } f_X(x) > 0$$

and analogously in the discrete case.

21 Lecture XXI: Conditional Expectation

Last class, we have defined conditional distribution via the fundamental definition of conditional probability, and have explored the ways to compute them.

Before we move further, notice that a *conditional distribution* is still a distribution, so we can ask: what is the random variable associated with it?

The answer is the random variable Y under the condition $X = x$. written as

$$Y \mid X = x.$$

This is to say, for each fixed x, $Y \mid X = x$ is a random variable (because Y can still take many values). This distinction is crucial between conditional random variable $Y \mid X = x$ and the unconditional one Y: the conditioning parameter x turns the distribution into a deterministic recipe; but the randomness of Y still remains within that recipe – it is just the universe in which you are exploring this randomness has changed (has at least become smaller since you provided the information $X = x$).

21.1 The Expected Value $E[Y \mid X = x]$

Since a conditional distribution is a proper probability distribution (recall from a while ago that you have shown in HW that the conditional probability is a probability that satisfies all three Axioms), it makes sense to proceed to define statistical quantities such as conditional expectation and variance.

The importance here lies in the word "value". The quantity we are about to compute is a real number (so it is not random).

Definition 21.1 (Conditional Expectation)**.** *If (X, Y) is a pair of random variables with conditional distribution of Y given $X = x$, then*

$$E[Y \mid X = x] = \begin{cases} \displaystyle\sum_y y\, p_{Y|X}(y \mid x), & \text{if discrete,} \\ \displaystyle\int_{-\infty}^{\infty} y\, f_{Y|X}(y \mid x)\, dy, & \text{if continuous.} \end{cases}$$

Definition 21.2 (Conditional Variance)**.** *Given $X = x$, the conditional variance of Y is*

$$\mathrm{Var}(Y \mid X = x) = E\Big[(Y - E[Y \mid X = x])^2 \,\Big|\, X = x\Big].$$

Example 21.1 (Discrete example cont.)**.** Recall the joint pmf table:

$p(x,y)$	$Y=1$	$Y=2$	$Y=3$	$Y=4$	$p_X(x)$
$X=1$	0.05	0.05	0.10	0.10	0.30
$X=2$	0.05	0.10	0.05	0.10	0.30
$X=3$	0.05	0.05	0.10	0.20	0.40
$p_Y(y)$	0.15	0.20	0.25	0.40	1

Then we can compute:

$$E[Y \mid X = 2] = \sum_{y=1}^{4} y\, p_{Y|X}(y \mid 2).$$

From the table:

$$p_{Y|X}(1 \mid 2) = \tfrac{0.05}{0.30}, \quad p_{Y|X}(2 \mid 2) = \tfrac{0.10}{0.30}, \quad p_{Y|X}(3 \mid 2) = \tfrac{0.05}{0.30}, \quad p_{Y|X}(4 \mid 2) = \tfrac{0.10}{0.30}.$$

So,

$$E[Y \mid X = 2] = 1 \cdot \tfrac{0.05}{0.30} + 2 \cdot \tfrac{0.10}{0.30} + 3 \cdot \tfrac{0.05}{0.30} + 4 \cdot \tfrac{0.10}{0.30} = \frac{1.25}{0.30} \approx 4.17.$$

Next, the conditional variance:

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

$$E[Y^2 \mid X = 2] = 1^2 \cdot \tfrac{0.05}{0.30} + 2^2 \cdot \tfrac{0.10}{0.30} + 3^2 \cdot \tfrac{0.05}{0.30} + 4^2 \cdot \tfrac{0.10}{0.30}.$$

This can be computed directly.

Mark down these conditional expectation and conditional probabilities. We will use them a little later.

21.2 The Random Variable $E[Y \mid X]$

Now, what if we also don't specify the conditional value, so X also gets its randomness back? What monstrosity is this?

At first glance, the notation $E[Y \mid X]$ seems bizarre. We are no longer conditioning on a specific value $X = x$, but instead on the entire random variable X itself.

- For each fixed x, $E[Y \mid X = x]$ is just a number.
- If we now let X be random again, then

  $$E[Y \mid X] := \big(E[Y \mid X = x]\big)\big|_{x=X}.$$

  That is, $E[Y \mid X]$ is the deterministic function of x evaluated at the random input X.
- Hence $E[Y \mid X]$ is itself a *random variable*, whose randomness comes entirely from X.

21.2.1 Distribution, Mean, and Variance of The Random Variable $E[Y \mid X]$.

Since $E[Y \mid X]$ is a random variable, it has its own distribution, and we can meaningfully ask for its mean and variance. In fact, these are governed by the following powerful laws:

Theorem 21.1 (Law of Total Expectation)**.** *For any pair of random variables* (X, Y)*,*

$$E[Y] = E\big[\, E[Y \mid X] \,\big].$$

Corollary 21.1 (Law of Total Expectation: **discrete** case)**.** *If* X *and* Y *are* ***discrete****, then*

$$E[Y] = E\big[\, E[Y \mid X] \,\big] = \sum_x E[Y \mid X = x]\, p_X(x).$$

Note that the summation is to take care of the outer expectation, while the inner expected value is retained inside as the summand. Note further that the pmf we use here is for X because that's precisely where the randomness is.

Corollary 21.2 (Law of Total Expectation: **continuous** case)**.** *Let* (X, Y) *be* ***continuous*** *random variables with joint pdf* $f_{X,Y}(x, y)$*. Then*

$$E[Y] \;=\; \int_{-\infty}^{\infty} E[Y \mid X = x]\, f_X(x)\, dx,$$

where

$$E[Y \mid X = x] = \int_{-\infty}^{\infty} y\, f_{Y|X}(y \mid x)\, dy.$$

In calculations (shown later), one may sometimes find immediately nice formulas to relate the expected value of Y and that of X by observing the data $E[Y \mid X = x]$; the integrals are often avoided.

Theorem 21.2 (Law of Total Variance)**.** *For any pair of random variables* (X, Y)*,*

$$\mathrm{Var}(Y) = E\big[\mathrm{Var}(Y \mid X)\big] + \mathrm{Var}\big(E[Y \mid X]\big).$$

Interpretation

- The first law says: if you first take expectations conditionally (given X) and then average over X, you recover the unconditional expectation of Y.
- The second law splits variability of Y into two parts: the *average conditional variability* and the *variability of the conditional means.*

So the apparent monstrosity $E[Y \mid X]$ is not so monstrous after all — it is simply a random variable derived from X, and its role is to link the conditional world back to the unconditional one.

21.3 Examples of Laws of Total Expectation and Total Variance

We demonstrate that the law of total expectation indeed provides an alternative to calculate the expected value of Y (the variable of interest).

Example 21.2 (Discrete)**.** Recall the joint pmf table (with numeric coding $Y \in \{1, 2, 3, 4\}$):

	$Y=1$	$Y=2$	$Y=3$	$Y=4$	$p_X(x)$
$X=1$	0.05	0.05	0.10	0.10	0.30
$X=2$	0.05	0.10	0.05	0.10	0.30
$X=3$	0.05	0.05	0.10	0.20	0.40
$p_Y(y)$	0.15	0.20	0.25	0.40	1

1. Conditional expectations.
   We first compute $E[Y \mid X = x]$ for each x using the conditional probabilities found in Example 21.1.

   For $X = 1$: conditional probs are

   $$p_{Y|X}(\cdot \mid 1) = (1/6,\ 1/6,\ 1/3,\ 1/3),$$

   so

   $$E[Y \mid X = 1] = 1 \cdot \tfrac{1}{6} + 2 \cdot \tfrac{1}{6} + 3 \cdot \tfrac{1}{3} + 4 \cdot \tfrac{1}{3} = \frac{17}{6} \approx 2.8333.$$

   For $X = 2$: conditional probs are $(1/6,\ 1/3,\ 1/6,\ 1/3)$, so

   $$E[Y \mid X = 2] = 1 \cdot \tfrac{1}{6} + 2 \cdot \tfrac{1}{3} + 3 \cdot \tfrac{1}{6} + 4 \cdot \tfrac{1}{3} = \frac{8}{3} \approx 2.6667.$$

   For $X = 3$: conditional probs are $(1/8,\ 1/8,\ 1/4,\ 1/2)$, so

   $$E[Y \mid X = 3] = 1 \cdot \tfrac{1}{8} + 2 \cdot \tfrac{1}{8} + 3 \cdot \tfrac{1}{4} + 4 \cdot \tfrac{1}{2} = \frac{25}{8} = 3.125.$$

2. Law of Total Expectation

   $$E[Y] = E\big[\,E[Y \mid X]\,\big] = \sum_{x=1}^{3} E[Y \mid X = x]\, p_X(x).$$

   Using $p_X = (0.3, 0.3, 0.4)$,

   $$E[Y] = \tfrac{17}{6} \cdot 0.3 + \tfrac{8}{3} \cdot 0.3 + \tfrac{25}{8} \cdot 0.4 = 2.9.$$

   (Checks with the direct marginal calculation $E[Y] = \sum_y y\, p_Y(y) = 1(0.15) + 2(0.2) + 3(0.25) + 4(0.4) = 2.9$.)

3. Law of Total Variance

$$\mathrm{Var}(Y) = E\big[\mathrm{Var}(Y \mid X)\big] + \mathrm{Var}\big(E[Y \mid X]\big).$$

Compute the two pieces:

$$E[\mathrm{Var}(Y \mid X)] = 0.3 \cdot \frac{41}{36} + 0.3 \cdot \frac{11}{9} + 0.4 \cdot \frac{71}{64} = \frac{553}{480} \approx 1.1520833,$$

$$\mathrm{Var}(E[Y \mid X]) = \Big(0.3 \cdot \frac{289}{36} + 0.3 \cdot \frac{64}{9} + 0.4 \cdot \frac{625}{64}\Big) - (2.9)^2 = \frac{91}{2400} \approx 0.0379167.$$

Hence

$$\mathrm{Var}(Y) = \frac{553}{480} + \frac{91}{2400} = \frac{2856}{2400} = 1.19.$$

(Direct check: $\mathrm{Var}(Y) = E[Y^2] - (E[Y])^2$ with $E[Y^2] = 9.6$ and $E[Y] = 2.9$ gives $9.6 - 8.41 = 1.19$, matching the total-variance result.)

Example 21.3 (Continuous)**.** Let (X, Y) have joint pdf

$$f_{X,Y}(x, y) = \begin{cases} 2, & 0 < x < 1,\ 0 < y < x, \\ 0, & \text{otherwise.} \end{cases}$$

(So (X, Y) is uniform on the triangle $0 < y < x < 1$.)

1. Marginal of X.

$$f_X(x) = \int_0^x 2\, dy = 2x, \qquad 0 < x < 1.$$

2. Conditional law of Y given $X = x$.

$$f_{Y|X}(y \mid x) = \frac{f_{X,Y}(x, y)}{f_X(x)} = \frac{2}{2x} = \frac{1}{x}, \qquad 0 < y < x,$$

so $Y \mid X = x \sim \mathrm{Unif}(0, x)$. Hence

$$E[Y \mid X = x] = \frac{x}{2}, \qquad \mathrm{Var}(Y \mid X = x) = \frac{x^2}{12}.$$

3. Law of total expectation.

$$E[Y] = E\big[\,E[Y \mid X]\,\big] = E\Big[\frac{X}{2}\Big] = \frac{1}{2}E[X].$$

Compute $E[X]$ using $f_X(x) = 2x$:

$$E[X] = \int_0^1 x \cdot 2x\,dx = 2\int_0^1 x^2\,dx = 2 \cdot \frac{1}{3} = \frac{2}{3}.$$

Thus

$$E[Y] = \frac{1}{2} \cdot \frac{2}{3} = \frac{1}{3}.$$

4. Law of total variance.

$$\mathrm{Var}(Y) = E\big[\mathrm{Var}(Y \mid X)\big] + \mathrm{Var}\big(E[Y \mid X]\big) = E\left[\frac{X^2}{12}\right] + \mathrm{Var}\left(\frac{X}{2}\right).$$

Compute the pieces:

$$E[X^2] = \int_0^1 x^2 \cdot 2x\,dx = 2\int_0^1 x^3\,dx = 2 \cdot \frac{1}{4} = \frac{1}{2},$$

so $E[X^2/12] = \frac{1}{12} \cdot \frac{1}{2} = \frac{1}{24}$.

Also

$$\mathrm{Var}\left(\frac{X}{2}\right) = \frac{1}{4}\mathrm{Var}(X), \qquad \mathrm{Var}(X) = E[X^2] - \big(E[X]\big)^2 = \frac{1}{2} - \left(\frac{2}{3}\right)^2 = \frac{1}{18}.$$

Therefore $\mathrm{Var}(X/2) = \frac{1}{4} \cdot \frac{1}{18} = \frac{1}{72}$.

So

$$\mathrm{Var}(Y) = \frac{1}{24} + \frac{1}{72} = \frac{1}{18}.$$

(As a sanity check one could compute $\mathrm{Var}(Y) = E[Y^2] - E[Y]^2$ directly; both routes give $\mathrm{Var}(Y) = 1/18$.)

Important Remark 21.1. *Note that in both examples, we did NOT need the marginal distribution of Y. This is a big gain in practice. See further motivation below.*

21.4 Motivation: Why Conditional Expectation?

At first sight, conditional expectation seems like needless complication: after all, we already knew how to compute $E[Y]$ or $\mathrm{Var}(Y)$ directly by summing or integrating with respect to Y. So why bring in conditional distributions at all?

1. Simplification through decomposition (mathematical motivation). The Laws of Total Expectation and Variance let us break a complicated problem into smaller, easier parts:

$$E[Y] = E\big[E[Y \mid X]\big], \qquad \mathrm{Var}(Y) = E[\mathrm{Var}(Y \mid X)] + \mathrm{Var}(E[Y \mid X]).$$

The strategy is always:

1. Analyze Y locally given $X = x$ (often much simpler).
2. Average over X to recover the global picture.

This is especially useful when the direct marginal of Y is complicated or unavailable.

2. Prediction and information (statistical motivation). Conditional expectation $E[Y \mid X]$ is the precise mathematical form of "best prediction of Y given knowledge of X." It is random because as X changes, our prediction for Y changes. This makes conditional expectation the language of learning from information — at the heart of statistics, forecasting, and machine learning.

3. Practical availability of conditional densities (practical motivation). Notice that to compute $E[Y \mid X]$ we do *not* need the marginal density of Y at all. We only require:

- the conditional density $f_{Y|X}(y \mid x)$,
- and the marginal density $f_X(x)$.

In practice, it may be much easier to observe or model the conditional distribution directly:

- In reliability studies: it is natural to record the lifetime of a component *given* environmental stress level X.
- In medical trials: it is more practical to model patient response Y *given* treatment type X, rather than try to obtain the overall marginal law of Y.
- In queueing systems: service time distribution Y is often studied *given* the number of servers X.

In all such cases, marginalizing Y directly would be very difficult, but conditioning on X produces a tractable path forward.

You may consider the following realistic examples where obtaining the marginal of Y (our variable of interest) is nearly impossible.

Practical scenarios where conditional densities are more accessible than marginals.

1. **Weather and energy consumption.**
   - Y = household electricity usage (kWh per day).
   - The marginal distribution of Y is extremely complicated: it depends on season, temperature, day of week, household size, income, appliance ownership, etc.
   - If we instead condition on daily average temperature X, then $f_{Y|X}(y \mid x)$ becomes much simpler: on hot days, usage shifts higher due to air conditioning; on mild days, usage is lower.
   - Measuring X (temperature) is trivial, and it "organizes" the data into conditional slices where modeling Y is manageable.

2. **Response time in computer networks.**
   - Y = response time of a packet traveling through a large network.
   - The unconditional distribution of Y is wildly complicated because it mixes over all possible traffic patterns, server loads, and routing paths.
   - If we condition on the current network load X (e.g., the number of concurrent users), then $Y \mid X = x$ often has a much simpler form, sometimes close to exponential.
   - Measuring X is simple (system logs provide it), while capturing the full marginal of Y would require tracking an astronomical number of network states.

3. **Medical trials.**
   - Y = recovery time (days until remission).
   - The overall marginal of Y across all patients is complicated because it averages over treatment groups, ages, and disease severity.
   - If we condition on treatment type X, then $Y \mid X = x$ is much easier to model and interpret.
   - Recording X is trivial (trial design provides it), while modeling the overall marginal of Y would obscure the differences between treatments.

22 Lecture XXI: Sequence of Random Variables

Today's main takeaway is

The sum of independent normal random variables is itself normal, with mean equal to the sum of the means and variance equal to the sum of the variances.

This statement may look almost obvious — but in fact, it reflects a very special and nontrivial property of the normal distribution. The following sections are to develop the technique to demonstrate the main takeaway.

22.1 Motivation of Technique: A Conditional Probability Exercise

We often want to understand sums of random variables:

$$S_n = X_1 + X_2 + \cdots + X_n.$$

For example:

- Total waiting time for n bus arrivals,
- Total demand from n customers,
- Total error from n repeated measurements.

Already for two variables, we face the problem of describing the distribution of $X + Y$. Let us see how to compute probabilities of events like $P(X + Y < 3)$ in two different ways.

Example 22.1 (Two Exponential Random Variables: Waiting for the 2nd bus (revisited))**.** Suppose X and Y are independent exponential random variables with rate 1, with joint pdf

$$f_{X,Y}(x, y) = e^{-x}e^{-y}, \qquad x, y > 0.$$

Method 1: Region of integration (already learned)

$$P(X + Y < 3) = \int_0^3 \int_0^{3-x} e^{-x}e^{-y}\, dy\, dx.$$

Method 2: Conditional probability

$$P(X + Y < 3) = \int_0^\infty P(Y < 3 - x \mid X = x)\, f_X(x)\, dx.$$

Since $Y \sim \text{Exp}(1)$, for $x < 3$,

$$P(Y < 3 - x \mid X = x) = 1 - e^{-(3-x)}.$$

Thus

$$P(X + Y < 3) = \int_0^3 \left(1 - e^{-(3-x)}\right) e^{-x}\, dx.$$

This produces the same result as Method 1, but highlights conditional probability as a tool.

The highlight here for Method 2 is in fact a restatement of **the Law of Total Probability** for continuous random variables. We state it here again in detail so you can say a nice "welcome back"!

Theorem 22.1 (Law of Total Probability: Discrete case)**.** *If $\{B_i\}$ is a partition of the sample space with $P(B_i) > 0$, then for any event A,*

$$P(A) = \sum_i P(A \mid B_i) P(B_i).$$

In the setting of random variables, we may take $B_i = \{X = x_i\}$ when X is discrete. Then

$$P(Y \in A) = \sum_x P(Y \in A \mid X = x)\, P(X = x).$$

Theorem 22.2 (Law of Total Probability: Continuous case)**.** *Let A be any event, and X a continuous random variable with density $f_X(x)$. Then*

$$P(A) = \int_{-\infty}^{\infty} P(A \mid X = x)\, f_X(x)\, dx.$$

Proof. First, convince yourself that $P(A) = \mathbb{E}\left[\mathbf{1}_A\right]$. Then, it follows by law of total expectation that

$$\begin{aligned} P(A) &= \mathbb{E}\left[\mathbf{1}_A\right] \\ &= \mathbb{E}_x\left[\mathbb{E}\left[\mathbf{1}_A \mid X\right]\right] \\ &= \int_{\mathbb{R}} \mathbb{E}\left[\mathbf{1}_A \mid X = x\right] f_X(x)\, dx \\ &= \int_{\mathbb{R}} P(A \mid X = x)\, f_X(x)\, dx \end{aligned}$$

□

22.2 General Structure: Convolution Formula

The example suggests a general rule.

Theorem 22.3 (Convolution Formula)**.** *If X and Y are independent continuous random variables with densities f_X and f_Y, then the sum $Z = X + Y$ has density*

$$f_Z(z) = \int_{-\infty}^{\infty} f_X(x) f_Y(z - x)\, dx.$$

Proof. Consider the CDF of Z. We apply the law of total probability right away and then use the fact that pdf is the derivative of cdf (Fundamental Theorem of Calculus).

$$\begin{aligned} F_Z(z) &= P(Z \leq z) \\ &= P(X + Y \leq z) \\ &= \int P(Y \leq z - X \mid X = x)\, f_X(x)\, dx \\ &= \int P(Y \leq z - x)\, f_X(x)\, dx \\ &= \int F_Y(z - x)\, f_X(x)\, dx \end{aligned}$$

Then, the density of Z follows by differentiating F_Z against z yields

$$\begin{aligned} f_Z(z) = \frac{dF_Z}{dz} &= \frac{d}{dz} \int_{\mathbb{R}} f_X(x)\, F_Y(z - x)\, dx \\ &= \int_{\mathbb{R}} f_X(x)\, \frac{dF_Y}{dz}(z - x)\, dx \\ &= \int_{\mathbb{R}} f_X(x)\, f_Y(z - x)\, dx \end{aligned}$$

□

Takeaway

- To compute probabilities involving $X + Y$, we can integrate directly over the region $\{x + y < z\}$ or use conditional probability.
- Both methods lead naturally to the convolution formula, which is the fundamental tool for analyzing sums of random variables.

Question 22.1. What if now we have a sum of three or more random variables? Will this convolution theorem hold? It looks like it is going to be messy!

If you are here for HW6 Problem 4, take a step back, and try to do the problem unassisted. The problem is a special case of the following, but the technique is practically the same: completing the square.

Example 22.2 (Sum of Independent Normal Variables)**.** Let $X \sim N(\mu_1, \sigma_1^2)$ and $Y \sim N(\mu_2, \sigma_2^2)$ be independent. Then

$$Z := X + Y \sim N(\mu_1 + \mu_2,\ \sigma_1^2 + \sigma_2^2).$$

But how do we verify this? Can we use the convolution theorem, and is the process pretty? Doesn't hurt to try it.

Proof via convolution (completing the square). Write the densities

$$f_X(x) = \frac{1}{\sqrt{2\pi}\sigma_1} \exp\Big(- \frac{(x-\mu_1)^2}{2\sigma_1^2}\Big), \quad f_Y(y) = \frac{1}{\sqrt{2\pi}\sigma_2} \exp\Big(- \frac{(y-\mu_2)^2}{2\sigma_2^2}\Big).$$

The density of $Z = X + Y$ is the convolution

$$f_Z(z) = \int_{-\infty}^{\infty} f_X(x)\, f_Y(z-x)\, dx.$$

Substitute and combine exponents:

$$f_Z(z) = \frac{1}{2\pi\sigma_1\sigma_2} \int_{-\infty}^{\infty} \exp\Big(- \frac{(x-\mu_1)^2}{2\sigma_1^2} - \frac{(z-x-\mu_2)^2}{2\sigma_2^2}\Big)\, dx.$$

The exponent is a quadratic in x. Collect terms and complete the square in x; after algebra one obtains the identity

$$-\frac{(x-\mu_1)^2}{2\sigma_1^2} - \frac{(z-x-\mu_2)^2}{2\sigma_2^2} = -\frac{(z-(\mu_1+\mu_2))^2}{2(\sigma_1^2+\sigma_2^2)} - \frac{\big(x - m(z)\big)^2}{2\,s^2},$$

for suitable

$$s^2 = \frac{\sigma_1^2\sigma_2^2}{\sigma_1^2+\sigma_2^2}, \qquad m(z) = \frac{\sigma_2^2\mu_1 + \sigma_1^2(z-\mu_2)}{\sigma_1^2+\sigma_2^2}.$$

(The precise form of $m(z)$ is not important for the conclusion; it is the center of the Gaussian in x.) Therefore

$$f_Z(z) = \frac{1}{2\pi\sigma_1\sigma_2} \exp\Big(- \frac{(z-(\mu_1+\mu_2))^2}{2(\sigma_1^2+\sigma_2^2)}\Big) \int_{-\infty}^{\infty} \exp\Big(- \frac{(x-m(z))^2}{2s^2}\Big)\, dx.$$

The integral equals $\sqrt{2\pi}\, s$. Simplifying the prefactor yields

$$f_Z(z) = \frac{1}{\sqrt{2\pi(\sigma_1^2+\sigma_2^2)}} \exp\Big(- \frac{(z-(\mu_1+\mu_2))^2}{2(\sigma_1^2+\sigma_2^2)}\Big),$$

i.e. $Z \sim N(\mu_1+\mu_2, \sigma_1^2+\sigma_2^2)$. □

Nooooooooo, this is so messy!

22.3 Moment Generating Function of A Sum of Random Variables

We saw just now that while the convolution theorem gets the job done, it ain't pretty. We must do something smart and efficient, now that we have endured the pain.

Theorem 22.4 (MGF of a sum of independent random variables)**.** *Let $X_1, \ldots, X_n$ be independent random variables and suppose each X_i has an mgf $M_{X_i}(t) = \mathbb{E}[e^{tX_i}]$ that exists in a neighborhood of $t = 0$. Define the sum*

$$S_n := \sum_{i=1}^{n} X_i.$$

Then the mgf of S_n exists in a neighborhood of 0 and satisfies

$$M_{S_n}(t) \;=\; \prod_{i=1}^{n} M_{X_i}(t).$$

Proof. By definition,

$$M_{S_n}(t) = \mathbb{E}\big[e^{tS_n}\big] = \mathbb{E}\Big[e^{t\sum_{i=1}^{n} X_i}\Big] = \mathbb{E}\Big[\prod_{i=1}^{n} e^{tX_i}\Big].$$

Independence allows factorization of expectations:

$$\mathbb{E}\Big[\prod_{i=1}^{n} e^{tX_i}\Big] = \prod_{i=1}^{n} \mathbb{E}\big[e^{tX_i}\big] = \prod_{i=1}^{n} M_{X_i}(t).$$

The product is finite and well-defined in any neighborhood of $t = 0$ where each M_{X_i} exists. □

Recall that the MGF of a normal random variable $X \sim N(\mu, \sigma)$ is $M_X(t) = e^{t\mu+\sigma^2 t^2/2}$. We make use of this MGF as an observation for the following corollary of the MGF theorem, revisiting a result from earlier where the proof is much more heinous.

Corollary 22.1 (Sum of independent normals)**.** *If $X_i \sim N(\mu_i, \sigma_i^2)$ are independent for $i = 1, \ldots, n$, then*

$$S_n = \sum_{i=1}^{n} X_i \sim N\Big(\sum_{i=1}^{n} \mu_i,\; \sum_{i=1}^{n} \sigma_i^2\Big).$$

In words, the sum of independent normal random variables is still normal, with mean and variance as the sum of means and variances of individual normal variables, respectively.

Proof. Each normal has mgf

$$M_{X_i}(t) = \exp\Big(\mu_i t + \tfrac{1}{2}\sigma_i^2 t^2\Big).$$

By the theorem,

$$\begin{aligned}
M_{S_n}(t) &= \mathbb{E}\left[e^{tS_n}\right] \\
&= \mathbb{E}\left[e^{t\sum_{i=1}^n X_i}\right] \\
&= \mathbb{E}\left[e^{tX_1}e^{tX_2}\cdots e^{tX_n}\right] \\
&\overset{\text{Independence}}{=} \mathbb{E}\left[e^{tX_1}\right]\mathbb{E}\left[e^{tX_2}\right]\cdots\mathbb{E}\left[e^{tX_n}\right] \\
&= \prod_{i=1}^n M_{X_i}(t) \\
&= \prod_{i=1}^n e^{t\mu_i + \frac{\sigma_i t^2}{2}} \\
&= \exp\Big(\Big(\sum_{i=1}^n \mu_i\Big)t + \tfrac{1}{2}\Big(\sum_{i=1}^n \sigma_i^2\Big)t^2\Big)
\end{aligned}$$

The right-hand side is the mgf of a normal with mean $\sum_i \mu_i$ and variance $\sum_i \sigma_i^2$. Hence S_n is normal with the stated parameters. □

Here is a fact that we don't aim to prove in this class, but one of the most astounding results in probability theory:

The normal distribution is the *only* distribution family that is closed under addition of independent random variables.

Part V
Limit Theorems

23 Lecture XXII: The Central Limit Theorem

In this class, we study what happens to **distribution** of a sequence of random variables $X_1, X_2, \ldots, X_n$ as n goes to infinity.

Suppose we collect repeated measurements of the same physical quantity — say, the height of a plant, or the waiting time between bus arrivals. Each measurement X_i is random because of noise and other sources of variability. We are interested in *aggregates* of these measurements, such as:

$$S_n = X_1 + X_2 + \cdots + X_n, \qquad \bar{X}_n = \frac{S_n}{n}.$$

Questions naturally arise:

- What happens to $\bar{X}_n$ as the number of samples increases?
- Does it approach a constant value (Law of Large Numbers, next class)?
- And what does the distribution of properly scaled sums look like? (Central Limit Theorem, this class)

This lecture ties together nearly everything we have seen so far: expectation, variance, independence, and moment generating functions.

23.1 Example: Why Study Sums of Random Variables?

Suppose we want to determine whether a coin is fair. We toss it once and record $X_1 = 1$ for heads and $X_1 = 0$ for tails. But a single trial tells us nothing — the outcome could easily have gone either way.

We toss it again, and again, say n times, recording independent Bernoulli random variables

$$X_1, X_2, \ldots, X_n \sim \text{Bernoulli}(p),$$

where p is the (unknown) probability of heads.

The total number of heads

$$S_n = X_1 + X_2 + \cdots + X_n$$

is a *sum of random variables*. From S_n, we form the sample mean

$$\bar{X}_n = \frac{S_n}{n},$$

which **estimates** the true probability p.

This setup is far more general than coins:

- In biology, we average repeated measurements of a cell's response.
- In engineering, we aggregate sensor readings to estimate a signal.
- In finance, we compute the average daily return of an asset.

In every case, a key question emerges:

What happens to the distribution of the sum (or average) as the number of samples increases?

We might expect the average to stabilize near the true mean — that's the idea behind the **Law of Large Numbers**. But there's a deeper question about the *shape* of the distribution: as we keep aggregating independent random effects, what form does the overall distribution take? Will the convolution theorem (Theorem 22.3) or the mgf theorem (Theorem 22.4) from last class help us?

That is the mystery addressed by the **Central Limit Theorem**. Before going into details, we first make a few definitions that assist us through the way to the big results.

23.2 Independent and Identically Distributed (IID) Random Variables

First, we make an official clarification of what we often said before involving the acronym "i.i.d".

Definition 23.1. *The r.v.s $X_1, X_2, \ldots, X_n$ are said to be **independent and identically distributed (i.i.d)** if*

1. *The X_i's are independent r.v.s.*
2. *Every X_i has the **same** probability distribution.*

*Such a colelction of r.v.s is also called a (simple) **random sample** of size n.*

23.3 Sample Mean and Variance

Definition 23.2 (Sample Mean and Sample Variance)**.** *For IID random variables $X_1, X_2, \ldots, X_n$ with mean $\mu = E[X_i]$ and variance $\sigma^2 = Var[X_i]$,*

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i, \qquad S_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2.$$

Proposition 23.1. *For the sample mean,*

$$E[\bar{X}_n] = \mu, \qquad Var(\bar{X}_n) = \frac{\sigma^2}{n}.$$

Proof. Standard calculations. You do it via linearity of expected value. □

Remark 23.1. *If the X_i are normal, say $X_i \sim N(\mu, \sigma^2)$, then $\bar{X}_n$ is also normal:*

$$\bar{X}_n \sim N\left(\mu, \frac{\sigma^2}{n}\right).$$

The normal distribution is the only family closed under addition and averaging.

23.4 The Central Limit Theorem

These visualizations reveal a universal pattern:

- As n increases, the shape of the standardized sum Z_n converges toward the standard normal curve.
- The underlying distribution (uniform, exponential, etc.) matters less — the bell curve emerges naturally.

Theorem 23.1 (Central Limit Theorem)**.** *Let $X_1, X_2, \ldots, X_n$ be IID random variables with mean μ and finite variance σ^2. Then the standardized sum*

$$Z_n = \frac{\sum_{i=1}^{n} X_i - n\mu}{\sqrt{n}\sigma}$$

converges in distribution to a standard normal random variable $Z \sim N(0,1)$ as $n \to \infty$. We write

$$Z_n \xrightarrow{d} Z.$$

Proof. Come take a second, more serious course, in probability theory, when you gain enough knowledge about calculus, and surprisingly, linear algebra, and complex variable. These things are simplifying an otherwise difficult result to show. Therefore, they are NOT extra work, but the essentials for survival in this industry. □

Example 23.1. A factory produces screws whose lengths (in mm) are independent random variables with mean $\mu = 50$ and standard deviation $\sigma = 2$. We draw a random sample of $n = 100$ screws and record their sample mean $\bar{X}_n$.

Question. What is the approximate probability that the sample mean lies between 49.5 and 50.5?

Step 1. Identify the ingredients.

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i, \qquad \mathbb{E}[\bar{X}_n] = \mu = 50, \qquad \mathrm{Var}(\bar{X}_n) = \frac{\sigma^2}{n} = \frac{4}{100} = 0.04.$$

Step 2. Apply the CLT. For large n, the sampling distribution of $\bar{X}_n$ is approximately normal:

$$\bar{X}_n \approx N\big(50,\, 0.2^2\big)\,.$$

Step 3. Standardize.

$$P(49.5 < \bar{X}_n < 50.5) = P\left(\frac{49.5-50}{0.2} < Z < \frac{50.5-50}{0.2}\right) = P(-2.5 < Z < 2.5),$$

where $Z \sim N(0,1)$.

Step 4. Look up from the standard normal table.

$$P(-2.5 < Z < 2.5) = \Phi(2.5) - \Phi(-2.5) = 0.9938 - 0.0062 = 0.9876.$$

Final answer:

$$\boxed{P(49.5 < \bar{X}_n < 50.5) \approx 0.9876.}$$

Interpretation. Even though individual screw lengths vary with standard deviation 2, the sample mean of 100 screws is tightly concentrated around 50 with standard deviation 0.2. The CLT justifies treating that mean as nearly normal — letting us use the normal CDF for quick probability estimates.

When sample size is large, sample mean is approximately normal.

24 Lecture XXIII: Weak Law of Large Numbers

We just saw that as we add more and more random variables, the distribution of their average becomes sharply concentrated around its mean — and even approximately normal. But we can make an even stronger statement: not only do those averages cluster near the mean, the event that the cluster steers away from the mean is ultimately impossible as the sample size increases. That's the Law of Large Numbers.

More precisely, the Central Limit Theorem tells us that, for large n, the **distribution** of the sample mean

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$$

is approximately normal, centered at the true mean μ, and with standard deviation $\sigma/\sqrt{n}$. As n increases, this spread $\sigma/\sqrt{n}$ shrinks — the distribution of $\bar{X}_n$ becomes increasingly concentrated near μ.

The Law of Large Numbers pushes this idea to the limit: not only does $\bar{X}_n$ *concentrate* near μ, it actually *converges to* μ. That is, with enough samples, the sample mean becomes a reliable estimator of the true mean.

24.1 Theorem Statement and Example

Theorem 24.1 (Weak Law of Large Numbers)**.** *Let $X_1, X_2, \ldots$ be i.i.d. random variables with mean μ and variance $\sigma^2 < \infty$. Then, for every $\varepsilon > 0$,*

$$P\big(|\bar{X}_n - \mu| > \varepsilon\big) \to 0 \quad \text{as } n \to \infty.$$

*Equivalently, we write $\bar{X}_n \xrightarrow{P} \mu$ (this reads: $\bar{X}_n$ converges to μ **in probability**).*

Proof. We need a technical lemma here that ended up becoming equally famous. We see an example first. □

Remark 24.1. *In simple terms, think of $|\bar{X}_n - \mu| > \varepsilon$ as a "bad set" or "bad event", in the sense that the sample mean $\bar{X}_n$ deviates away from the true mean μ by some distance ε. The Weak Law of Large Numbers says that the probability of these "bad events" is small when you take enough samples.*

Example 24.1. Let X_i be the outcome of the i-th coin toss:

$$X_i = \begin{cases} 1, & \text{if Heads,} \\ 0, & \text{if Tails.} \end{cases}$$

Then $\mathbb{E}[X_i] = p$ (the probability of Heads), and $\mathrm{Var}(X_i) = p(1-p)$.

Define the sample mean

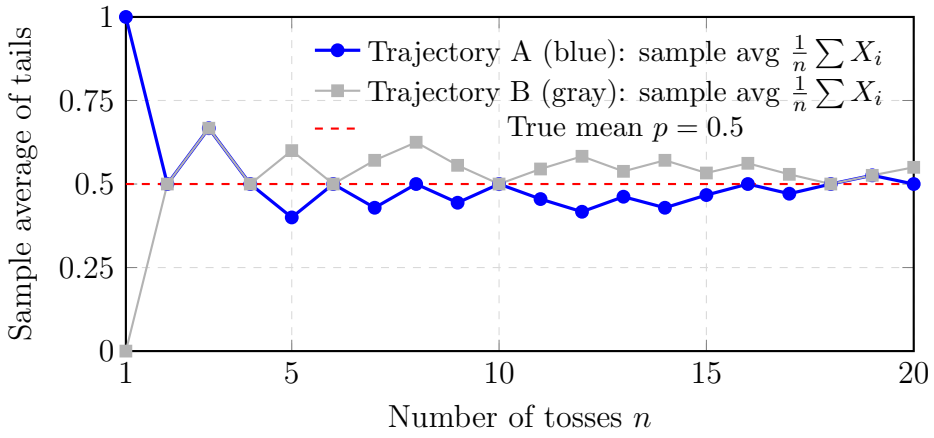
$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

This is simply the *sample proportion* of Heads after n tosses.

By the LLN,

$$\bar{X}_n \xrightarrow{P} p.$$

That is, as we toss the coin more and more times, the fraction of Heads stabilizes near the true probability p.



Generating sequences (1 = Tail, 0 = Head):

Trajectory A (blue):

{ 1, 0, 1, 0, 0, 1, 0, 1, 0, 1, 0, 0, 1, 0, 1, 1, 0, 1, 1, 0 }

Trajectory B (gray):

{ 0, 1, 1, 0, 1, 0, 1, 1, 0, 0, 1, 1, 0, 1, 0, 1, 0, 0, 1, 1 }

Figure 8: Two sample paths of the running sample average $\frac{1}{n}\sum_{i=1}^{n} X_i$ for coin tosses (1 = tail). Trajectory A (blue dots) and Trajectory B (gray squares) are two different realizations of 20 tosses each; their exact generating sequences are printed below. By toss 20 both trajectories are close to the true mean $p = 0.5$ (red dashed line), illustrating the Weak Law of Large Numbers.

The LLN tells us that averaging many independent random outcomes "cancels out" random fluctuations. In practice:

- Repeated measurements of the same quantity converge to the true value.
- The empirical proportion in large samples approximates the true probability.
- Random noise averages away.
- The LLN guarantees that averages are reliable in the long run.

24.2 Proof of WLLN

Before we prove the weak law of large numbers, we establish two of the most important technical lemmas in probability theory. It is also instructive here to note that a mathematical result may be set up in several stages.

Lemma 24.1 (Markov's Inequality)**.** *Let X be a non-negative random variable with finite expectation $\mathbb{E}[X] < \infty$. Then, for any $a > 0$,*

$$\mathbb{P}(X \geq a) \leq \frac{\mathbb{E}[X]}{a}.$$

Proof. Define the indicator function

$$\mathbf{1}_{\{X \geq a\}} = \begin{cases} 1, & X \geq a, \\ 0, & X < a. \end{cases}$$

Since $X \geq 0$, we have

$$X \geq a\mathbf{1}_{\{X \geq a\}}.$$

Taking expectations on both sides gives

$$\mathbb{E}[X] \geq a\,\mathbb{E}[\mathbf{1}_{\{X \geq a\}}] = a\,\mathbb{P}(X \geq a),$$

which yields the desired inequality. □

Lemma 24.2 (Chebyshev's Inequality)**.** *Let X be a random variable with finite mean $\mu = \mathbb{E}[X]$ and variance $\sigma^2 = \mathrm{Var}(X)$. Then, for any $k > 0$,*

$$\mathbb{P}(|X - \mu| \geq k) \leq \frac{\sigma^2}{k^2}.$$

Proof. Apply Markov's inequality to the non-negative random variable $Y = (X - \mu)^2$ with threshold $a = k^2$:

$$\mathbb{P}(|X - \mu| \geq k) = \mathbb{P}(Y \geq k^2) \leq \frac{\mathbb{E}[Y]}{k^2} = \frac{\mathrm{Var}(X)}{k^2}.$$

□

Theorem 24.2 (Weak Law of Large Numbers — iid, finite variance)**.** *Let $X_1, X_2, \ldots$ be i.i.d. random variables with mean $\mathbb{E}[X_i] = \mu$ and variance $\mathrm{Var}(X_i) = \sigma^2 < \infty$. Define the sample mean*

$$\overline{X}_n := \frac{1}{n}\sum_{i=1}^{n} X_i.$$

Then for every $\varepsilon > 0$,

$$P\big(|\overline{X}_n - \mu| > \varepsilon\big) \longrightarrow 0 \quad \text{as } n \to \infty,$$

i.e. $\overline{X}_n \xrightarrow{P} \mu$.

Proof. Compute the mean and variance of $\overline{X}_n$. By linearity,

$$\mathbb{E}[\overline{X}_n] = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}[X_i] = \mu.$$

Since the X_i are independent,

$$\mathrm{Var}(\overline{X}_n) = \mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\sum_{i=1}^{n} \mathrm{Var}(X_i) = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}.$$

Now apply Chebyshev's inequality with $V = \overline{X}_n$:

$$P\big(|\overline{X}_n - \mu| \geq \varepsilon\big) \leq \frac{\mathrm{Var}(\overline{X}_n)}{\varepsilon^2} = \frac{\sigma^2}{n\varepsilon^2}.$$

The right-hand side goes to 0 as $n \to \infty$; hence $P(|\overline{X}_n - \mu| > \varepsilon) \to 0$, which proves the Weak Law. □

Important Remark 24.1. *This proof only uses finiteness of the second moment and independence (or, more generally, pairwise uncorrelatedness together with a suitable variance bound). It provides a quantitative rate: the Chebyshev bound is $O(1/n)$, which is often loose but sufficient to show convergence in probability.*

24.3 What the WLLN says and does NOT say

The Weak Law of Large Numbers tells us that

$$P(|\overline{X}_n - \mu| > \varepsilon) \to 0$$

that is, the probability of a large deviation from the mean becomes small *as n grows large.*

However, it is crucial to understand what the WLLN *does not* say:

- It does **not** guarantee that the sample average will be close to μ for any finite n. Even for $n = 100$ or $n = 1{,}000$, large deviations can still occur.
- It does **not** rule out **infinitely many** large deviations. It only says that the *probability* of such deviations eventually becomes small.
- In other words, the WLLN ensures that the "bad" events

  $$S_n(\varepsilon) = \{\omega : |\overline{X}_n(\omega) - \mu| > \varepsilon\}$$

  become *rare*, but does not prevent them from happening infinitely often across the sequence.

The Strong Law of Large Numbers, which you will study in the future, strengthens this statement to an almost sure convergence:

$$\overline{X}_n \xrightarrow{\text{a.s.}} \mu,$$

assuming the second moment of each X_n is finite. The proof is much more delicate and requires deeper knowledge of mathematics. Let's meet there!

The Strong Law of Large Numbers strengthens the Weak Law by saying that, after some point, the sample average will stay close to the true mean forever (except possibly for a few early wobbles). In other words, while the Weak Law says that large deviations become rare, the Strong Law says that large deviations happen only finitely many times.