IEOR 3658: Probability for Engineers Fall 2022

Probability Theory for Engineers

Prof. Daniel Lacker *Scribe: Ekene Ezeunala*

1 Sample Spaces (06/09)

1.1 Introduction

Definition 1.1 (Probability Space)**.** *An ordered triple* $(\Omega, \Sigma, \mathbb{P})$ *is called a probability space if*

- Ω *is a sample space;*
- Σ *is a* σ*-algebra of measurable subsets (events) of* Σ *(note to self: study measure theory soon)*
- $\mathbb{P}$ *is a probability measure on* Σ*; that is,* $\mathbb{P}$ *satisfies the following axioms:*
  1. *For any* $A \in \Sigma$ *there exists a number* $\mathbb{P}(A) \geq 0$*,*
  2. *The probability measure is normalized, that is,* $\mathbb{P}(\Sigma) = 1$*,*
  3. $\mathbb{P}$ *is a countable additive measure satisfying* $\mathbb{P}\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} \mathbb{P}(A_i)$*.*

We will discuss this definition in future classes. For now, we call the sample space Ω the *set* of *all* possible *outcomes* in some experiment.

Example 1.1. Flip a coin. Then $\Omega = \{\text{heads}, \text{tails}\}$, or $\Sigma = \{\text{H}, \text{T}\}$.

Example 1.2. Roll a (standard, six-sided) die. Then $\Omega = \{1, 2, 3, 4, 5, 6\}$.

Example 1.3. Complete IEOR 3658 with a grade, where this grade is a standard letter grade. Then $\Omega = \{$A+, A, A-, B+, B, B-, C+, C, C-, D, F$\}$.

Example 1.4. Flip three coins. Then Ω is the set of sequences or strings of length 3 made of characters H and T, or

$$\Omega = \{\text{HHH, HHT, HTH, HTT, THH, THT, TTH, TTT}\} = \{\text{H, T}\}^3$$

Example 1.5. Flip n coins, where n is a given integer > 0. Then $\Omega = \{\text{H, T}\}^n$. Note that Ω has 2^n elements. (The proof is left as an exercise.)

1.1.1 Language

- Exactly *one* of the *outcomes*—elements of Ω—will *occur* or *happen*.
- A *set* of *outcomes*—that is, a subset of Ω—is called an *event*.

- We write $E \subset \Omega$ to mean E *is a proper subset of* Ω, and likewise write $E \subseteq \Omega$ to mean E *is a subset of* Ω.
- Two extreme cases worthy of note: $\Omega \subseteq \Omega$, and $\emptyset \subset \Omega$.

Suppose we rolled a boring fair die with $\Omega = \{1, 2, 3, 4, 5, 6\}$. Then:

- 2 is an outcome, not an event.
- 7 is not an outcome.
- The event that we roll an even number is $A = \{2, 4, 6\}$.
- The event that we roll at least three is $B = \{3, 4, 5, 6\}$.
- $\{2\}$ is an event, not an outcome.

Events can overlap, or occur simultaneously; outcomes cannot. This is another way of saying that two things cannot happen at the same time, but two things of the same kind can. We say an event occurs if the outcome that results belongs to the event. For example, if I roll a 4, then both events A and B occur.

Example 1.6. Flip 4 coins, so that $\Omega = \{\text{H, T}\}^4$. Then the event that I get the same number of heads and tails is {HHTT, HTHT, HTTH, THTH, TTHH, THHT}.

1.2 Manipulating events

The complement of some event A with respect to the universe Ω is the set $\{x \in \Omega \mid x \notin A\}$ of all outcomes of Ω that do not belong to A, and is written A^c. Note that $\Omega^c = \emptyset$.

The union of two events A and B is the set of all outcomes that belong to at least A or B, and is denoted $A \cup B = \{x \mid x \in A \text{ or } x \in B\}$.

The intersection of two events A and B is the set of all outcomes that belong to both A or B, and is denoted $A \cap B = \{x \mid x \in A \text{ and } x \in B\}$.

The difference between two events $A - B$ is the intersection of the complement of B and A, or $A - B = A \cap B^c$.

The exclusive or of two events is the union of the two events excluding the intersection of the events, or $(A \cup B) \cap (A \cap B)^c$.

Example 1.7. Roll a die. Consider the events that we obtain an even number, A, and that we obtain a number greater than or equal to 3, B. Then A^c is the event that we obtain an odd number, with outcome set $\{1, 3, 5\}$, and $A \cap B$ is the event that we obtain an even number ≥ 3, with outcome set $\{4, 6\}$.

Note that we always have $A \cup A^c = \Omega$ and $A \cap A^c = \emptyset$. Furthermore, for some events A, B, if $A \subset B$, we say A implies B.

We might extend some of these notions to n events $A_1, A_2, \ldots, A_n$:

- **Union**: If at least one of the A_i occurs, then

$$A_1 \cup A_2 \cup \ldots \cup A_n = \bigcup_{i=1}^{n} A_i.$$

- **Intersection**: If all of the A_i occur, then

$$A_1 \cap A_2 \cap \ldots \cap A_n = \bigcap_{i=1}^{n} A_i.$$

1.3 Event properties

Let A, B, and C be events.

- **Commutativity/symmetry**: Order of union and intersection is invariant.

$$A \cup B = B \cup A, \qquad A \cap B = B \cap A.$$

- **Associativity**: Union and intersection associate over the events.

$$A \cup (B \cup C) = (A \cup B) \cup C = A \cup B \cup C$$

$$A \cap (B \cap C) = (A \cap B) \cap C = A \cap B \cap C$$

- **Distributivity**: Union and intersection distribute over each other.

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

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

- **De Morgan's Laws**: Consider events A_i. Then,

$$\left(\bigcup_n A_i\right)^c = \bigcap_n A_i^c, \qquad \left(\bigcap_n A_i\right)^c = \bigcup_n A_i^c.$$

Definition 1.2 (Disjoint events)**.** *Events A and B are said to be disjoint—or mutually exclusive—if $A \cap B = \emptyset$. We say that A_i events are disjoint if, for every i, j such that $1 \leq i, j \leq n$, it is true that $A_i \cap A_j = \emptyset$.*

Example 1.8. If all the partitioned events A_i are disjoint *and* they exhaust all possibilities, then $A_1 \cup A_2 \cup \ldots \cup A_n = \Omega$. In other words, exactly one of these events will occur.

Example 1.9. Roll a die. Then there are two possible outcomes that can result: $A = \{\text{even-numbered outcomes}\}$ and $B = \{\text{odd-numbered outcomes}\}$. At most one of these events can occur, so A and B are disjoint.

Example 1.10. In general, A and A^c are disjoint, and in fact form a partition. Additionally, if $A \cap B = \emptyset$, then $A \subset B^c$ and $B \subset A^c$.

2 Axioms of probability (13/09)

Let Ω be a sample space. A *probability measure* on Ω is a function $\mathbb{P}$ which assigns to each *event* a real number (called a *probability*) satisfying:

1. **(Nonnegativity.)** $\mathbb{P}(A) \geq 0$ for every event A.[1]
2. **(Normalization.)** The probability of the entire sample space Ω is equal to 1; that is, $\mathbb{P}(\Omega) = 1$.
3. **(Additivity.)** If A_i is a sequence of n disjoint events, then the probability of their union satisfies
$$\mathbb{P}(A_1 \cup A_2 \cup \ldots \cup A_n) = \mathbb{P}(A_1) + \mathbb{P}(A_2) + \ldots + \mathbb{P}(A_n).$$

2.1 Properties

- $\mathbb{P}(A) + \mathbb{P}(A^c) = 1$, or $\mathbb{P}(A) = 1 - \mathbb{P}(A^c)$.

  Proof. Since A and A^c are disjoint, $\mathbb{P}(A \cup A^c) = \mathbb{P}(A) + \mathbb{P}(A^c)$ by (3). On the other hand, $A \cup A^c = \Omega$, so $\mathbb{P}(A) + \mathbb{P}(A^c) = \mathbb{P}(\Omega) = 1$. ■

- If A implies B (B is guaranteed to happen if A happens), that is, $A \subset B$, then $\mathbb{P}(A) \leq \mathbb{P}(B)$.

  Proof. Since A and A^c are disjoint and $A \subset B$, we might write B as a union of these disjoint events, so that $B = A \cup (B \cap A^c)$. Thus:
  $$\mathbb{P}(B) = \mathbb{P}(A \cup (B \cap A^c)) = \mathbb{P}(A) + \mathbb{P}(B \cap A^c) \geq \mathbb{P}(A),$$
  since $\mathbb{P}(B \cap A^c) \geq 0$ by (1). ■

- For events A and B which are not necessarily disjoint, we have
$$\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B).$$

  Proof. The events $A \cap B^c$, $B \cap A^c$, and $A \cap B$ are disjoint. (Much of the motivation for the relevance of this in the proof comes by drawing the Venn diagram for the events.) We observe that
  $$\begin{aligned} A &= (A \cap B^c) \cup (A \cap B) \\ B &= (B \cap A^c) \cup (A \cap B) \\ A \cup B &= (A \cap B^c) \cup (B \cap A^c) \cup (A \cap B) \end{aligned}$$
  By additivity,
  $$\begin{aligned} \mathbb{P}(A) &= \mathbb{P}(A \cap B^c) + \mathbb{P}(A \cap B) \\ \mathbb{P}(B) &= \mathbb{P}(B \cap A^c) + \mathbb{P}(A \cap B) \\ \mathbb{P}(A \cup B) &= \mathbb{P}(A \cap B^c) + \mathbb{P}(B \cap A^c) + \mathbb{P}(A \cap B), \end{aligned}$$
  from which we obtain $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$, as desired. ■

[1] $\mathbb{P}(A) = 0 \not\Longrightarrow A$ is an impossible event. More on this later.

- $\mathbb{P}(A \cup B \cup C) = \mathbb{P}(A) + \mathbb{P}(A^c \cap B) + \mathbb{P}(A^c \cap B^c \cap C)$. The proof is left as an exercise to the reader.

Example 2.1. Flip a coin. Then $\Omega = \{\text{H}, \text{T}\}$. There are $2^{|\Omega|} = 4$ events, namely $\emptyset, \Omega, \{\text{H}\}, \{\text{T}\}$. The most natural probability measure is $\mathbb{P}(\{\text{H}\}) = \mathbb{P}(\{\text{T}\}) = 1/2$, $\mathbb{P}(\emptyset) = 0$, and $\mathbb{P}(\Omega) = 1$.
Another choice would be to model a *biased* coin, so that for $q \in [0, 1]$, we have $\mathbb{P}(\{\text{H}\}) = q$, $\mathbb{P}(\{\text{T}\}) = 1 - q$, $\mathbb{P}(\emptyset) = 0$, and $\mathbb{P}(\Omega) = 1$.

Example 2.2. *Carl tests positive for COVID-19. He has no symptoms and suspects it might be a false positive. Since a year and a half ago (78 weeks ago), he has gotten a rapid test every week, as required by his job. The false positive rate for rapid tests is* $1/250 = 0.4\%$. *Carl argues that, assuming he does not have COVID-19, then the probability that he gets a false positive at some point is* $78/250 = 0.39\%$. *This is because, he says, each of the 78 tests has a* $1/250$ *chance of a false positive. What is wrong with Carl's logic?*
First we see that extending his logic to, say, 500 tests, Carl would say that the probability is $500/250 = 2$, which is nonsense.
A better answer is that Carl's claim that

$$\mathbb{P}(A_1 \cup A_2 \cup \ldots \cup A_{78}) = \mathbb{P}(A_1) + \mathbb{P}(A_2) + \ldots + \mathbb{P}(A_{78})$$

is not valid because A_1 and A_2 are not disjoint events.

3 Probability basics; Conditioning (15/09)

A simple recipe for defining a probability measure. For finite sample spaces Ω, just specify a probability for each individual outcome, as long as the probabilities are non-negative and add up to 1. Then compute the probability of an event $A \subset \Omega$ by summing probabilities of outcomes belonging to the event A.

Example 3.1. In a coin flip, $\Omega = \{\text{H}, \text{T}\}$. Define $\mathbb{P}(H) = \mathbb{P}(T) = 1/2$. Then $\mathbb{P}(\Omega) = \mathbb{P}(\{\text{H}, \text{T}\}) = \mathbb{P}(H) + \mathbb{P}(T) = 1/2 + 1/2 = 1$.

Two popular problems for which defining a nice sample space and working systematically are immensely helpful include the boy/girl problem and the Monty Hall problem.

Example 3.2. Three pairs of socks, coloured red, green, or blue, are unsorted in a drawer. Suppose we wanted to know the probability that two socks grabbed at random are of the same colour. To do this, we model six distinct socks named $R_1, R_2, G_1, G_2, B_1, B_2$. There are $\binom{6}{2} = \frac{6!}{4! \cdot 2!} = 15$ ways to grab 2 socks. All grabbing outcomes are equally likely in this scenario, so we might say $\mathbb{P}(\omega) = 1/15$ for each outcome ω.
The event that I get two matching socks is then $A = \{R_1R_2, B_1B_2, G_1G_2\}$. Thus

$$\mathbb{P}(A) = \mathbb{P}(R_1R_2) + \mathbb{P}(G_1G_2) + \mathbb{P}(B_1B_2) = 3 \cdot 1/15 = 3/15.$$

Example 3.3. Flip 2 biased coins with heads probability q, where $0 \leq q \leq 1$. Model this by defining $\Omega = \{H, T\}^2 = \{HH, TT, TH, HT\}$, and

$$\mathbb{P}(HH) = q \cdot q = q^2, \qquad \mathbb{P}(TT) = (1-q) \cdot (1-q) = (1-q)^2, \qquad \mathbb{P}(HT) = \mathbb{P}(TH) = q(1-q).$$

Then:

$$\mathbb{P}(HH) + \mathbb{P}(TT) + \mathbb{P}(HT) + \mathbb{P}(TH) = q^2 + (1-q)^2 + 2q(1-q) = (q + 1 - q)^2 = 1.$$

Suppose we define A to be the event that the first flip is a heads. Then $\mathbb{P}(A) = \mathbb{P}(HH) + \mathbb{P}(HT) = q^2 + q(1-q) = q$.
Suppose we inquired about the probability of getting at least 1 heads. One acceptable method is to define B as the event $\{HH, HT, TH\}$, so that

$$\mathbb{P}(B) = \mathbb{P}(HH) + \mathbb{P}(HT) + \mathbb{P}(TH) = q^2 + 2q(1-q) = 2q - q^2.$$

Alternatively we might work with $B^c = \{TT\}$, and observe that $\mathbb{P}(B^c) = \mathbb{P}(TT) = (1-q)^2$. Thus the original event of interest has $\mathbb{P}(B) = 1 - \mathbb{P}(B^c) = 1 - (1-q)^2$.

Example 3.4. Consider Exercise 3.2 above. The probability of getting two reds is $\mathbb{P}(R_1R_2) = 1/15$ (or $2/6 \cdot 1/5 = 1/15$; initially there are 2 red socks among all the 6 of them, but after the first is drawn without replacement there is 1 red sock out of 5).
Similarly, the probability of getting a red and a green is $\mathbb{P}(R_1G_1, R_1G_2, R_2G_1, R_2G_2) = 4 \cdot 1/15 = 4/15$ (or $4/6 \cdot 2/5 = 4/15$; can you explain why?).

3.1 Conditional Probability

Consider the motivating example below.

Example 3.5. Flip 3 coins, so $\Omega = \{H, T\}^3$ and $\mathbb{P}(\omega) = 1/8$ for each outcome ω. Suppose we know that there is at least one heads. Then our sample size "changes":

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

The condition that there is at least one heads removes the TTT outcome from consideration. Thus, the probability that there are at least two heads is then $4/7$.

Definition 3.1 (Conditional probability)**.** *For events A and B, with $\mathbb{P}(B) > 0$, the conditional probability of A given B, or $\mathbb{P}(A \mid B)$, is defined by*

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

Note that $\mathbb{P}(B \mid B) = \dfrac{\mathbb{P}(B \cap B)}{\mathbb{P}(B)} = 1$. *Additionally,* $\mathbb{P}(A \mid B) \neq \mathbb{P}(B \mid A)$.

Example 3.6. *On what day of the week do you do your IEOR 3658 homework due on Tuesday?*

Outcomes	*Mon.*	*Tues.*	*Wed.*	*Thurs.*	*Fri.*	*Sat.*	*Sun.*
Probabilities	*0.20*	*0.50*	*0.01*	*0.04*	*0.07*	*0.08*	*0.10*

Suppose your friend tells you he has not started yet, and it is Sunday morning. What is the probability that he does the homework before Tuesday, that is, on Sunday or Monday? Probabilistically speaking, this is asking for the probability of $A = \{\text{Sun.}, \text{Mon.}\}$ given $B = \{\text{Sun.}, \text{Mon.}, \text{Tues.}\}$, i.e. $\mathbb{P}(A \mid B)$. As $A \subset B$, $A \cap B = A$, so

$$\mathbb{P}(A \mid B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)} = \frac{\mathbb{P}(A)}{\mathbb{P}(B)} = \frac{\mathbb{P}(\text{Sun.}) + \mathbb{P}(\text{Mon.})}{\mathbb{P}(\text{Sun.}) + \mathbb{P}(\text{Mon.}) + \mathbb{P}(\text{Tues.})} = \frac{0.10 + 0.20}{0.10 + 0.20 + 0.50} = \frac{3}{8}.$$

4 Conditional probability (20/09)

4.1 Properties of $\mathbb{P}(A \mid B)$

- **(Measure.)** As a function of the event A, $\mathbb{P}(A \mid B)$ is itself a probability measure. That is, given events B, E, F, it satisfies the same rules of probability:

  1. $\mathbb{P}(\Omega \mid B) = 1$. Check that $\mathbb{P}(\Omega \mid B) = \dfrac{\mathbb{P}(\Omega \cap B)}{\mathbb{P}(B)} = \dfrac{\mathbb{P}(B)}{\mathbb{P}(B)} = 1$.
  2. $\mathbb{P}(A \mid B) \geq 0$ for any event A.
  3. If E and F are *disjoint*, then $\mathbb{P}(E \cap F \mid B) = \mathbb{P}(E \mid B) + \mathbb{P}(F \mid B)$.

     Proof. The key step is distributing over the union and intersection of the sets involved:

$$\begin{aligned}
\mathbb{P}(E \cap F \mid B) &= \frac{\mathbb{P}((E \cup F) \cap B)}{\mathbb{P}(B)} \\
&= \frac{\mathbb{P}((E \cap B) \cup (F \cap B))}{\mathbb{P}(B)} \\
&= \frac{\mathbb{P}((E \cap B) + \mathbb{P}(F \cap B)}{\mathbb{P}(B)} \\
&= \frac{\mathbb{P}((E \cap B)}{\mathbb{P}(B)} + \frac{\mathbb{P}((F \cap B)}{\mathbb{P}(B)} \\
&= \mathbb{P}(E \mid B) + \mathbb{P}(F \mid B).
\end{aligned}$$

     and the proof is done. ■

     Similarly, $\mathbb{P}(A^c \mid B) = 1 - \mathbb{P}(A \mid B)$. However, $\mathbb{P}(A \mid B^c) \neq 1 - \mathbb{P}(A \mid B)$.

- **(Multiplicativity.)** $\mathbb{P}(A \cap B) = \mathbb{P}(B \cap A) = \mathbb{P}(B \mid A)\mathbb{P}(A) = \mathbb{P}(A \mid B)\mathbb{P}(B)$.

- **(Bayes' Theorem.)** A beautiful result when you meditate on it: $\mathbb{P}(A \mid B) = \dfrac{\mathbb{P}(B \mid A)\mathbb{P}(A)}{\mathbb{P}(B)}$.

- **(Law of total probability.)** Let us attempt to develop a relationship between *conditional and unconditional probabilities.* Suppose A and B are two events with $\mathbb{P}(B) \neq 0$. Then $A = (A \cap B^c) \cup (A \cap B)$, so that

$$\begin{aligned}\mathbb{P}(A) &= \mathbb{P}(A \cap B^c) + \mathbb{P}(A \cap B) \\ &= \mathbb{P}(A \mid B^c)\mathbb{P}(B^c) + \mathbb{P}(A \mid B)\mathbb{P}(B) \\ &= \mathbb{P}(A \mid B^c)(1 - \mathbb{P}(B)) + \mathbb{P}(A \mid B)\mathbb{P}(B) \\ &= \mathbb{P}(B)(\mathbb{P}(A \mid B) - \mathbb{P}(A \mid B^c)) + \mathbb{P}(A \mid B^c).\end{aligned}$$

  Extension to multiple events. Let $B_1, B_2, \ldots, B_i, \ldots, B_n$ events form a partition, so that $\bigcap_{i=1}^{n} B_i = \emptyset$ (i.e. the B_i are disjoint) and $\bigcup_{i=1}^{n} B_i = \Omega$ (i.e. the B_i are exhaustive), then $A = \bigcup_{i=1}^{n} A \cap B_i$ (with all the $A \cap B_i$ disjoint), so that

$$\mathbb{P}(A) = \sum_{i=1}^{n} \mathbb{P}(A \cap B_i) = \sum_{i=1}^{n} \mathbb{P}(A \mid B_i)\mathbb{P}(B_i).$$

4.2 Example: Medical Diagnostic Tests

Given that I test positive for COVID, what is the probability that I actually have COVID?

Let A be the event that I have COVID and B the event that I test positive for COVID. Suppose that the false positive rate (the probability that I test positive given that I have COVID) is $q = \mathbb{P}(B \mid A^c)$, and that the false negative rate (the probability that I do not test positive for COVID given that I actually have COVID) is $r = \mathbb{P}(B^c \mid A)$. Furthermore, let the fraction of the population who already have COVID be $p = \mathbb{P}(A)$. The *goal* is then to find the probability that I have COVID given that I test positive, or $\mathbb{P}(A \mid B)$.

From Bayes' rule, we know that $\mathbb{P}(A \mid B) = \dfrac{\mathbb{P}(B \mid A)\mathbb{P}(A)}{\mathbb{P}(B)}$.

We are not given $\mathbb{P}(A \mid B)$, but we can compute it easily: $\mathbb{P}(B \mid A) = 1 - \mathbb{P}(B^c \mid A) = 1 - r$.

We are not given $\mathbb{P}(B)$, so we compute it using the law of total probability:

$$\begin{aligned}\mathbb{P}(B) &= \mathbb{P}(B \mid A)\mathbb{P}(A) + \mathbb{P}(B \mid A^c)\mathbb{P}(A^c) \\ &= (1 - \mathbb{P}(B^c \mid A))\mathbb{P}(A) + \mathbb{P}(B \mid A^c)(1 - \mathbb{P}(A)) \\ &= (1 - r)p + q(1 - p).\end{aligned}$$

Putting it all together, we get

$$\boxed{\mathbb{P}(A \mid B) = \frac{p(1 - r)}{p(1 - r) + q(1 - p)}.}$$

Example 4.1. If $p = q = r =$ say $1/10000$ on day 1, then

$$\mathbb{P}(A \mid B) = \frac{p(1-p)}{p(1-p) + p(1-p)} = \frac{1}{2}.$$

Example 4.2. Examine the role of prevalence in the value of $\mathbb{P}(A \mid B)$ by plotting its graph out as a function of p for values like $q = r = 1/2$, $q = r = 1/100 < 1/2$, and $q = r = 99/100 > 1/2$. This is left as an exercise.

5 Independence (22/09)

Let A and B be events on some sample space Ω. What does it mean to say that *A and B are independent*?

Definition 5.1 (Independence)**.** *We say A and B are independent events if* $\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B)$.

Note that independence is symmetric, i.e.

$$A \text{ is independent of } B \iff B \text{ is independent of } A \iff A \text{ and } B \text{ are independent.}$$

Observation. If $\mathbb{P}(B) > 0$, recall the definition of conditional probabilities, namely $\mathbb{P}(A \mid B) = \dfrac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}$. If A and B are *independent*, then $\mathbb{P}(A \mid B) = \dfrac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)} = \dfrac{\mathbb{P}(A)\mathbb{P}(B)}{\mathbb{P}(B)} = \mathbb{P}(A)$.
In other words, if $\mathbb{P}(B) > 0$, then *A and B are independent* if and only if $\mathbb{P}(A \mid B) = \mathbb{P}(A)$. This makes sense intuitively, because if A and B are independent, knowing that B has occurred does not make A any more or less likely to occur.
If $\mathbb{P}(A), \mathbb{P}(B) > 0$, then the following are equivalent:

$$\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B), \qquad \mathbb{P}(A \mid B) = \mathbb{P}(A), \qquad \mathbb{P}(B \mid A) = \mathbb{P}(B).$$

Example 5.1. Roll two dice. The sample space is $\Omega = \{1, 2, 3, 4, 5, 6\}^2$. All outcomes $\omega \in \Omega$ are equally likely, so $\mathbb{P}(\omega) = 1/36$. Let A be the event that the first roll is a 3 and B the event that the second row is a 4.
Are A and B independent? Intuitively, *yes*. Mathematically,

$$A = \{(3,1), (3,2), (3,3), (3,4), (3,5), (3,6)\}, \qquad B = \{(1,4), (2,4), (3,4), (4,4), (5,4), (6,4)\}.$$

Each event has six outcomes, so $\mathbb{P}(A) = \mathbb{P}(B) = 6 \cdot 1/36 = 1/6$. Alternatively, the event $A \cap B = \{(3,4)\}$, so $\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B)$, from which we get $\mathbb{P}(A \cap B) = 1/36$.

Example 5.2. Let C be the event that the first roll is even, with the same event A. Are A and C independent? *No.* Instead, A and C are disjoint (convince yourself of this), so $\mathbb{P}(A \mid C) = 0$. Note that $\mathbb{P}(A) = 1/6 \neq 0 = \mathbb{P}(A \mid C)$, so these events are not independent.

Example 5.3. Let D be the event that the sum of the two rolls equals 7, with the same event A. Are A and D independent? *Yes!* But this is not immediately obvious. Let us convince ourselves of this using two arguments.

Argument 1: Check that, by listing all the possible outcomes,

$$A = \{(3,1),(3,2),(3,3),(3,4),(3,5),(3,6)\} \text{ and } D = \{(1,6),(2,5),(3,4),(4,3),(5,2),(6,1)\}.$$

Both events have 6 outcomes, so $\mathbb{P}(A) = \mathbb{P}(D) = 6 \cdot 1/36 = 1/6$. Also, $\mathbb{P}(A \cap D) = 1/36 = 1/6 \cdot 1/6 = \mathbb{P}(A)\mathbb{P}(D)$, verifying independence.

Argument 2: Intuitively, $\mathbb{P}(D \mid A) = \mathbb{P}(\text{sum} = 7 \mid \text{the first roll is 3})$, which, by simple logic is the same as saying that $\mathbb{P}(D \mid A) = \mathbb{P}(\text{the second roll is 4} \mid \text{the first roll is 3}) = \mathbb{P}(B \mid A) = \mathbb{P}(B)$ as previously shown. Now $\mathbb{P}(D) = \mathbb{P}(B) = 1/6$, so independent.

Interesting point. If E is the event that the sum is 8, or any other number from 2 through 12 other than 7, then A and E are not independent.

Definition 5.2 (Pairwise independence)**.** *Let $E_1, E_2, \ldots, E_n$ be events on the same sample space Ω. We say that they are pairwise independent if the pair E_i and E_j are independent for each $i, j = 1, \ldots, n$ with $i \neq j$, that is, $\mathbb{P}(E_i \cap E_j) = \mathbb{P}(E_i)\mathbb{P}(E_j)$.*

Definition 5.3 (Joint independence)**.** *We say that the E_i events are jointly independent if*

$$\mathbb{P}\left(\bigcap_{\tau=1}^{k} E_{i_\tau}\right) = \prod_{\tau=1}^{k} \mathbb{P}(E_{i_\tau})$$

for any $k \in [1, n]$ (any number of the n events) and any $i_\tau \in [1, n]$ with $\tau \in [1, k]$ (the indices of the n events).

Fact. Joint independence implies pairwise independence, but *not* vice versa. We might intuit this by observing that A, B, and C can be pairwise independent, yet the knowledge of C occurring could influence the "dependence" between A and B.

Example 5.4. Roll two dice as before, with events A, B, D as before. We saw that $\mathbb{P}(A) = \mathbb{P}(B) = \mathbb{P}(D) = 1/6$. Furthermore, A and B are independent, and A and D are independent, so B and D are independent. Thus, events A, B, D are pairwise independent.
However, they are not jointly independent, because $\mathbb{P}(A \cap B \cap D) = \mathbb{P}(\{3,4\}) = 1/36$, whereas $\mathbb{P}(A)\mathbb{P}(B)\mathbb{P}(D) = 1/6 \cdot 1/6 \cdot 1/6 = 1/216$.
Check that, in this case, $\mathbb{P}(D \mid A \cap B) = 1 \neq 1/6 = \mathbb{P}(D)$.

6 Counting problems (27/09)

6.1 Equally likely outcomes

Suppose we had a finite sample space Ω, and $\mathbb{P}$ make all outcomes equally likely—that is, $\mathbb{P}(\omega) = 1/|\Omega|$ for each $\omega \in \Omega$, where $|\Omega|$ is the number of elements of the set Ω, or the number of possible

outcomes. Then for any event E,

$$\mathbb{P}(E) = \text{sum of } \mathbb{P}(\omega) \text{ over all } w \in E = \frac{\text{no. of outcomes in } E}{\text{no. of outcomes total}} = \frac{|E|}{|\Omega|}.$$

6.2 Counting problems: basics

In obtaining the probability of discrete events where we must systematically count, it is often either to ask questions like *Is our sample space ordered (as in the case of social security numbers) or unordered (as in the case of poker hands)? Are outcomes obtained with (as in the case of rolling dice) or without (as in the case of drawing cards from a deck) replacement?*

Now suppose we have N "objects" which can be one of k different "states".

Example 6.1. N coins, each of which can be H or T, so $k = 2$.

Example 6.2. Roll N dice, each can result in $1, 2, 3, 4, 5, 6$, so $k = 6$.

Example 6.3. Each social security number (SSN) has $N = 9$ digits, each of which can be $0, 1, 2, \ldots, 9$, so $k = 10$.

Example 6.4. In how many ways can we assign k states to N ordered objects with replacement? Each state is constantly available for each object, so there are k^N ways.

Example 6.5. How many sequences of N coin flips are available? There are 2 states available for each of the N objects, so the result is 2^N ways.

Example 6.6. How many distinct SSN's are there? There are 10 choices for the first digit, 10 for the second, ..., 10 for the ninth. So there are $10 \cdot 10 \cdot \ldots \cdot 10 = 10^9$ distinct SSNs.

Example 6.7. In how many ways can we choose k distinct objects *in order and without replacement* from the N objects? For example, in how many ways can k students out of N students line up to exit a room? There are N choices for the first person who joins the line, $N-1$ choices for the second person (since the students are down by 1), $N-2$ choices for the third person, and so on until there are $N-(k-1) = N-k+1$ choices for the last person who joins the line. Thus the required number of ways is

$$\begin{aligned} N(N-1)(N-2)\cdots(N-k+1) &= \frac{N(N-1)\cdots(N-k+1)(N-k)(N-k-1)\cdots 3\cdot 2\cdot 1}{(N-k)(N-k-1)\cdots 3\cdot 2\cdot 1} \\ &= \frac{N!}{(N-k)!}. \end{aligned}$$

We can also write this as $k!\dfrac{N!}{(N-k)!k!} = k!\binom{N}{k}$, where the binomial coefficient N *choose* k is defined as $\binom{N}{k} = \dfrac{N!}{(N-k)!k!}$.

Example 6.8. In how many ways can k *unordered* objects be drawn/selected *without replacement* from the N total? After some thought, the result is $\binom{n}{k} = \dfrac{N!}{(N-k)!k!}$

Example 6.9. How many 5-card hands are possible in a standard 52-card deck? This is just choosing 5 states from 52 objects, or $\binom{52}{5} = \dfrac{52!}{47!\cdot 5!}$.

Note. $\binom{n}{k}$ differs from $k!\binom{N}{k}$ by a factor of $k!$

6.3 Example: Birthday paradox

There are N people in class. What is the probability that at least two people share the same birthday?

This is higher than you might expect! For $N = 80$, the probability is $\geq 99.9\%$. For $N = 23$, it is $\geq 50\%$. Let us compute this probability in terms of N. Let

$$p_N = \mathbb{P}(\text{at least 2 people out of } N \text{ have the same birthday}).$$

This is a bit messy to work with, so we try something simpler:

$$q_N = 1 - p_N = \mathbb{P}(\text{nobody has the same birthday out of } N).$$

Now the task becomes to choose $k = 365$ birthdays for N people. We may check that

$$\begin{aligned} q_N &= \frac{\text{no. of ways to choose } N \text{ distinct birthdays}}{\text{no. of ways to choose } N \text{ birthdays}} \\ &= \frac{365 \cdot 364 \cdot 363 \cdot \ldots \cdot (365 - N + 1)}{365^N} \\ &= \left(\frac{365!}{(365-N)!}\right) \Big/ 365^N, \end{aligned}$$

so $p_N = 1 - q_N = 1 - \left(\frac{365!}{(365-N)!}\right) \Big/ 365^N$. We can't get a simpler form of this, but it is easy to graph on a computer. Note that when N is large, 365^{-N} is small and $\frac{365!}{(365-N)!}$ is relatively smaller, so we might expect the probability to be larger. The behaviour of the probabilistic function is anomalous, and should spark some interest. This spark is left as an exercise.

6.4 Example: Poker hands

A standard deck has 52 cards: *4 suits*, namely diamonds, hearts, clubs, and spades, and *13 states* per suit, namely A, 2, 3, 4, 5, 6, 7, 8, 9, 10, J, Q, K.

Example 6.10. *What is the probability of drawing 3 of a kind? (More precisely, three cards have the same number, but not four, and the last two cards are not a pair.)*
We proceed in stages, bearing in mind that drawing 3 of a kind looks like drawing something of the form $\{c_1, c_1, c_1, c_2, c_3\}$, with $c_1, c_2, c_3 \in \{A, 2, \ldots, Q, K\}$.

- There are five cards in a hand. The number of ways to select unordered hands from a full deck is then $\binom{52}{5}$.
- There are $\binom{13}{1}$ ways to choose c_1, obviously.
- There are $\binom{12}{2}$ ways to choose c_2 and c_3 as there are 12 cards left after c_1 is selected.

- There are $\binom{4}{3}$ ways to choose the three suits for c_1, and $\binom{4}{1}$ ways to choose the suits for each of c_2 and c_3.

Putting everything together, we find that

$$\mathbb{P}(\text{3 of a kind}) = \frac{\text{no. of pair hands}}{\text{no. of possible hands total}} = \binom{13}{1}\binom{12}{2}\binom{4}{3}\binom{4}{1}\binom{4}{1}\Big/\binom{52}{5},$$

and we are done.

7 Discrete random variables (29/09)

Definition 7.1 (Random variable)**.** *Suppose we are operating in a sample space Ω. Then a random variable is a function from Ω to $\mathbb{R}$, $\mathcal{X} : \Omega \to \mathbb{R}$ that maps event outcomes in the sample space to the real number space.*

Example 7.1. Flip 2 coins. Then $\Omega = \{HH, HT, TH, TT\}$. Let $\mathcal{X}_1$ be the number of heads obtained, conceptually. This function maps the event outcomes in Ω to the real numbers:

$$\mathcal{X}_1(HH) = 2, \qquad \mathcal{X}_1(HT) = 1, \qquad \mathcal{X}_1(TH) = 1, \qquad \mathcal{X}_1(TT) = 0.$$

Another random variable on Ω is $\mathcal{X}_2$, which could represent the number of tails obtained. Mathematically, then:

$$\mathcal{X}_2(HH) = 0, \qquad \mathcal{X}_2(HT) = 1, \qquad \mathcal{X}_2(TH) = 1, \qquad \mathcal{X}_2(TT) = 2.$$

Example 7.2. Consider the event that $\mathcal{X}_1 = 1$, or $\{\mathcal{X}_1 = 1\}$, or $\{\omega \in \Omega \mid \mathcal{X}_1(\omega) = 1\}$, or $\{HT, TH\}$, or the set of outcomes for which $\mathcal{X}_1$ maps to 1—all of these are equivalent.

Probabilities of events of this form are typically written as $\mathbb{P}(\mathcal{X} = 1)$ rather than $\mathbb{P}(\{\mathcal{X} = 1\})$.

Example 7.3. Roll two dice. Then the sample space is $\Omega = \{1, 2, 3, 4, 5, 6\}^2$. Obviously there are 36 outcomes, each equally likely with probability $1/36$. Let $\mathcal{X}$ be the random variable pointing to the value of the first of both rolls and $\mathcal{Y}$ be the random variable pointing to the maximum of the two rolls. Then for $(i, j) \in \Omega$,

$$\mathcal{X}(i, j) = i \qquad \text{and} \qquad \mathcal{Y}(i, j) = \max(i, j).$$

Example 7.4. Consider the event $\{\mathcal{X} = 1 \text{ or } \mathcal{Y} = 1\}$ from the example above. Convince yourself that $\mathbb{P}(\{\mathcal{X} = 1\} \cup \{\mathcal{Y} = 1\}) = 1/6$.

Definition 7.2 (Discrete R.V.)**.** *A discrete random variable is one taking values in a finite (such as $\mathcal{X} : \Omega \to \{1, 2\}$) or countably infinite (such as $\mathcal{X} : \Omega \to \mathbb{N}$ or $\mathcal{X} : \Omega \to \mathbb{Z}$) set.*

Definition 7.3 (Probability mass function, PMF)**.** *The probability mass function of a random variable $\mathcal{X}$ is the function $p : \mathbb{R} \to \mathbb{R}$ defined by $p_{\mathcal{X}}(k) = \mathbb{P}(\mathcal{X} = k)$ for $k \in \mathbb{R}$.*

Remark. The PMF summarizes all of the probabilities related to $\mathcal{X}$.
Remark. The PMF is a *distribution*. More on this later.

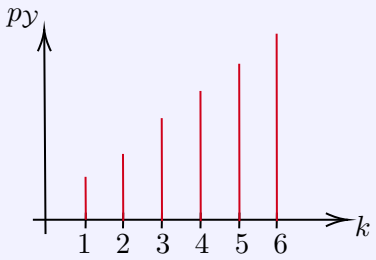
Example 7.5. Recall the random variable $\mathcal{Y}$ from Example 7.3. To find the PMF of $\mathcal{Y}$, we need to compute $\mathbb{P}(\mathcal{Y} = K)$ for any number k. Here, the possible values are $k \in \{1,2,3,4,5,6\}$, thus $p_{\mathcal{Y}}(k) = \mathbb{P}(\mathcal{Y} = k) = 0$ for $k \notin \{1,2,3,4,5,6\}$, i.e. for all other k values. It is not hard to check that

$$p_{\mathcal{Y}}(1) = \frac{1}{36}, \qquad p_{\mathcal{Y}}(2) = \frac{3}{36}, \qquad p_{\mathcal{Y}}(3) = \frac{5}{36}, \qquad p_{\mathcal{Y}}(4) = \frac{7}{36}, \qquad p_{\mathcal{Y}}(5) = \frac{9}{36}, \qquad p_{\mathcal{Y}}(6) = \frac{11}{36}.$$

If we wanted to summarize this PMF, we could write

$$p_{\mathcal{Y}}(k) = \begin{cases} \dfrac{2k-1}{36} & \text{if } k \in \{1,2,3,4,5,6\} \\ 0 & \text{otherwise.} \end{cases}$$

The plot for this PMF is discretized:



Remark. To calculate *any* probability involving an R.V., say $\mathcal{Y}$, we *only* need to know its PMF. More generally, we can compute a probability involving $\mathcal{Y}$ by summing up the values of PMF $p_{\mathcal{Y}}(k)$ over the relevant values of k.

We often define an R.V. in terms of its PMF, without necessarily mentioning the sample space.

Example 7.6. Let $\mathcal{X}$ be a random variable with PMF given by

$$p_{\mathcal{X}} = \begin{cases} 1/5 & \text{if } k \in \{-2,1,0,1,2\} \\ 0 & \text{otherwise.} \end{cases}$$

We can find a few probabilities involving $\mathcal{X}$:

- $\mathbb{P}(\mathcal{X} = 1) = p_{\mathcal{X}}(1) = 1/5$.
- Similarly,
$$\mathbb{P}(\mathcal{X}^2 = 1) = \mathbb{P}(\mathcal{X} = 1 \text{ or } \mathcal{X} = -1) = p_{\mathcal{X}}(1) + p_{\mathcal{X}}(-1) = 1/5 + 1/5 = 2/5,$$
- and
$$\mathbb{P}(|\mathcal{X} - 2| = 1) = \mathbb{P}(\mathcal{X} = 1 \text{ or } \mathcal{X} = 3) = p_{\mathcal{X}}(1) + p_{\mathcal{X}}(3) = 1/5 + 0 = 1/5.$$

Remark. In general, a valid PMF is any function $p : \mathbb{R} \to \mathbb{R}$ such that $p(k) \geq 0$ for every k and all the output values sum up to 1 (and hence form what is called a *probability simplex*).

8 Common random variables (04/10)

Recap. A *discrete random variable* $\mathcal{X}$ is

- **Formally**, a function $\mathcal{X} : \Omega \to \mathbb{R}$ on the sample space (with finite/countably infinite values).
- **Practically**, specified in terms of its probability mass function (PMF),

$$p_{\mathcal{X}}(k) = \mathbb{P}(\mathcal{X} = k) \qquad \text{for } k \in \mathbb{R}.$$

  A valid PMF is nonnegative and has values that sum to 1.
- We can calculate any probabilities involving $\mathcal{X}$ using its PMF. For example, $\mathbb{P}(-3 \leq \mathcal{X} \leq 3)$ is equal to the sum of $\mathbb{P}(\mathcal{X} = k)$ over those values of k with $-3 \leq k \leq 3$ and $\mathbb{P}(X = k) \neq 0$.

8.1 Common families of RVs specified in terms of their PMF

1. **The Bernoulli Random Variable.** The Bernoulli random variable $\mathcal{X}$ takes values 0 or 1, and has a single parameter p called the "success probability". We write $\mathcal{X} \sim \mathsf{Bernoulli}(p)$ or $\mathcal{X} \sim \mathsf{Bern}(p)$ to mean $\mathcal{X}$ has the Bernoulli distribution with parameter p. Defined in terms of its PMF, the Bernoulli distribution is

$$p_k(\mathcal{X}) = \begin{cases} p & \text{if } k = 1 \\ 1 - p & \text{if } k = 0 \\ 0 & \text{otherwise.} \end{cases}$$

   Example 8.1. Toss a coin. Let $\mathcal{X} = 1$ if I toss heads, and $\mathcal{X} = 0$ if I toss tails. Then $\mathcal{X} \sim \mathsf{Bern}(1/2)$. The parameter $1/2$ is derived from the success probability $p = \mathbb{P}(\text{heads}) = \mathbb{P}(\mathcal{X} = 1) = 1/2$.

   Example 8.2. Roll a die. Let $\mathcal{X} = 1$ if I roll a 6, and $\mathcal{X} = 0$ otherwise. Then $\mathcal{X} \sim \mathsf{Bern}(1/6)$. The parameter $1/6$ is derived from the success probability $p = \mathbb{P}(\text{roll a } 6) = \mathbb{P}(\mathcal{X} = 1) = 1/6$.

   Although this is a simple r.v., the Bernoulli r.v. is very important. We typically use it in practice to model generic probabilistic situations with just two outcomes, such as:

   (a) The state of a telephone at a given time that can be either free or busy.

   (b) A person who can be either healthy or sick with a certain disease.

   (c) The preference of a person who can be either for or against a certain political candidate.

   By combining multiple Bernoulli random variables, we can construct more complicated random variables.

2. **The Geometric Random Variable.** The geometric random variable $\mathcal{X}$ takes values k in $\mathbb{N}$, and has a single parameter p once again called the "success probability". We write $\mathcal{X} \sim \mathsf{Geometric}(p)$ or $\mathcal{X} \sim \mathsf{Geom}(p)$ to mean $\mathcal{X}$ has the Bernoulli distribution with parameter p. We think of the geometric r.v. as modelling the number of independent trials until the

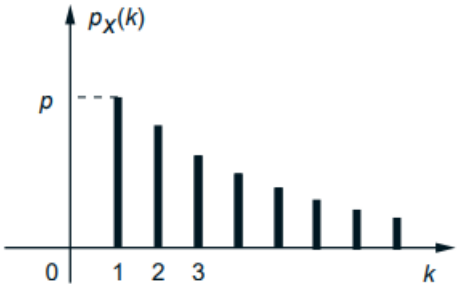
first "success" occurs (inclusive). To this end, we can say that the PMF of a geometric r.v. is the product of the probability that the kth trial succeeds and all the $k-1$ trials before it are all failures, i.e.

$$p_{\mathcal{X}}(k) = \mathbb{P}(\mathcal{X} = k) = p(1-p)^{k-1}.$$

We can check that this PMF is valid by noting that

$$\sum_{k=1}^{\infty} p_{\mathcal{X}}(k) = \sum_{k=1}^{\infty} p(1-p)^{k-1} = p\sum_{k=1}^{\infty}(1-p)^{k-1} = p\sum_{j=0}^{\infty}(1-p)^{j} = p \cdot \frac{1}{1-(1-p)} = 1,$$

where we set $j = k-1$ somewhere in there. This PMF decreases geometrically with common ratio $1-p$:



Example 8.3. If $p = 1/2$, as in the instance where $\mathcal{X}$ models the number of coin flips before the first head appears, then

$$\mathbb{P}(\mathcal{X} = k) = \frac{1}{2}\left(1 - \frac{1}{2}\right)^{k-1} = 2^{-k}.$$

Example 8.4. If $p = 1/6$, as in the instance where $\mathcal{X}$ models the number of die rolls before the first head appears, then

$$\mathbb{P}(\mathcal{X} = k) = \frac{1}{6}\left(1 - \frac{1}{6}\right)^{k-1} = \frac{1}{6}\left(\frac{5}{6}\right)^{k-1}.$$

3. **The Binomial Random Variable.** The binomial random variable $\mathcal{X}$ has parameters p and n, where $0 \leq p \leq 1$ is again the "success probability" and n a non-negative integer is the number of trials. $\mathcal{X}$ takes values k in $\{0, 1, 2, \ldots, n\}$. We write $\mathcal{X} \sim \mathsf{Binomial}(n, p)$ or simply $\mathcal{X} \sim \mathsf{Binom}(n, p)$ to mean that $\mathcal{X}$ has the binomial distribution with parameters n and p.

   The binomial distribution models the number of successes out of n independent trials each of which has success probability p. Let us then find the PMF of the binomial r.v. The probability of k successful trials out of the n trials is p^k, and the probability of the other $n-k$ failed trials out of n is $(1-p)^{n-k}$. Additionally, the number of ways of selecting these k successful trials out of all the n is $\binom{n}{k}$. The PMF is the product of these three quantities:

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

Example 8.5. The number of heads I get out of 100 coin flips is modelled by $X \sim \mathsf{Binomial}(100, 1/2)$.

Example 8.6. We have that

$$\mathbb{P}(\mathcal{X} = 0) = \mathbb{P}(\text{all } n \text{ trials are failures}) = \binom{n}{0}(1-p)^k(1-p)^{n-k} = (1-p)^n,$$

$$\mathbb{P}(\mathcal{X} = 1) = \mathbb{P}(\text{exactly one of the } n \text{ trials succeeds}) = \binom{n}{1}p(1-p)^{n-1} = np(1-p)^{n-1}.$$

We can check that our PMF above is valid. Thanks to the binomial theorem $(a+b)^n = \sum_{k=0}^{n}\binom{n}{k}a^k b^{n-k}$, we have

$$\sum_{k=0}^{n} p_{\mathcal{X}}(k) = \sum_{k=0}^{n}\binom{n}{k}p^k(1-p)^{n-k} = (p+[1-p])^n = 1^n = 1.$$

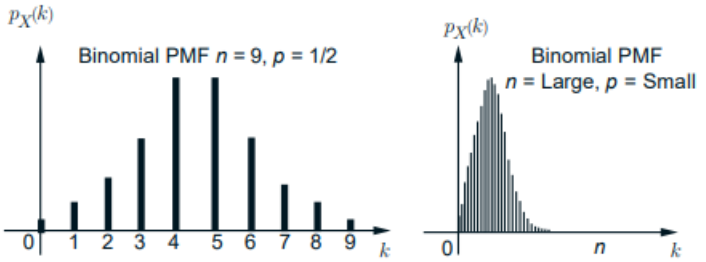


Figure 2.3: The PMF of a binomial random variable. If $p = 1/2$, the PMF is symmetric around $n/2$. Otherwise, the PMF is skewed towards 0 if $p < 1/2$, and towards n if $p > 1/2$.

8.2 Example: Decisive Vote

Does my vote really count? By how much and with what probability?

Let n be a positive integer and $0 < p < 1$. Suppose there are $2n$ people in my state, so that there are $2n+1$ people in total (including me). The voting rules are constructed such that there is a two-party system of Party A and Party B. Each person votes independently, and vote for A with probability p and B with probability $1-p$.

Let the random variable $\mathcal{X}$ represent the number of votes for A out of the other $2n$ people in the state, excluding myself. It is clear that $\mathcal{X} \sim \mathsf{Binomial}(2n, p)$.

My vote is *decisive* if $\mathcal{X} = n$—that is, the number of votes of the other people are split practically evenly—which happens with probability

$$\mathbb{P}(\mathcal{X} = n) = \binom{2n}{n}p^n(1-p)^{2n-n} = \binom{2n}{n}p^n(1-p)^n.$$

Case 1: Set $p = 1/2$ to model a *polarized* community. Then

$$\mathbb{P}(\mathcal{X} = n) = \binom{2n}{n}\left(\frac{1}{2}\right)^{2n} = \frac{(2n)!}{n!(2n-n)!}2^{-2n} = 2^{-2n}\frac{(2n)!}{(n!)^2}.$$

Recall Stirling's formula:

$$m! \approx \sqrt{2\pi m}\left(\frac{m}{e}\right)^m \qquad \text{for sufficiently large } m.$$

Use this formula with both $m = n$ and $m = 2n$ to get

$$\mathbb{P}(\mathcal{X} = n) = 2^{-2n} \cdot \frac{\sqrt{2\pi \cdot 2n}\left(\frac{2n}{e}\right)^{2n}}{\left(\sqrt{2\pi \cdot n}\left(\frac{n}{e}\right)^n\right)^2} = \frac{1}{\sqrt{\pi n}}.$$

This expression decays very slowly in n: if $2n = 100,000$, then $\mathbb{P}(\mathcal{X} = n) \approx 1/400$; if $2n = 1,000,000$, then $\mathbb{P}(\mathcal{X} = n) \approx 1/1250$.

Case 2: If $p \neq 1/2$, then $\mathbb{P}(\mathcal{X} = n)$ decays exponentially—much more quickly than $1/\sqrt{\pi n}$—as $n \to \infty$. Exercise: Convince yourself of this.

9 Transformation; Poisson (11/10)

9.1 Transformations—Functions of a Random Variable

If $\mathcal{Y} = f(\mathcal{X})$ is a function of a random variable $\mathcal{X}$, then $\mathcal{Y}$ is also a random variable, since it provides a numerical value for each possible outcome. This is because every outcome in the sample space defines a numerical value k for $\mathcal{X}$ and hence also the numerical value $n = f(k)$ for $\mathcal{Y}$. If $\mathcal{X}$ is discrete with PMF $p_{\mathcal{X}}$, then $\mathcal{Y}$ is also discrete, and its PMF $p_{\mathcal{Y}}$ can be calculated using the PMF of $\mathcal{X}$. In particular, to obtain $p_{\mathcal{Y}}(y)$ for any y, we add the probabilities of all values of k such that $f(k) = n$:

$$\begin{aligned} p_{\mathcal{Y}}(n) &= \mathbb{P}(\mathcal{Y} = n) \\ &= \mathbb{P}(f(k) = n) \\ &= \sum_{\{k \mid f(k) = n\}} p_{\mathcal{X}}(k) \\ &= \text{sum of } \mathbb{P}(\mathcal{X} = k) \text{ over } k \text{ such that } f(k) = n \end{aligned}$$

Example 9.1. Pick a random number from $-2, -1, 0, 1, 2$. Call it $\mathcal{X}$. The PMF of $\mathcal{X}$ is

$$p_{\mathcal{X}}(k) = \begin{cases} 1/5 & \text{if } k = -2, -1, 0, 1, 2 \\ 0 & \text{otherwise.} \end{cases}$$

Now define a new random variable $\mathcal{Y} = \mathcal{X}^2$. To find the PMF of $\mathcal{Y}$, we note that all its possible values are 0, 1, and 4, so that

$$\mathbb{P}(\mathcal{Y} = 0) = \mathbb{P}(\mathcal{X}^2 = 0) = \mathbb{P}(\mathcal{X} = 0) = 1/5,$$

$$\begin{aligned}
\mathbb{P}(\mathcal{Y}=1) &= \mathbb{P}(\mathcal{X}^2=1) \\
&= \mathbb{P}(\mathcal{X}=1 \cup \mathcal{X}=-1) \\
&= \mathbb{P}(\mathcal{X}=1) + \mathbb{P}(\mathcal{X}=-1) \\
&= 1/5 + 1/5 \\
&= 2/5, \text{ and,} \\
\mathbb{P}(\mathcal{Y}=4) &= \mathbb{P}(\mathcal{X}^2=4) \\
&= \mathbb{P}(\mathcal{X}=2 \cup \mathcal{X}=-2) \\
&= \mathbb{P}(\mathcal{X}=2) + \mathbb{P}(\mathcal{X}=-2) \\
&= 1/5 + 1/5 \\
&= 2/5.
\end{aligned}$$

Summarising, the PMF is

$$p_{\mathcal{Y}}(n) = \begin{cases} 1/5 & \text{if } n = 0 \\ 2/5 & \text{if } n = 1, 4 \\ 0 & \text{otherwise.} \end{cases}$$

9.2 The Poisson distribution

A Poisson random variable $\mathcal{X} \sim \mathsf{Poisson}(\lambda)$ takes non-negative integer values. Its PMF is given by

$$p_{\mathcal{X}}(k) = \mathbb{P}(\mathcal{X}=k) = e^{-\lambda}\frac{\lambda^k}{k!} \qquad \text{for } k = 0, 1, 2, \ldots$$

where λ is a positive rate parameter characterising the PMF. This PMF is valid, since

$$\sum_{k=0}^{\infty} e^{-\lambda}\frac{\lambda^k}{k!} = e^{-\lambda}\sum_{k=0}^{\infty}\frac{\lambda^k}{k!} = e^{-\lambda}\left(1 + \lambda + \frac{\lambda^2}{2!} + \frac{\lambda^3}{3!} + \ldots\right) = e^{-\lambda}e^{\lambda} = 1.$$

The Poisson distribution typically models:

(a) the number of "arrivals" (at average rate λ per unit time) within a given unit of time. For example, the number of trains that arrive in one hour

(b) "rare" events. For example, deaths by horse kick

(c) the number of typos in a book with a total of n words, when the probability p that any one word is misspelled is very small (associate a word with a coin toss which comes a head when the word is misspelled)

(d) the number of cars involved in accidents in a city on a given day (associate a car with a coin toss which comes a head when the car has an accident), etc.

Remark. We can think of the Poisson random variable as a model of a binomial random variable with very large n and very small p. More precisely, if $\mathcal{X}_n \sim \mathsf{Binomial}\left(n, \frac{\lambda}{n}\right)$ for each integer $n \geq \lambda$,

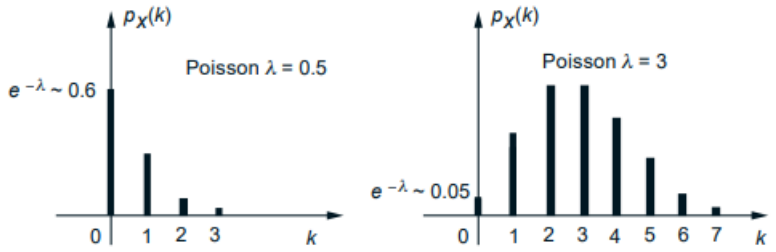


Figure 2.5: The PMF $e^{-\lambda}\frac{\lambda^k}{k!}$ of the Poisson random variable for different values of λ. Note that if $\lambda < 1$, then the PMF is monotonically decreasing, while if $\lambda > 1$, the PMF first increases and then decreases as the value of k increases (this is shown in the end-of-chapter problems).

then for any k, we can show that (doing this is left as an exercise):

$$\lim_{n\to\infty} \mathbb{P}(\mathcal{X}_n = k) = e^{-\lambda} \cdot \frac{\lambda^k}{k!}.$$

A common-sense interpretation of this is that the Poisson random variable is the limiting case of the binomial random variable with many trials, each of which is unlikely/rare.

10 Expectation and Variance (13/10)

Definition 10.1 (Expected value)**.** *Let $\mathcal{X}$ be a discrete r.v. Then the expectation of $\mathcal{X}$ is the sum of the product of every feasible value k by its corresponding probability $\mathbb{P}(\mathcal{X} = k)$ over all the k, that is,*

$$\mathbb{E}[X] = \sum_k k \cdot \mathbb{P}(\mathcal{X} = k).$$

Example 10.1. Let $\mathcal{X}$ be the outcome of the roll of a fair six-sided die. Then

$$\mathbb{E}[\mathcal{X}] = \sum_k k \cdot \mathbb{P}(\mathcal{X} = k) = \sum_{k=1}^{6} k \cdot \frac{1}{6} = 3.5$$

Note that $\mathbb{E}[\mathcal{X}]$ does not need to be one of the values $\mathcal{X}$ can take.

Example 10.2 (Mega-millions lottery ticket)**.** Suppose that everyone in the US entered for a fair lottery ticket with a payout of \$1 billion to one person. Then the probability p that I win is $\frac{1}{302,000,000}$. Let $\mathcal{X}$ represent the amount that I win, so that $\mathbb{P}(\mathcal{X} = \$1 \text{ billion}) =$

$\frac{1}{302,000,000} = p$, and $\mathbb{P}(\mathcal{X} = \$0) = 1 - p$. Then,

$$\begin{aligned}\mathbb{E}[\mathcal{X}] &= 0 \cdot \mathbb{P}(\mathcal{X} = 0) + \$1 \text{ billion} \cdot \frac{1}{302,000,000} \\ &= \frac{\$10^9}{3.02 \cdot 10^8} \\ &\approx \$3.\end{aligned}$$

So we only have a good deal if we enter an amount $< \$3$!

10.1 Expectation of common random variables

1. $\mathcal{X} \sim \mathsf{Bernoulli}(p)$: The expectation is $\mathbb{E}[\mathcal{X}] = 0 \cdot \mathbb{P}(\mathcal{X} = 0) + 1 \cdot \mathbb{P}(\mathcal{X} = 1) = p$.
2. $\mathcal{X} \sim \mathsf{Geometric}(p)$: Recall that the PMF is $\mathbb{P}(\mathcal{X} = k) = p(1-p)^{k-1}$ for $k = 1, 2, 3, \ldots$. Then

   $$\mathbb{E}[\mathcal{X}] = \sum_{k=1}^{\infty} kp(1-p)^{k-1} = p\sum_{k=1}^{\infty} k(1-p)^{k-1} = p\sum_{k=1}^{\infty} -\frac{d}{dp}(1-p)^k = -p \cdot \frac{d}{dp}\left(\sum_{k=1}^{\infty}(1-p)^k\right) = \frac{1}{p}.$$

   Note that the expectation decreases with increasing p—which makes sense intuitively.
3. $\mathcal{X} \sim \mathsf{Binomial}(n, p)$: Recall that the PMF is $\mathbb{P}(\mathcal{X} = k) = \binom{n}{k}p^k(1-p)^{n-k}$ for $k = 0, 1, \ldots, n$. As a preparatory step, we compute, for $k \geq 1$:

   $$k\binom{n}{k} = k \cdot \frac{n!}{k!(n-k)!} = \frac{n(n-1)!}{(k-1)!(n-k)!} \cdot n\frac{(n-1)!}{(k-1)!(n-1-(k-1))!} = n\binom{n-1}{k-1}.$$

   Then, with $q = 1 - p$, we have

   $$\mathbb{E}[\mathcal{X}] = \sum_{k=0}^{n} k\binom{n}{k}p^k q^{n-k} = \sum_{k=1}^{n} k\binom{n}{k}p^k q^{n-k} = \sum_{k=0}^{n} n\binom{n-1}{k-1}pp^{k-1}q^{n-1-(k-1)},$$

   so that

   $$\mathbb{E}[\mathcal{X}] = \sum_{k=0}^{n} n\binom{n-1}{k-1}pp^{k-1}q^{n-1-(k-1)} = np\sum_{k=0}^{n-1} n\binom{n-1}{k}p^k q^{n-k} = np.$$

4. $\mathcal{X} \sim \mathsf{Poisson}(\lambda)$: Recall that the PMF is $\mathbb{P}(\mathcal{X} = k) = e^{-\lambda}\frac{\lambda^k}{k!}$ for $k \geq 0$. Now,

   $$\mathbb{E}[\mathcal{X}] = \sum_{k=0}^{\infty} ke^{-\lambda} \cdot \frac{\lambda^k}{k!} = \lambda\sum_{k=1}^{\infty} e^{-\lambda} \cdot \frac{\lambda^{k-1}}{(k-1)!} = \lambda\sum_{k=0}^{\infty} ke^{-\lambda}\frac{\lambda^k}{k!} = \lambda.$$

   This computation helps justify calling λ the expected arrival rate. Note that the expectation increases with λ—it is λ after all. Additionally, $\mathrm{Var}(\mathcal{X}) = \lambda$.

Remark. *Expectation* is distinct from *mode*, which is defined as the value of k for which $\mathbb{P}(\mathcal{X} = k)$ is largest, i.e. the most likely outcome.

10.2 Transformed expectations

Given a random variable $\mathcal{X}$ together with its PMF and a function $f : \mathbb{R} \to \mathbb{R}$ generating a random variable $\mathcal{Y}$, we can compute $\mathbb{E}[\mathcal{Y}]$ in two ways:

(a) Find the PMF $\mathbb{P}(\mathcal{Y} = k)$ for all $k \in \mathbb{R}$, and then use the definition $\mathbb{E}[\mathcal{Y}] = \sum_k k \cdot \mathbb{P}(\mathcal{Y} = k)$, or

(b) (Law of the unconscious statistician.) Use the direct formula $\mathbb{E}[f(\mathcal{X})] = \sum_k f(k) \cdot \mathbb{P}(\mathcal{X} = k)$.

Example 10.3. Recall the example from the previous lecture, where we found that the PMF is

$$p_{\mathcal{Y}}(n) = \begin{cases} 1/5 & \text{if } n = 0 \\ 2/5 & \text{if } n = 1, 4 \\ 0 & \text{otherwise.} \end{cases}$$

To compute the PMF, we could

- use the transformed PMF to get $\mathbb{E}[\mathcal{Y}] = 0 \cdot \mathbb{P}(\mathcal{Y} = 0) + 1 \cdot \mathbb{P}(\mathcal{Y} = 1) + 4 \cdot \mathbb{P}(\mathcal{Y} = 4)$, which works out to $1 \cdot 2/5 + 4 \cdot 2/5 = 2$.
- use the direct formula to get

$$\mathbb{E}[\mathcal{Y}] = \mathbb{E}[\mathcal{X}^2] = \sum_{k=-2}^{2} k^2 \mathbb{P}(\mathcal{X} = k) = \sum_{k=-2}^{2} k^2 \cdot \frac{1}{5} = 2.$$

10.3 Variance

The *variance* of a distribution is a measure of how "spread out" the distribution is around the mean.

Definition 10.2 (Variance)**.** *The variance of a random variable $\mathcal{X}$ is defined by the expected/mean squared error* $\operatorname{Var}(\mathcal{X}) = \mathbb{E}[(\mathcal{X} - \mathbb{E}[\mathcal{X}])^2]$ *and can be calculated as*

$$\operatorname{Var}(\mathcal{X}) = \sum_k (\mathcal{X} - \mathbb{E}[\mathcal{X}])^2 p_{\mathcal{X}}(k).$$

Example 10.4. Let $\mathcal{X} \sim \mathsf{Bernoulli}(p)$, and recall that $\mathbb{E}[\mathcal{X}] = p$. Then,

$$\begin{aligned} \operatorname{Var}(\mathcal{X}) &= \mathbb{E}[(\mathcal{X} - \mathbb{E}[\mathcal{X}])^2] \\ &= \mathbb{E}[(\mathcal{X} - p)^2] \\ &= (0 - p)^2 \mathbb{P}(\mathcal{X} = 0) + (1 - p)^2 \mathbb{P}(\mathcal{X} = 1) \\ &= p^2(1 - p) + (1 - p)^2 p \\ &= p(1 - p). \end{aligned}$$

10.4 Properties of expectation and variance

We now begin to itemize a few properties of expectation and variance as we go along, including proofs where necessary.

- In general, $\text{Var}(\mathcal{X}) \geq 0$. Additionally, $\text{Var}(\mathcal{X}) > 0$ except when $\mathcal{X}$ is *degenerate*, meaning all of its probability is localised on one point (i.e. $\mathbb{P}(\mathcal{X} = c) = 1$ for some $c \in \mathbb{R}$).
- The square in the definition "changes the units" of the r.v. To get back to original units, we define the standard deviation $\sigma_{\mathcal{X}}$ to be the standard deviation of the random variable $\mathcal{X}$.
- *Warning:* In general, $\mathbb{E}[f(\mathcal{X})] \neq f(\mathbb{E}[\mathcal{X}])$—you cannot switch the order of expectation and (nonlinear) functions! For example, $\mathbb{E}[\mathcal{X}^2] \neq (\mathbb{E}[\mathcal{X}])^2$.
- Expectation preserves constants, i.e. $\mathbb{E}[a] = a$ for constant a.
- Variance destroys constants, i.e. $\text{Var}(a) = 0$ for constant a.
- (Linearity of expectations.) For constants a and b and r.v. $\mathcal{X}$, $\mathbb{E}[a\mathcal{X} + b] = a\mathbb{E}[\mathcal{X}] + b$.

  Proof. Recall that $\mathbb{E}[\mathcal{X}] = \sum_k k\mathbb{P}(\mathcal{X} = k)$. Then,

  $$\begin{aligned}\mathbb{E}[a\mathcal{X} + b] &= \sum_k (ak + b)\mathbb{P}(\mathcal{X} = k) \\ &= \sum_k ak\mathbb{P}(\mathcal{X} = k) + \sum_k b\mathbb{P}(\mathcal{X} = k) \\ &= a\sum_k k\mathbb{P}(\mathcal{X} = k) + b\sum_k \mathbb{P}(\mathcal{X} = k) \\ &= a\mathbb{E}[\mathcal{X}] + b\end{aligned}$$

  Notable special case: Set $b = 0$. Then $\mathbb{E}[a\mathcal{X}] = a\mathbb{E}[\mathcal{X}]$. ■
- (Variance scaling and shifting.) For constants a and b, we have that $\text{Var}(a\mathcal{X}+b) = a^2\text{Var}(\mathcal{X})$.

  Proof. By definition,

  $$\begin{aligned}\text{Var}(a\mathcal{X} + b) &= \mathbb{E}[(a\mathcal{X} + b - \mathbb{E}[a\mathcal{X} + b])^2] \\ &= \mathbb{E}[(a\mathcal{X} + b - a\mathbb{E}[\mathcal{X}] - b)^2] \\ &= \mathbb{E}[a^2(\mathcal{X} - \mathbb{E}[\mathcal{X}])^2] \\ &= a^2\mathbb{E}[(\mathcal{X} - \mathbb{E}[\mathcal{X}])^2] \\ &= a^2\text{Var}(\mathcal{X})\end{aligned}$$

  Notable special case: Set $a = -1, b = 0$. Then $\text{Var}(-\mathcal{X}) = \text{Var}(\mathcal{X})$. ■
- (Alternative formula.) The variance of a random variable $\mathcal{X}$ is also given by $\text{Var}(\mathcal{X}) = \mathbb{E}[\mathcal{X}^2] - (\mathbb{E}[X])^2$.

Proof. Let $m = \mathbb{E}[\mathcal{X}]$. Then,

$$\begin{aligned}\mathrm{Var}(\mathcal{X}) &= \mathbb{E}[(\mathcal{X}-m)^2] \\ &= \mathbb{E}[\mathcal{X}^2 - 2m\mathcal{X} + m^2] \\ &= \mathbb{E}[\mathcal{X}^2] - \mathbb{E}[2m\mathcal{X}] + \mathbb{E}[m^2] \\ &= \mathbb{E}[\mathcal{X}^2] - 2m\mathbb{E}[\mathcal{X}] + m^2 \\ &= \mathbb{E}[\mathcal{X}^2] - 2m^2 + m^2 \\ &= \mathbb{E}[\mathcal{X}^2] - m^2 \\ &= \mathbb{E}[\mathcal{X}^2] - (\mathbb{E}[X])^2\end{aligned}$$

■

- We have that $\mathrm{Var}(\mathcal{X}) = \mathbb{E}[\mathcal{X}^2] - (\mathbb{E}[X])^2 \geq 0$. Rearrange to get $\mathbb{E}[\mathcal{X}^2] \geq (\mathbb{E}[X])^2$; these are only equal when $\mathcal{X}$ is nonrandom.
  A more general fact: If $f : \mathbb{R} \to \mathbb{R}$ is convex (i.e. $f'' \geq 0$), then $\mathbb{E}[f(\mathcal{X})] \geq f(\mathbb{E}[\mathcal{X}])$. This is an important inequality—it is called **Jensen's inequality**.

11 Joint Distributions (10/18)

Suppose $\mathcal{X} : \Omega \to \mathbb{R}$ and $\mathcal{Y} : \Omega \to \mathbb{R}$ are discrete random variables defined on a sample space Ω.

Definition 11.1 (Joint PMF)**.** *The joint probability mass function of $\mathcal{X}$ and $\mathcal{Y}$ is the function $p_{\mathcal{X},\mathcal{Y}} : \mathbb{R}^2 \to \mathbb{R}$ defined by*

$$p_{\mathcal{X},\mathcal{Y}}(x,y) = \mathbb{P}(\mathcal{X} = x, \mathcal{Y} = y)$$

for $x, y \in \mathbb{R}$.

Example 11.1.

$$p_{\mathcal{X},\mathcal{Y}} = \mathbb{P}(\mathcal{X} = x, \mathcal{Y} = y) = \mathbb{P}(\{\mathcal{X} = 3\} \cap \{\mathcal{Y} = 8\})$$

is an example of a joint PMF. We often illustrate a joint PMF using a table.

Example 11.2. Roll two 3-sided dice. Then as usual, $\Omega = \{1,2,3\}^2$. Each outcome is equally likely with probability $1/9$.
Let $\mathcal{X}$ be the sum of the two rolls and $\mathcal{Y}$ be the maximum of the two rolls. Thus $\mathcal{X}(i,j) = i + j$ and $\mathcal{Y} = \max(i,j)$. We desire the joint PMF of these two random variables, namely $\mathbb{P}(\mathcal{X} = x, \mathcal{Y} = y)$ for every $x, y \in \mathbb{R}$. Note that $\mathcal{X}$ takes values in $\{2,3,4,5,6\}$ and $\mathcal{Y}$ takes values in $\{1,2,3\}$.

		$\mathcal{X}$				
$\mathbb{P}(x,y)$		2	3	4	5	6
$\mathcal{Y}$	1	1/9	0	0	0	0
	2	0	2/9	1/9	0	0
	3	0	0	2/9	2/9	1/9

Convince yourself that the above table of the joint PMF is accurate. As a sanity check, notice that the sum of all the numbers in all entries of the table is 1. The joint PMF allows us to calculate *any* probability involving $\mathcal{X}$ and $\mathcal{Y}$.

Example 11.3. The probability that $\mathcal{X} = \mathcal{Y} + 1$ is given by

$$\begin{aligned}\mathbb{P}(\mathcal{X} = \mathcal{Y}+1) &= \text{sum of } p_{\mathcal{X},\mathcal{Y}}(x,y) \text{ over all } (x,y)\text{-pairs satisfying } \mathcal{X}-\mathcal{Y}=1 \\ &= \mathbb{P}(\mathcal{X}=2,\mathcal{Y}=1) + \mathbb{P}(\mathcal{X}=3,\mathcal{Y}=2) + \mathbb{P}(\mathcal{X}=4,\mathcal{Y}=3) \\ &= 1/9 + 2/9 + 2/9 \\ &= 5/9.\end{aligned}$$

Definition 11.2 (Marginal PMF)**.** *The marginal PMF of $\mathcal{X}$ is just the PMF of $\mathcal{X}$ itself, i.e. the function $p_{\mathcal{X}} : \mathbb{R} \to \mathbb{R}$ given by $p_{\mathcal{X}}(x) = \mathbb{P}(\mathcal{X} = x)$. Likewise, the marginal of $\mathcal{Y}$ is $p_{\mathcal{Y}}(y) = \mathbb{P}(\mathcal{Y} = y)$.*

We can compute the marginals from the joint PMF viz:

$$p_{\mathcal{X}}(x) = \sum_y p_{\mathcal{X},\mathcal{Y}}(x,y), \qquad \text{or equivalently,} \qquad \mathbb{P}(\mathcal{X}=x) = \sum_y \mathbb{P}(\mathcal{X}=x,\mathcal{Y}=y) \text{ for each } x \in \mathbb{R}.$$

This is essentially summing along the columns of the joint PMF table to find $p_{\mathcal{X}}(x)$, or summing along the rows to find $p_{\mathcal{Y}}(y)$.

Motivation for marginals. Suppose $y_1, \ldots, y_n$ are the possible $\mathcal{Y}$ values. Then $B_i = \{Y = y_i\}$ for $i = 1, \ldots, n$ form a partition of Ω. As a result,

$$\begin{aligned}\mathbb{P}(\{\mathcal{X}=x\}) &= \mathbb{P}\left(\bigcup_{i=1}^n (\{\mathcal{X}=x\} \cap B_i)\right) \\ &= \sum_{i=1}^n \mathbb{P}(\{\mathcal{X}=x\} \cap B_i) \\ &= \sum_{i=1}^n \mathbb{P}(\mathcal{X}=x, \mathcal{Y}=y_i)\end{aligned}$$

Remark 1. Joint PMFs "contain more information" then the two marginals. We can derive marginals from the joint PMF, but cannot derive the joint PMF from the marginals.

Remark 2. There can be $(\mathcal{X},\mathcal{Y})$ and $(\hat{\mathcal{X}},\hat{\mathcal{Y}})$ such that the joint PMFs, i.e. $p_{\mathcal{X},\mathcal{Y}}$ and $p_{\hat{\mathcal{X}},\hat{\mathcal{Y}}}$, are different but the marginals are the same, i.e. $p_{\mathcal{X}} = p_{\hat{\mathcal{X}}}$ and $p_{\mathcal{Y}} = p_{\hat{\mathcal{Y}}}$. Exercise: can you come up with an example?

12 Expectations, linearity, and conditioning (20/10)

Example 12.1. Run a food truck. Every customer chooses a burrito, a taco, meat, or vegetables on each. We charge \$1 per taco, \$2 per burrito, \$2 per meat, and \$1 per vegetable. Furthermore, let $\mathcal{X} = 0$ for a burrito, $\mathcal{X} = 1$ for a taco, $\mathcal{Y} = 0$ for vegetables, and $\mathcal{Y} = 1$ for meat. The joint PMF of $\mathcal{X}$ and $\mathcal{Y}$ is given as

$\mathbb{P}(x,y)$		$\mathcal{X}$ 0	1
$\mathcal{Y}$	0	1/6	1/6
	1	1/2	1/6

Let $f(x,y)$ represent the amount of money spent. Then

$$f(0,0) = 2+1 = 3, \qquad f(0,1) = 2+2 = 4, \qquad f(1,0) = 1+1 = 2, \qquad f(1,1) = 1+2 = 3,$$

and

$$\begin{aligned}
\mathbb{E}[f(\mathcal{X},\mathcal{Y})] &= \sum_{\mathcal{X},\mathcal{Y}} f(x,y)\mathbb{P}(\mathcal{X}=x,\mathcal{Y}=y) \\
&= f(0,0)p_{\mathcal{X},\mathcal{Y}}(0,0) + f(0,1)p_{\mathcal{X},\mathcal{Y}}(0,1) + f(1,0)p_{\mathcal{X},\mathcal{Y}}(1,0) + f(1,1)p_{\mathcal{X},\mathcal{Y}}(1,1) \\
&= 3\cdot 1/6 + 4\cdot 1/2 + 2\cdot 1/6 + 3\cdot 1/6 \\
&= 10/3
\end{aligned}$$

Linearity of expectation. Recall that

$$\sum_{\mathcal{X},\mathcal{Y}} (x+y)p_{\mathcal{X},\mathcal{Y}}(x,y) = \sum_{\mathcal{X},\mathcal{Y}} xp_{\mathcal{X},\mathcal{Y}}(x,y) + \sum_{\mathcal{X},\mathcal{Y}} yp_{\mathcal{X},\mathcal{Y}}(x,y).$$

With a bit of thought, this leads to

$$\mathbb{E}[\mathcal{X}+\mathcal{Y}] = \mathbb{E}[\mathcal{X}] + \mathbb{E}[\mathcal{Y}].$$

This fact is called the *linearity of expectation*, and naturally generalises to n random variables $\mathcal{X}_1, \mathcal{X}_2, \ldots, \mathcal{X}_n$:

$$\mathbb{E}\left[\sum_{i=1}^{n} \mathcal{X}_i\right] = \sum_{i=1}^{n} \mathbb{E}[\mathcal{X}_i].$$

Note that the left hand side requires the joint PMF, and the right hand side requires the individual marginal PMFs. The linearity of expectation is a statement of their equivalence through sum.

12.1 Example: "Hat problem"

Suppose n people put their phones in a box. At the end of class, everyone takes a phone from the box randomly.

Q1. *What is the expected random number of people who get back their own phone?*
The key idea here is to introduce indicator random variables to check whether someone gets their own phone back. Label everyone $i = 1, 2, \ldots, n$, and define the indicator r.v.

$$\mathcal{X}_i = \begin{cases} 1 & \text{if person } i \text{ gets their own phone} \\ 0 & \text{otherwise.} \end{cases}$$

With this specification, the number of people who get their own phone is given by the random variable $\mathcal{Y} = \sum_{i=1}^{n} \mathcal{X}_i$. By linearity then, it is true that

$$\mathbb{E}[\mathcal{Y}] = \mathbb{E}\left[\sum_{i=1}^{n} \mathcal{X}_i\right] = \sum_{i=1}^{n} \mathbb{E}[\mathcal{X}_i].$$

For each i,

$$\begin{aligned}\mathbb{E}[\mathcal{X}_i] &= 1 \cdot \mathbb{P}(\mathcal{X}_i = 1) + 0 \cdot \mathbb{P}(\mathcal{X}_i = 0) \\ &= \mathbb{P}(\text{person } i \text{ gets their own phone}) \\ &= 1/n,\end{aligned}$$

since person i is equally likely to get any of the phones. Thus

$$\mathbb{E}[\mathcal{Y}] = \sum_{i=1}^{n} \mathbb{E}[\mathcal{X}_i] = \sum_{i=1}^{n} \frac{1}{n} = n \cdot \frac{1}{n} = 1.$$

Q2. (Much harder.) *What is the probability that nobody gets their own phone?*
In this case, we first find the number of different ways to distribute n phones to n people such that no single person receives the phone that they started with. Suppose there are $D(n)$ ways to do this. The first person has $n-1$ phones that they could receive that aren't their own. Let's say they receive the kth person's phone. Now, however, the number of phone options for the next person depends on whose phone the kth person received. For this, there are two options:

(1) The kth person received the first person's phone. In this case, we know they didn't receive their own, so now there are $n-2$ people without phones and $n-2$ phones left. To distribute these to make sure no one else gets their own phone, this is the same problem we started with, except with 2 less phones. So now we must solve for $D(n-2)$.

(2) The kth person did not receive the first person's phone. They could still receive their own phone, though, so we want to make sure they don't. Thus we still need to distribute $n-1$ phones to $n-1$ people without giving anyone their own phone, so this problem is now $D(n-1)$.

We can combine both cases into one expression for $D(n)$. There are $n-1$ options for the first person, and after that there are $D(n-2) + D(n-1)$ ways left. Thus we have the recurrence relation with initial conditions $D(0) = 1$, $D(1) = 0$, and

$$D(n) = (n-1)(D(n-2) + D(n-1)).$$

We proceed to solve this recurrence. Subtract $nD(n-1)$ from both sides of the recurrence relation to get

$$D(n) - nD(n-1) = -(D(n-1) - (n-1)D(n-2)).$$

Now let $A(n) = D(n) - nD(n-1)$, and the recurrence relation becomes $A(n) = -A(n-1)$. With $A(0) = 1$, this recurrence is easy, and we get $D(n) - nD(n-1) = A(n) = (-1)^n$. Divide both sides of this recurrence by $n!$ to obtain

$$\frac{D(n)}{n!} - \frac{D(n-1)}{(n-1)!} = \frac{(-1)^n}{n!}.$$

Now let $B(n) = \dfrac{D(n)}{n!}$. Thus we have $B(n) = B(n-1) + \dfrac{(-1)^n}{n!}$. With $B(0) = 1$, this recurrence is also easy, and we get

$$\frac{D(n)}{n!} = B(n) = \sum_{k=0}^{n} \frac{(-1)^k}{k!} \implies D(n) = n! \sum_{k=0}^{n} \frac{(-1)^k}{k!}$$

This expression gives the number of derangements $D(n)$ on n elements; the desired probability is then $D(n)/n!$, i.e. $\displaystyle\sum_{k=0}^{n} \frac{(-1)^k}{k!}$. This probability $\to 1/e$ as $n \to \infty$.

12.2 Conditioning with random variables

Definition 12.1 (Conditional PMF)**.** *Let $\mathcal{X}$ be an r.v. and A an event with $\mathbb{P}(A) > 0$. The conditional PMF of $\mathcal{X}$ given A is the function $p_{\mathcal{X}|A}(x) = \mathbb{P}(\mathcal{X} = x \mid A) = \dfrac{\mathbb{P}(\{\mathcal{X} = x\} \cap A)}{\mathbb{P}(A)}$ for $x \in \mathbb{R}$.*

Remarks. The following are true:

- The conditional PMF also forms a probability simplex: $\displaystyle\sum_x p_{\mathcal{X},A}(x) = 1$.
- We can use it to calculate other probabilities, e.g. $\mathbb{P}(\mathcal{X} \leq 8 \mid A) = \displaystyle\sum_{x \leq 8} p_{\mathcal{X}|A}(x)$.
- The *conditional expectation* of $\mathcal{X}$ given A is $\mathbb{E}[\mathcal{X} \mid A] = \displaystyle\sum_x x\mathbb{P}(\mathcal{X} = x \mid A)$.

Example 12.2. Roll a die. Let the r.v. $\mathcal{X}$ be the value rolled. Then the PMF is

$$\mathbb{P}(\mathcal{X} = x) = \begin{cases} 1/6 & \text{if } x = 1, 2, 3, 4, 5, 6 \\ 0 & \text{otherwise.} \end{cases}$$

Let $A = \{\mathcal{X} \text{ is even}\}$. Note that $\mathbb{P}(A) = 1/2$. The conditional PMF of $\mathcal{X}$ given $\{\mathcal{X} \text{ is even}\}$ is

$$\mathbb{P}(\mathcal{X} = 2 \mid A) = \frac{\mathbb{P}(\{\mathcal{X} = x\} \cap \{\mathcal{X} \text{ is even}\})}{\mathbb{P}(\mathcal{X} \text{ is even})} = \frac{\mathbb{P}(\mathcal{X} = 2)}{\mathbb{P}(\mathcal{X} \text{ is even})} = \frac{1/6}{1/2} = \frac{1}{3},$$

and 0 otherwise. (Note that the conditional PMF is the same if $x = 2, 4, 6$.) The tabular plot of the conditional PMF $\mathbb{P}(\mathcal{X} = x \mid \mathcal{X} \text{ is even})$ is shown below:

x	1	2	3	4	5	6
$\mathbb{P}(\mathcal{X} = x)$	1/6	1/6	1/6	1/6	1/6	1/6
$\mathbb{P}(\mathcal{X} = x \mid \mathcal{X} \text{ is even})$	0	1/3	0	1/3	0	1/3

The conditional expectation is then

$$\begin{aligned}\mathbb{E}[\mathcal{X} \mid \mathcal{X} \text{ is even}] &= \sum_x x\mathbb{P}(\mathcal{X} = x \mid \mathcal{X} \text{ is even}) \\ &= 2 \cdot \mathbb{P}(\mathcal{X} = 2 \mid \mathcal{X} \text{ is even}) + 4 \cdot \mathbb{P}(\mathcal{X} = 4 \mid \mathcal{X} \text{ is even}) + 6 \cdot \mathbb{P}(\mathcal{X} = 6 \mid \mathcal{X} \text{ is even}) \\ &= 2 \cdot 1/3 + 4 \cdot 1/3 + 6 \cdot 1/3 \\ &= 4.\end{aligned}$$

12.2.1 Conditioning with two random variables

Definition 12.2 (Conditional PMF of conditional)**.** *Let $\mathcal{X}$ and $\mathcal{Y}$ be random variables with joint PMF $p_{\mathcal{X},\mathcal{Y}}$. Recall the definitions of the joint PMF—$p_{\mathcal{X},\mathcal{Y}}(x,y) = \mathbb{P}(\mathcal{X} = x, \mathcal{Y} = y)$—and the marginal PMFs—$p_{\mathcal{X}}(x) = \sum_y p_{\mathcal{X},\mathcal{Y}}(x,y)$ and $p_{\mathcal{Y}}(y) = \sum_x p_{\mathcal{X},\mathcal{Y}}(x,y)$.*

Then the conditional PMF of $\mathcal{X}$ given $\mathcal{Y}$ is the function of two variables $p_{\mathcal{X}|\mathcal{Y}}$ defined by

$$p_{\mathcal{X}|\mathcal{Y}}(x \mid y) = \mathbb{P}(\mathcal{X} = x \mid \mathcal{Y} = y) \qquad \text{for all } x, y \in \mathbb{R} \text{ with } \mathbb{P}(\mathcal{Y} = y) > 0.$$

Note that

$$p_{\mathcal{X}|\mathcal{Y}}(x \mid y) = \frac{\mathbb{P}(\mathcal{X} = x, \mathcal{Y} = y)}{\mathbb{P}(\mathcal{Y} = y)} = \frac{p_{\mathcal{X},\mathcal{Y}}(x,y)}{p_{\mathcal{Y}}(y)} = \frac{\text{joint PMF}}{\text{marginal PMF}}.$$

Example 12.3. Recall the joint PMF from last class:

$\mathbb{P}(x,y)$		$\mathcal{X}$ = 2	3	4	5	6
$\mathcal{Y}$	1	1/9	0	0	0	0
	2	0	2/9	1/9	0	0
	3	0	0	2/9	2/9	1/9

We can represent the conditional PMF of $\mathcal{X}$ given $\mathcal{Y}$, $p_{\mathcal{X}|\mathcal{Y}}(x \mid y)$, as a table by dividing each entry of the joint PMF table by the corresponding $p_{\mathcal{Y}}(y)$ value (the sum of the rows along y). As an example,

$$p_{\mathcal{X}|\mathcal{Y}}(4 \mid 2) = \frac{p_{\mathcal{X},\mathcal{Y}}(4,2)}{p_{\mathcal{Y}}(2)} = \frac{1/9}{3/9} = \frac{1}{3}.$$

The conditional PMF table for this example is

$\mathbb{P}(x,y)$		$\mathcal{X}$ 2	3	4	5	6
$\mathcal{Y}$	1	1	0	0	0	0
	2	0	2/3	1/3	0	0
	3	0	0	2/5	2/5	1/5

Note that the row entries now sum to 1.

13 Conditioning and Independence (25/10)

13.1 More Conditioning

Recap: For two discrete random variables $\mathcal{X}$ and $\mathcal{Y}$,

- the **joint PMF** is $p_{\mathcal{X},\mathcal{Y}}(x,y) = \mathbb{P}(\mathcal{X}=x, \mathcal{Y}=y)$ for $x, y \in \mathbb{R}$.
- the **marginals** are $p_{\mathcal{X}}(x) = \mathbb{P}(\mathcal{X}=x) = \sum_y p_{\mathcal{X},\mathcal{Y}}(x,y)$ and $p_{\mathcal{Y}}(y) = \mathbb{P}(\mathcal{Y}=y) = \sum_x p_{\mathcal{X},\mathcal{Y}}(x,y)$
- the **conditional PMFs** are $p_{\mathcal{X}|\mathcal{Y}}(x \mid y) = \mathbb{P}(\mathcal{X}=x \mid \mathcal{Y}=y) = \dfrac{p_{\mathcal{X},\mathcal{Y}}(x,y)}{p_{\mathcal{Y}}(y)}$ and $p_{\mathcal{Y}|\mathcal{X}}(y \mid x) = \mathbb{P}(\mathcal{Y}=y \mid \mathcal{X}=x) = \dfrac{p_{\mathcal{X},\mathcal{Y}}(x,y)}{p_{\mathcal{X}}(x)}$.
- the **conditional expectation** is $\mathbb{E}[\mathcal{X} \mid \mathcal{Y}=y] = \sum_x \mathbb{P}(\mathcal{X}=x \mid \mathcal{Y}=y) = \sum_x p_{\mathcal{X}|\mathcal{Y}}(x \mid y)$.

Multiplication rule. It is true that

$$p_{\mathcal{X},\mathcal{Y}}(x,y) = p_{\mathcal{X}|\mathcal{Y}}(x \mid y) p_{\mathcal{Y}}(y).$$

Likewise, $p_{\mathcal{X},\mathcal{Y}}(x,y) = p_{\mathcal{Y}|\mathcal{X}}(y \mid x) p_{\mathcal{X}}(x)$, both of which lead to a form of Bayes' rule:

$$p_{\mathcal{Y}|\mathcal{X}}(y \mid x) = \frac{p_{\mathcal{X},\mathcal{Y}}(x \mid y) p_{\mathcal{Y}}(y)}{p_{\mathcal{X}}(x)}.$$

Example 13.1. Flip n coins. Let $\mathcal{Y}$ be the number of heads. Re-flip the coins that show heads. Let $\mathcal{X}$ be the number of remaining heads.

- *Find the joint PMF of* $(\mathcal{X}, \mathcal{Y})$. We are given the marginal of $\mathcal{Y} \sim \mathsf{Binomial}(n, 1/2)$ as $p_{\mathcal{Y}}(y) = \binom{n}{y} 2^{-n}$ for $y = 0, 1, \ldots, n$. Also, we are given the conditional of $\mathcal{X} \sim \mathsf{Binomial}(y, 1/2)$ given $\mathcal{Y}$ as $p_{\mathcal{X}|\mathcal{Y}}(x \mid y) = \binom{y}{x} 2^{-y}$ for $x = 0, 1, \ldots, y$.
  By the multiplication rule, we get the joint PMF
  $$p_{\mathcal{X},\mathcal{Y}}(x,y) = p_{\mathcal{X}|\mathcal{Y}}(x \mid y) p_{\mathcal{Y}}(y) = \binom{n}{y} 2^{-n} \cdot \binom{y}{x} 2^{-y} = \binom{n}{y}\binom{y}{x} 2^{-(n+y)}$$
  for $y = 0, 1, \ldots, n$ and $x = 0, 1, \ldots, y$ and $p_{\mathcal{X},\mathcal{Y}}(x,y) = 0$ otherwise.
- *Compute the conditional expectation.* We could easily say
  $$\mathbb{E}[\mathcal{X} \mid \mathcal{Y}=y] = \sum_x x p_{\mathcal{X}|\mathcal{Y}}(x \mid y) = \sum_{x=0}^{y} x \binom{y}{x} 2^{-y}$$
  which is hard to simplify. Alternatively, however, we could recognise that $\mathcal{X} \mid \mathcal{Y}=y \sim \mathsf{Binomial}(y, 1/2)$, so that
  $$\mathbb{E}[\mathcal{X} \mid \mathcal{Y}=y] = \mathbb{E}[\mathsf{Binomial}(y, 1/2)] = y/2,$$

since the mean of a $\mathsf{Binomial}(m, q)$ is mq.

- *Find the marginal of $\mathcal{X}$.* We have that

$$p_{\mathcal{X}}(x) = \sum_y p_{\mathcal{X},\mathcal{Y}}(x, y) = \sum_{y=x}^{n} \binom{n}{y}\binom{y}{x} 2^{-(n+y)},$$

which is hard to simplify.

- *Compute the expectation.* Note here that the direct approach for computing the expectation over x is hard because the formula for $p_{\mathcal{X}}(x)$ is not simple. We might intuitively guess that $\mathbb{E}[X] = n/2$ since we are making binary observations twice. Regardless, let us introduce a new tool, the law of total expectation, to help us with these computations:

$$\mathbb{E}[\mathcal{X}] = \sum_y \frac{y}{2} p_{\mathcal{Y}}(y) = \frac{1}{2} \sum_y y p_{\mathcal{Y}}(y),$$

which simplifies to

$$\frac{1}{2} \cdot \mathbb{E}[\mathcal{Y}] = 1/2 \cdot n/2 = n/4.$$

Law of total expectation. The unconditional average of an event can be obtained by averaging the conditional averages, i.e.

$$\mathbb{E}[\mathcal{X}] = \sum_y \mathbb{E}[\mathcal{X} \mid \mathcal{Y} = y] \cdot p_{\mathcal{Y}}(y)$$

for random variables $\mathcal{X}$ and $\mathcal{Y}$.

Proof. We have that

$$\begin{aligned}
\sum_y \mathbb{E}[\mathcal{X} \mid \mathcal{Y} = y] p_{\mathcal{Y}}(y) &= \sum_y \left(\sum_x x p_{\mathcal{X}|\mathcal{Y}}(x \mid y) \right) p_{\mathcal{Y}}(y) \\
&= \sum_y \left(\sum_x x p_{\mathcal{X}|\mathcal{Y}}(x \mid y) p_{\mathcal{Y}}(y) \right) \\
&= \sum_y \left(\sum_x x p_{\mathcal{X},\mathcal{Y}}(x, y) \right) \\
&= \sum_x \left(\sum_y x p_{\mathcal{X},\mathcal{Y}}(x, y) \right) \\
&= \sum_x x \left(\sum_y p_{\mathcal{X},\mathcal{Y}}(x, y) \right) \\
&= \sum_x x p_{\mathcal{X}}(y) \\
&= \mathbb{E}[\mathcal{X}],
\end{aligned}$$

as desired. ■

13.2 Independence

Definition 13.1 (Independence)**.** *Two random variables $\mathcal{X}$ and $\mathcal{Y}$ are independent if*

$$p_{\mathcal{X},\mathcal{Y}}(x,y) = p_{\mathcal{X}}(x)p_{\mathcal{Y}}(y), \text{ i.e. } \mathbb{P}(\mathcal{X}=x,\mathcal{Y}=y) = \mathbb{P}(\mathcal{X}=x)\mathbb{P}(\mathcal{Y}=y) \text{ for all } x,y \in \mathbb{R}.$$

In terms of conditional PMFs, independence means $p_{\mathcal{X}|\mathcal{Y}}(x \mid y) = p_{\mathcal{X}}(x)$ for all $x, y \in \mathbb{R}$. Similarly, $p_{\mathcal{Y}|\mathcal{X}}(y \mid x) = p_{\mathcal{Y}}(y)$ for all $x, y \in \mathbb{R}$.

Remark. If we are given marginals $p_{\mathcal{X}}$ and $p_{\mathcal{Y}}$ and are told that $\mathcal{X}$ and $\mathcal{Y}$ are independent, then we know the joint PMF immediately: $p_{\mathcal{X},\mathcal{Y}}(x,y) = p_{\mathcal{X}}(x)p_{\mathcal{Y}}(y)$.

Example 13.2. Let $\mathcal{X}$ and $\mathcal{Y}$ be independent $\mathsf{Geometric}(p)$ random variables, so that

$$p_{\mathcal{X}}(x) = p(1-p)^{x-1} \qquad \text{for } x = 1,2,\dots$$
$$p_{\mathcal{Y}}(y) = p(1-p)^{y-1} \qquad \text{for } y = 1,2,\dots$$

so that $p_{\mathcal{X},\mathcal{Y}}(x,y) = p_{\mathcal{X}}(x)p_{\mathcal{Y}}(y) = p^2(1-p)^{x+y-2}$ for $x, y = 1, 2, \dots$.

13.2.1 Properties

Let $\mathcal{X}$ and $\mathcal{Y}$ be independent, meaning $p_{\mathcal{X},\mathcal{Y}}(x,y) = p_{\mathcal{X}}(x)p_{\mathcal{Y}}(y)$. Then,

(1) For sets $A \subset \mathbb{R}$ and $B \subset \mathbb{R}$, $\mathbb{P}(\mathcal{X} \in A, \mathcal{Y} \in B) = \mathbb{P}(\mathcal{X} \in A)\mathbb{P}(\mathcal{Y} \in B)$—probabilities factorise.

Example 13.3. $\mathbb{P}(1 \le \mathcal{X} \le 4, \mathcal{Y} \ge 2) = \mathbb{P}(1 \le \mathcal{X} \le 4)\mathbb{P}(\mathcal{Y} \ge 2)$.

Caveat: We cannot factorise probabilities for non-rectangular regions, i.e. regions that have $\mathcal{X}$ and $\mathcal{Y}$ mixed together like $\mathbb{P}(\mathcal{X} > \mathcal{Y})$ or $\mathbb{P}(\mathcal{X} + \mathcal{Y} \le 4)$.

(2) For any functions $f, g : \mathbb{R} \to \mathbb{R}$, multiplicity of independence holds every time: $\mathbb{E}[f(\mathcal{X})g(\mathcal{Y})] = \mathbb{E}[f(\mathcal{X})]\mathbb{E}[g(\mathcal{Y})]$.

Example 13.4. $\mathbb{E}[\mathcal{X}\mathcal{Y}] = \mathbb{E}[\mathcal{X}]\mathbb{E}[\mathcal{Y}]$.

Caveat: This really needs independence! If $\mathcal{X}$ and $\mathcal{Y}$ are not independent, $\mathbb{E}[\mathcal{X}\mathcal{Y}]$ and $\mathbb{E}[\mathcal{X}]\mathbb{E}[\mathcal{Y}]$ may be different. For example, if $\mathcal{Y} = \mathcal{X}$ then $\mathbb{E}[\mathcal{X}\mathcal{Y}] = \mathbb{E}[\mathcal{X}^2]$ whereas $\mathbb{E}[\mathcal{X}]\mathbb{E}[\mathcal{Y}] = (\mathbb{E}[\mathcal{X}])^2$, and we know $\mathrm{Var}(\mathcal{X}) = \mathbb{E}[\mathcal{X}^2] - (\mathbb{E}[\mathcal{X}])^2 \ge 0$, a strict inequality that always holds unless $\mathcal{X}$ is constant.
By contrast, linearity of expectations holds regardless of the independence criterion.

(3) (Linearity of variance.) $\mathrm{Var}(\mathcal{X} + \mathcal{Y}) = \mathrm{Var}(\mathcal{X}) + \mathrm{Var}(\mathcal{Y})$. This fails if $\mathcal{X}$ and $\mathcal{Y}$ are not independent.

Proof. Let $\hat{\mathcal{X}} = \mathcal{X} - \mathbb{E}[\mathcal{X}]$ and $\hat{\mathcal{Y}} = \mathcal{Y} - \mathbb{E}[\mathcal{Y}]$. Then

$$\mathbb{E}[\hat{\mathcal{X}}] = \mathbb{E}[\mathcal{X} - \mathbb{E}[\mathcal{X}]] = \mathbb{E}[\mathcal{X}] - \mathbb{E}[\mathcal{X}] = 0,$$

and similarly $\mathbb{E}[\hat{\mathcal{Y}}] = 0$. In addition,

$$\mathrm{Var}(\mathcal{X}) = \mathbb{E}[(\mathcal{X} - \mathbb{E}[\mathcal{X}])^2] = \mathbb{E}[\hat{\mathcal{X}}^2]$$

and

$$\mathrm{Var}(\mathcal{Y}) = \mathbb{E}[(\mathcal{Y} - \mathbb{E}[\mathcal{Y}])^2] = \mathbb{E}[\hat{\mathcal{Y}}^2].$$

Finally,

$$\begin{aligned}
\mathrm{Var}(\mathcal{X}+\mathcal{Y}) &= \mathbb{E}[(\mathcal{X}+\mathcal{Y} - \mathbb{E}[\mathcal{X}+\mathcal{Y}])^2] \\
&= \mathbb{E}[(\mathcal{X}+\mathcal{Y} - \mathbb{E}[\mathcal{X}] - \mathbb{E}[\mathcal{Y}])^2] \\
&= \mathbb{E}[(\hat{\mathcal{X}}+\hat{\mathcal{Y}})^2] \\
&= \mathbb{E}[\hat{\mathcal{X}}^2 + \hat{\mathcal{Y}}^2 + 2\hat{\mathcal{X}}\hat{\mathcal{Y}}] \\
&= \mathbb{E}[\hat{\mathcal{X}}^2] + \mathbb{E}[\hat{\mathcal{Y}}^2] + 2\mathbb{E}[\hat{\mathcal{X}}\hat{\mathcal{Y}}] \\
&= \mathrm{Var}(\mathcal{X}) + \mathrm{Var}(\mathcal{Y}) + 2\mathbb{E}[\hat{\mathcal{X}}\hat{\mathcal{Y}}]
\end{aligned}$$

The last term is zero by independence: $\mathbb{E}[\hat{\mathcal{X}}\hat{\mathcal{Y}}] = \mathbb{E}[\hat{\mathcal{X}}]\mathbb{E}[\hat{\mathcal{Y}}] = 0$, and we are done. ■

14 Independence, covariance, and correlation (27/10)

Example 14.1. Let $0 < p < 1$ and n a positive integer be given. Let $\mathcal{X}_1, \mathcal{X}_2, \ldots, \mathcal{X}_n$ be independent $\mathsf{Bernoulli}(p)$. We might interpret these random variables as

$$\mathcal{X}_i = \begin{cases} 1 & \text{if } i\text{th trial is a success} \\ 0 & \text{otherwise.} \end{cases}$$

Let $\mathcal{X} = \sum_{i=1}^{n} \mathcal{X}_i$. This counts the total number of successes out of the n trials, thus $\mathcal{X} \sim \mathsf{Binomial}(n, p)$. This representation of the binomial distribution—as the sum of independent Bernoulli distributions—is very convenient for calculations.

Example 14.2. To find the mean, we can simply do

$$\mathbb{E}[\mathcal{X}] = \mathbb{E}\left[\sum_{i=1}^{n} \mathcal{X}_i\right] = \sum_{i=1}^{n} \mathbb{E}\left[\mathcal{X}_i\right] = \sum_{i=1}^{n} p = np,$$

and the variance is

$$\mathrm{Var}(\mathcal{X}) = \mathrm{Var}\left(\sum_{i=1}^{n} \mathcal{X}_i\right) = \sum_{i=1}^{n} \mathrm{Var}(\mathcal{X}_i) = \sum_{i=1}^{n} p(1-p) = np(1-p).$$

Example 14.3. *Let the independent random variables $\mathcal{X} \sim \mathsf{Poisson}(\lambda)$ and $\mathcal{Y} \sim \mathsf{Poisson}(\mu)$ for $\lambda, \mu > 0$. Find the PMF of $\mathcal{Z} = \mathcal{X} + \mathcal{Y}$.*
Intuitively, here are two possible interpretations of the random variables. $\mathcal{X}$ represents the

total number of local trains per hour coming in with rate λ and $\mathcal{Y}$ represents the total number of express trains per hour coming in with rate μ, so that $\mathcal{Z}$ gives the number of trains (of any kind) coming in per hour with rate $\lambda + \mu$. It is obvious then that $\mathcal{Z} \sim \mathsf{Poisson}(\lambda + \mu)$. Recall that $\mathcal{X}$ and $\mathcal{Y}$ have PMFs

$$\mathbb{P}(\mathcal{X} = k) = e^{-\lambda}\frac{\lambda^k}{k!}, \mathbb{P}(\mathcal{Y} = k) = e^{-\mu}\frac{\mu^k}{k!},$$

for $k = 0, 1, \ldots$. Thus $\mathcal{Z}$ must always be non-negative, so that

$$\mathbb{P}(\mathcal{Z} = n) = \mathbb{P}(\mathcal{X} + \mathcal{Y} = n) = \sum_{\substack{(x,y) \text{ s.t.} \\ x+y=n}} \mathbb{P}(\mathcal{X} = x, \mathcal{Y} = y) = \sum_{k=0}^{n} \mathbb{P}(\mathcal{X} = k, \mathcal{Y} = n - k).$$

Simplifying using independence and the given PMFs, we have

$$\begin{aligned}\mathbb{P}(\mathcal{X} = k, \mathcal{Y} = n - k) = \mathbb{P}(\mathcal{X} = k)\mathbb{P}(\mathcal{Y} = n - k) &= \left(e^{-\lambda}\frac{\lambda^k}{k!}\right)\left(e^{-\mu}\frac{\mu^{n-k}}{(n-k)!}\right) \\ &= e^{-(\lambda+\mu)}\frac{1}{k!(n-k)!}\lambda^k\mu^{n-k} = \frac{e^{-(\lambda+\mu)}}{n!}\binom{n}{k}\lambda^k\mu^{n-k}\end{aligned}$$

Plugging this in, we have that

$$\begin{aligned}\mathbb{P}(\mathcal{Z} = n) = \mathbb{P}(\mathcal{X} = k, \mathcal{Y} = n - k) &= \sum_{k=0}^{n} \frac{e^{-(\lambda+\mu)}}{n!}\binom{n}{k}\lambda^k\mu^{n-k} \\ &= \frac{e^{-(\lambda+\mu)}}{n!}\sum_{k=0}^{n}\binom{n}{k}\lambda^k\mu^{n-k} = \frac{e^{-(\lambda+\mu)}}{n!}(\lambda + \mu)^n,\end{aligned}$$

which is exactly $\mathbb{P}(\mathsf{Poisson}(\lambda + \mu) = n)$.

This is a special property of Poissons: the sum of independent Poissons is again a Poisson.

14.1 Measures of dependence—covariance and correlation

Definition 14.1 (Covariance)**.** *For random variables $\mathcal{X}$ and $\mathcal{Y}$, the covariance is defined by*

$$\mathrm{Cov}(\mathcal{X}, \mathcal{Y}) = \mathbb{E}[(\mathcal{X} - \mathbb{E}[\mathcal{X}])(\mathcal{Y} - \mathbb{E}[\mathcal{Y}])],$$

and quantifies how much $\mathcal{X}$ and $\mathcal{Y}$ tend to deviate from their means in the same direction. Note that $\mathrm{Cov}(\mathcal{X}, \mathcal{X})$ can be positive, negative, or zero, and that $\mathrm{Cov}(\mathcal{X}, \mathcal{X}) = \mathrm{Var}(\mathcal{X})$.

The following are a few properties of covariance. Their proofs are left as exercises.

- (Symmetry.) $\mathrm{Cov}(\mathcal{X}, \mathcal{Y}) = \mathrm{Cov}(\mathcal{Y}, \mathcal{X})$.
- For constants $a, b, c, d \in \mathbb{R}$, $\mathrm{Cov}(a\mathcal{X} + b, c\mathcal{Y} + d) = ac\mathrm{Cov}(\mathcal{X}, \mathcal{Y})$.
- (Bilinearity.) $\mathrm{Cov}(\mathcal{X} + \mathcal{Y}, \mathcal{Z}) = \mathrm{Cov}(\mathcal{X}, \mathcal{Z}) + \mathrm{Cov}(\mathcal{Y}, \mathcal{Z})$.

- (Alternative formula.) $\text{Cov}(\mathcal{X}, \mathcal{Y}) = \mathbb{E}[\mathcal{X}\mathcal{Y}] - \mathbb{E}[\mathcal{X}]\,\mathbb{E}[\mathcal{Y}]$.
- If $\mathcal{X}$ and $\mathcal{Y}$ are independent random variables, then $\text{Cov}(\mathcal{X}, \mathcal{Y}) = 0$, because independence implies $\mathbb{E}[\mathcal{X}\mathcal{Y}] = \mathbb{E}[\mathcal{X}]\,\mathbb{E}[\mathcal{Y}]$.

Remark. We say that $\mathcal{X}$ and $\mathcal{Y}$ are *uncorrelated* if $\text{Cov}(\mathcal{X}, \mathcal{Y}) = 0$. In fact,

$$\begin{aligned} &\mathcal{X} \text{ and } \mathcal{Y} \text{ are independent } \implies \mathcal{X} \text{ and } \mathcal{Y} \text{ are uncorrelated;} \\ &\mathcal{X} \text{ and } \mathcal{Y} \text{ are uncorrelated } \not\implies \mathcal{X} \text{ and } \mathcal{Y} \text{ are independent.} \end{aligned}$$

(To get an idea of why, check that uncorrelated means $\mathbb{E}[\mathcal{X}\mathcal{Y}] = \mathbb{E}[\mathcal{X}]\,\mathbb{E}[\mathcal{Y}]$, while independence means $\mathbb{E}[f(\mathcal{X})g(\mathcal{Y})] = \mathbb{E}[f(\mathcal{X})]\,\mathbb{E}[g(\mathcal{Y})]$ for all functions f and g.)

As in variance, covariance is in (units)^2 if X is in (units). The unitless version of covariance is *correlation*. (Covariance measures correlation in addition to size.)

Definition 14.2. *The correlation between two random variables $\mathcal{X}$ and $\mathcal{Y}$ is*

$$\rho(\mathcal{X}, \mathcal{Y}) = \frac{\text{Cov}(\mathcal{X}, \mathcal{Y})}{\sigma(\mathcal{X})\sigma(\mathcal{Y})} = \frac{\text{Cov}(\mathcal{X}, \mathcal{Y})}{\sqrt{\text{Var}(\mathcal{X})\text{Var}(\mathcal{Y})}}.$$

Remark. From the Cauchy-Schwarz inequality, we can show that $|\text{Cov}(\mathcal{X}, \mathcal{Y})| \le \sigma(\mathcal{X})\sigma(\mathcal{Y})$ from which we get that $-1 \le \rho(\mathcal{X}, \mathcal{Y}) \le 1$.

Some intuition for understanding correlation coefficients. If $\rho(\mathcal{X}, \mathcal{Y}) > 0$, we say $\mathcal{X}$ and $\mathcal{Y}$ are positively correlated. If $\rho(\mathcal{X}, \mathcal{Y}) < 0$, we say that $\mathcal{X}$ and $\mathcal{Y}$ are negatively correlated. If $\rho(\mathcal{X}, \mathcal{Y}) = 0$, we say that they are uncorrelated.

Example 14.4. Extreme values of ± 1 are perfect positive/negative correlation.

Fact. If $\mathcal{Y}$ is a linear transformation of $\mathcal{X}$, eg $\mathcal{Y} = a\mathcal{X} + b$ for constants a and b, then

$$\rho(\mathcal{X}, \mathcal{Y}) = \begin{cases} 1 & \text{if } a > 0 \\ -1 & \text{if } a < 0. \end{cases}$$

Proof. Recall the following properties of variance and covariance:

$$\text{Var}(\mathcal{Y}) = \text{Var}(a\mathcal{X} + b) = a^2\text{Var}(\mathcal{X})$$

$$\text{Cov}(\mathcal{X}, \mathcal{Y}) = \text{Cov}(\mathcal{X}, a\mathcal{X} + b) = a\text{Cov}(\mathcal{X}, \mathcal{X}) = a\text{Var}(\mathcal{X}).$$

Thus

$$\rho(\mathcal{X}, \mathcal{Y}) = \frac{\text{Cov}(\mathcal{X}, \mathcal{Y})}{\sqrt{\text{Var}(\mathcal{X})\text{Var}(\mathcal{Y})}} = \frac{a\text{Var}(\mathcal{X})}{\sqrt{\text{Var}(\mathcal{X})a^2\text{Var}(\mathcal{X})}} = \frac{a}{\sqrt{a^2}} = \frac{a}{|a|} = \begin{cases} 1 & \text{if } a > 0 \\ -1 & \text{if } a < 0 \end{cases}$$

as desired. ∎

Example 14.5. Flip n coins. Let $\mathcal{X}$ represent the number of heads, and $\mathcal{Y}$ represent the number of tails. Then $\mathcal{X}$ and $\mathcal{Y}$ are perfectly negatively correlated, and $\rho(\mathcal{X}, \mathcal{Y}) = -1$. Mathematically, this is because $\mathcal{X} + \mathcal{Y} = n$, or $\mathcal{Y} = n - \mathcal{X}$—now use the previous fact with $a = -1$ and $b = n$. Intuitively, this is because a larger $\mathcal{X}$ corresponds to a smaller $\mathcal{Y}$.

Example 14.6. Flip 10 coins. Let $\mathcal{X}$ be the number of heads, and $\mathcal{Y}$ be the number of heads out of the first five flips. Intuitively, $0 < \rho(\mathcal{X}, \mathcal{Y}) < 1$.

15 Continuous random variables (01/11)

Recall our earlier definition of a probability measure $\mathbb{P}$ as a function from *events* (i.e. subsets of the sample space Ω). Along with this description came the axioms of probability, including $\mathbb{P}(\Omega) = 1$, $\mathbb{P}(E) \geq 0$ for any event E, and $\mathbb{P}(E \cup F) = \mathbb{P}(E) + \mathbb{P}(F)$ for two disjoint events E and F

Now suppose this sample space Ω is continuous (the sample space is uncountably infinite), for example, say $\Omega = [0, 1]$. Interesting things can happen here—hang tight.

Example 15.1. Suppose Ω contains all the real numbers from 0 to 1. Then the probability measure $\mathbb{P}$ is defined as $\mathbb{P}([a, b]) = b - a$ for $0 \leq a \leq b \leq 1$, i.e. the length of the interval from a to b on the number line.

Example 15.2 (Additivity)**.** Suppose we have $0 \leq a \leq b \leq c \leq 1$. Then

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

Example 15.3. With this description of $\mathbb{P}$, individual points have zero probability: take $b = a$, so that

$$\mathbb{P}(\{a\}) = \mathbb{P}([a, a]) = a - a = 0,$$

for each possible $a \in \Omega$.

Example 15.4. Let $\Omega \in \mathbb{R}^2$ be some region. Then the 'uniform' probability measure on Ω is defined by

$$\mathbb{P}(A) = \frac{\text{Area}(A)}{\text{Area}(\Omega)} \quad \text{for } A \subseteq \Omega.$$

We can think of many natural situations for which this model may come in handy.

Example 15.5. Suppose you are playing a strange game of darts where the winning area is the region in $\mathbb{R}^2$ bounded by $y = 0$, $x = 1$, and $y = 2x$, and that the probability of winning is higher in the region nearest the centroid of this triangle. Take $\Omega = [0, 1]$; we want to model this biased $\mathbb{P}$.
Define, for $0 \leq a \leq b \leq 1$, $\mathbb{P}([a, b])$ as the area under the line $y = 2x$ between a and b.[a]The natural strategy is to sum up these weighted probability chunks via the integral:

$$\mathbb{P}([a, b]) = \int_a^b 2x \, \mathrm{d}x = b^2 - a^2.$$

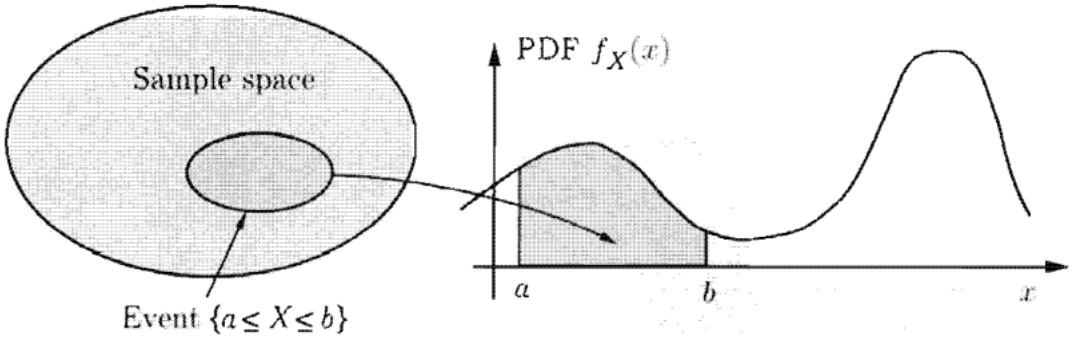
[a] $\mathbb{P}([0,1])$ is the area of a triangle with base 1 and height 2, namely 1.

Note that the statement $\int_a^b f(x\,\mathrm{d}x) + \int_b^c f(x)\,\mathrm{d}x = \int_a^c f(x)\,\mathrm{d}x$ harkens to the additivity axiom of probability.

Definition 15.1 (Probability density function)**.** *A continuous random variable $\mathcal{X}$ is a random variable $\mathcal{X} : \Omega \to \mathbb{R}$ for which there exists a function $f : \mathbb{R} \to \mathbb{R}$, with $f \geq 0$, such that*

$$\mathbb{P}(a \leq \mathcal{X} \leq b) = \int_a^b f(x)\,\mathrm{d}x \qquad \text{for all } a \leq b.$$

We call this function the probability density function or the PDF:



We can take $b = \infty$ or $a = -\infty$ here:

$$\mathbb{P}(\mathcal{X} \geq a) = \int_a^\infty f(x)\,\mathrm{d}x, \qquad \mathbb{P}(\mathcal{X} \leq b) = \int_{-\infty}^b f(x)\,\mathrm{d}x.$$

The latter is a *cumulative distribution function.*

Definition 15.2 (Cumulative distribution function)**.** *The cumulative distribution function of $\mathcal{X}$ is the function $F(x) = \mathbb{P}(\mathcal{X} \leq x) = \int_{-\infty}^x f(y)\,\mathrm{d}y$.*

Note that the derivative of the CDF is the PDF—this is only the fundamental theorem of calculus:

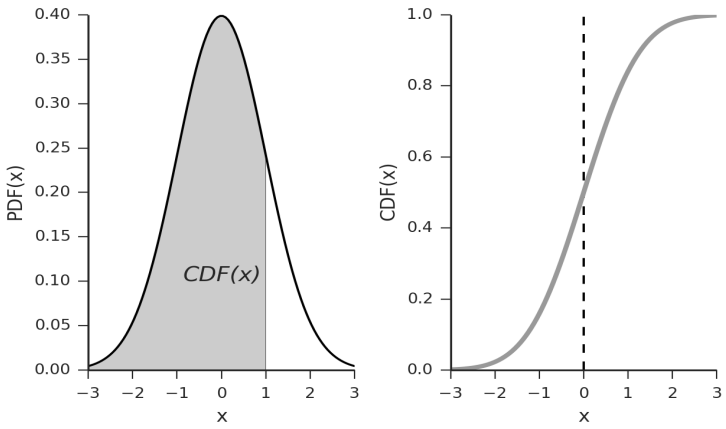
$$F'(x) = \frac{d}{dx}F(x) = \frac{d}{dx}\int_{-\infty}^x f(y)dy = f(x).$$

Definition 15.3 (Expectation and variance of continuous RV)**.** *The expectation of a continuous random variable $\mathcal{X}$ is defined by (and we can convince ourselves that this is indeed a natural definition for the expectation)*

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

for a function $g : \mathbb{R} \to \mathbb{R}$. The variance of a continuous random variable follows similarly:

$$\mathrm{Var}(\mathcal{X}) = \mathbb{E}[(\mathcal{X} - \mathbb{E}[\mathcal{X}])^2] = \int_{-\infty}^\infty (x - \mathbb{E}[\mathcal{X}])^2 f_{\mathcal{X}}(x)$$



15.1 Basic properties of the probability density function

(1) The defining property of a valid PDF is:

$$f \geq 0, \qquad \int_{-\infty}^{\infty} f(x)dx = 1.$$

Often the range of admissible values of $\mathcal{X}$ is bounded over an interval; outside this interval, $f(x) = 0$. This is helpful, because the vanishing values of f mean that all integrals can be restricted to this interval, so that if our interval was $[0, 3]$, we would have $\mathbb{P}(\mathcal{X} \leq 3) = \int_{-\infty}^{3} f(x)dx = \int_{0}^{3} f(x)dx$ and $\mathbb{E}[\mathcal{X}] = \int_{-\infty}^{\infty} xf(x)dx = \int_{0}^{3} xf(x)dx$.
We typically capture this situation using a piecewise function. In Example 15.5 above, the PDF would take the form

$$f(x) = \begin{cases} 2x & \text{if } 0 \leq x \leq 1 \\ 0 & \text{otherwise.} \end{cases}$$

(2) For any $a \in \mathbb{R}$, $\mathbb{P}(\mathcal{X} = a) = 0$. This is true because

$$\mathbb{P}(\mathcal{X} = a) = \mathbb{P}(a \leq \mathcal{X} \leq a) = \int_{a}^{a} f(x)dx = 0.$$

A consequence of this is that we can be lax about the endpoints of intervals:

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

(3) **Intuition for PDF.** Let $a \in \mathbb{R}$ and $\delta > 0$, with δ *small*. Then

$$\mathbb{P}([a, a + \delta]) = \mathbb{P}(a \leq \mathcal{X} \leq a + \delta) = \int_{a}^{a+\delta} f(x)dx \approx \delta f(a),$$

so we may interpret $f(a)$ as (roughly) the "probability per unit length" of $\mathcal{X}$ lying near the point a.

Warning. The PDF $f(x)$ is **not** a probability of any single event! (It may even exceed 1.)

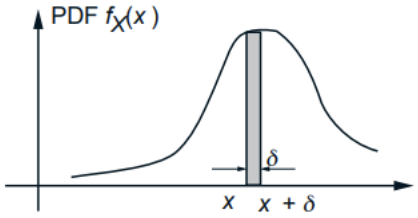


Figure 3.2: Interpretation of the PDF $f_X(x)$ as "probability mass per unit length" around x. If δ is very small, the probability that X takes value in the interval $[x, x+\delta]$ is the shaded area in the figure, which is approximately equal to $f_X(x) \cdot \delta$.

15.2 Basic properties of the cumulative distribution function

(1) It applies to random variables in general, and not just to continuous random variables:

$$F(x) = \mathbb{P}(\mathcal{X} \leq x) = \begin{cases} \sum_{k \leq x} p_{\mathcal{X}}(k) & \text{if } \mathcal{X} \text{ is discrete,} \\ \int_{-\infty}^{x} f(t)dt & \text{if } \mathcal{X} \text{ is continuous.} \end{cases}$$

We may think of the CDF as a function that "accumulates" probability "up to" the value x. Note that the CDF itself is a probability (thus $0 \leq F(x) \leq 1$), and $f(x)$ is its slope.

(2) The CDF is continuous and non-decreasing, i.e. $x \leq y \implies F(x) \leq F(y)$ for all x, y. This is because $F' = f \geq 0$.

(3) $F(x)$ tends to 0 as $x \to -\infty$, and tends to 1 as $x \to \infty$.

(4) If $a \leq b$, then

$$\mathbb{P}(a \leq \mathcal{X} \leq b) = \int_a^b f(x)dx = F(b) - F(a) = \mathbb{P}(\mathcal{X} \leq b) - \mathbb{P}(\mathcal{X} \leq a).$$

Example 15.6. Let

$$f(x) = \begin{cases} \dfrac{1}{2\sqrt{x}} & \text{if } 0 < x < 1, \\ 0 & \text{otherwise.} \end{cases}$$

This is a valid PDF because $f \geq 0$ and

$$\int_{-\infty}^{\infty} f(x)dx = \int_0^1 f(x)dx = \int_0^1 \frac{1}{2\sqrt{x}}dx = 1.$$

Suppose we wished to find the CDF $F(x) = \mathbb{P}(\mathcal{X} \leq x)$. Note that since $\mathcal{X}$ only takes values in $[0, 1]$, $F(x) = 0$ for $x \leq 0$ and $F(x) = 1$ for $x \geq 1$. For $0 \leq x \leq 1$, we then have that

$$F(x) = \int_{-\infty}^{x} f(y)dy = \int_0^x f(y)dy = \int_0^x \frac{1}{2\sqrt{y}}dy = \sqrt{x}.$$

Thus the CDF is exactly

$$F(x) = \begin{cases} 0 & \text{if } x \leq 0 \\ \sqrt{x} & \text{if } 0 < x < 1 \\ 1 & \text{if } x \geq 1. \end{cases}$$

15.3 Common Continuous RVs.

(1) **The Uniform Distribution.** Let $a \leq b$; these are the parameters of the distribution. We write $\mathcal{X} \sim \mathsf{Unif}[a, b]$ to represent the uniform random variable. This distribution eliminates biases specific to certain regions of the interval $[a, b]$, and as such we might expect its probability density function to take the form

$$f(x) = \begin{cases} \dfrac{1}{b-a} & \text{if } a \leq x \leq b \\ 0 & \text{otherwise.} \end{cases}$$

This is evidently a valid PDF, because

$$\int_{-\infty}^{\infty} f(x)dx = \int_a^b \frac{1}{b-a}dx = \frac{b-a}{b-a} = 1.$$

This PDF is uniform, so we might expect the CDF within the interval $[a, b]$ to move linearly from $(a, 0)$ to $(b, 1)$. Thus this CDF is

$$F(x) = \begin{cases} 0 & \text{if } x < a \\ \dfrac{x-a}{b-a} & \text{if } a \leq x \leq b \\ 1 & \text{if } x > b. \end{cases}$$

Because of how we defined the expectation and the uniform random variable, we might intuitively expect the expectation of the uniform random variable to be the midpoint between a and b, namely $(a + b)/2$. We can formally justify that this is true:

$$\mathbb{E}[\mathcal{X}] = \int_{-\infty}^{\infty} xf(x)dx = \int_a^b xf(x)dx = \int_a^b \frac{x}{b-a}dx = \frac{b^2-a^2}{2(b-a)} = \frac{a+b}{2}.$$

We can compute the variance of continuous random variables just like we did the discrete random variables. So for the uniform distribution, we start from

$$\mathbb{E}[\mathcal{X}^2] = \int_{-\infty}^{\infty} x^2 f(x)dx = \int_a^b \frac{x^2}{b-a}dx = \frac{b^3-a^3}{3(b-a)},$$

so that

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

Note that the standard deviation of $\mathcal{X}$ is proportional to the length of the interval—can you come up with an intuitive explanation for why this is the case?
A special case worth mentioning is that of the *standard uniform distribution* with r.v. $\mathcal{X} \sim \mathsf{Unif}[0, 1]$ (set $a = 0$ and $b = 1$). The PDF of this distribution is

$$f(x) = \begin{cases} 1 & \text{if } 0 \leq x \leq 1 \\ 0 & \text{otherwise,} \end{cases}$$

and the CDF is

$$F(x) = \begin{cases} 0 & \text{if } x < 0 \\ x & \text{if } 0 \leq x \leq 1 \\ 1 & \text{if } x > 1. \end{cases}$$

The mean and variance of the standard uniform distribution are $\mathbb{E}[\mathcal{X}] = 1/2$, $\text{Var}(\mathcal{X}) = 1/12$. A nice property of the uniform random variable is that we can derive any other continuous random variable by the application of a sufficient transformation to the uniform random variable.

(2) **The Exponential Distribution.** We write $\mathcal{X} \sim \mathsf{Exponential}(\lambda)$ or simply $\mathcal{X} \sim \mathsf{Exp}(\lambda)$ (where λ is a parameter called the rate parameter) to say that $\mathcal{X}$ is an exponential random variable. This distribution typically models waiting times (*How long do I need to wait until the next train arrives?*), in which case we might think of λ as the average number of arrivals per unit time. (Contrast this with $\mathsf{Poisson}(\lambda)$, which models the *number of arrivals* in a unit of time.) We can also use the exponential random variable to model the amount of time until a piece of equipment breaks down, or a light bulb goes out, or until an accident occurs.
The PDF of an exponential random variable is

$$f(x) = \begin{cases} \lambda e^{-\lambda x} & \text{if } x \geq 0 \\ 0 & \text{if } x < 0. \end{cases}$$

It is not hard to check that this is a valid PDF, since $f \geq 0$ and $\displaystyle\int_{-\infty}^{\infty} f(x)dx = \int_0^{\infty} \lambda e^{-\lambda x}dx =$ 1. The CDF is derived directly from this PDF as $F(x) = \mathbb{P}(\mathcal{X} \leq x) = \displaystyle\int_{-\infty}^{x} f(t)dt$. If $x < 0$, then $F(x) = 0$. If $x \geq 0$, then

$$F(x) = \int_{-\infty}^{x} f(t)dt = \int_{-\infty}^{x} \lambda e^{-\lambda t}dt = 1 - e^{-\lambda x}.$$

Thus the CDF is

$$F(x) = \begin{cases} 0 & \text{if } x < 0 \\ 1 - e^{-\lambda x} & \text{if } x \geq 0. \end{cases}$$

It may also be of interest to consider the *tail CDF*: for $x > 0$,

$$\mathbb{P}(\mathcal{X} > x) = 1 - \mathbb{P}(\mathcal{X} \leq x) = 1 - F(x) = e^{-\lambda x}.$$

Memoryless property of $\mathsf{Exponential}(\lambda)$. For $t, s > 0$:

$$\mathbb{P}(\mathcal{X} > t + s \mid \mathcal{X} > s) = \mathbb{P}(\mathcal{X} > t)$$

$\mathbb{P}$(wait an additional t units of time | already waited s units of time) = $\mathbb{P}$(wait t units of time).

Proof. We have that

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

Thus waiting for even longer after you have already waited for long is but an instance of the sunk-cost fallacy. ■

It is not hard to compute the mean and variance of this distribution. We can check that

$$\mathbb{E}[\mathcal{X}] = \int_0^{\infty} x\lambda e^{-\lambda x}dx = (-xe^{-\lambda x})|_0^{\infty} + \int_0^{\infty} e^{-\lambda x} = \frac{1}{\lambda},$$

and that

$$\text{Var}(\mathcal{X}) = \mathbb{E}[\mathcal{X}^2] - \mathbb{E}[\mathcal{X}]^2 = \int_0^{\infty} x^2\lambda e^{-\lambda x}dx - \int_0^{\infty} x\lambda e^{-\lambda x}dx = \frac{2}{\lambda}\,\mathbb{E}[\mathcal{X}] = \frac{2}{\lambda^2}.$$

16 Transformations and the normal distribution (15/11)

16.1 Transformations of continuous random variables

Given a continuous $\mathcal{X}$ with PDF and CDF $f_{\mathcal{X}}$ and $F_{\mathcal{X}}$, and given $\mathcal{Y} = g(\mathcal{X})$ for some function $g : \mathbb{R} \to \mathbb{R}$, how do we find the PDF or CDF of $\mathcal{Y}$?
Rule of thumb. Always start with the CDF (and not the PDF) of $\mathcal{X}$, because $F_{\mathcal{X}}(x) = \mathbb{P}(\mathcal{X} \leq x)$ is a probability, while the PDF is not.
Strategy. The CDF of $\mathcal{Y}$ is computed using $F_{\mathcal{Y}} = \mathbb{P}(\mathcal{Y} \leq y) = \mathbb{P}(g(\mathcal{X}) \leq y$ for some given $y \in \mathbb{R}$.

Example 16.1. Let $\mathcal{X} \sim \mathsf{Unif}[0, 1]$ and $\mathcal{Y} = 2\mathcal{X}$. Intuitively, $\mathcal{Y}$ should be $\mathsf{Unif}[0, 2]$, since we are but doubling a uniform region on the real line. Mathematically, recall that the PDF and CDF of $\mathcal{X}$ are

$$f_{\mathcal{X}}(x) = \begin{cases} 1 & \text{if } 0 \leq x \leq 1 \\ 0 & \text{otherwise.} \end{cases} \qquad F_{\mathcal{X}}(x) = \begin{cases} 0 & \text{if } x < 0 \\ x & \text{if } 0 \leq x \leq 1 \\ 1 & \text{if } x > 1. \end{cases}$$

Suppose now that we wished to find the CDF of $\mathcal{Y} = 2\mathcal{X}$. For $y \in \mathbb{R}$,

$$F_{\mathcal{Y}}(y) = \mathbb{P}(\mathcal{Y} \leq y) = \mathbb{P}(2\mathcal{X} \leq y) = \mathbb{P}(\mathcal{X} \leq y/2) = F_{\mathcal{X}}\left(\frac{y}{2}\right) = \begin{cases} 0 & \text{if } y/2 < 0, \\ y/2 & \text{if } 0 \leq y/2 \leq 1 \\ 1 & \text{if } y/2 > 1 \end{cases}$$
$$= \begin{cases} 0 & \text{if } y < 0, \\ y/2 & \text{if } 0 \leq y \leq 2 \\ 1 & \text{if } y > 2. \end{cases}$$

To get the PDF of $\mathcal{Y}$, we only need differentiate this form of the CDF:

$$f_{\mathcal{Y}}(y) = F'_{\mathcal{Y}}(y) = \begin{cases} 1/2 & \text{if } 0 \leq y \leq 2 \\ 0 & \text{otherwise.} \end{cases}$$

Example 16.2. Let $\mathcal{X}$ be an exponential random variable with rate parameter 1, i.e. $\mathcal{X} \sim \mathsf{Exp}(1)$. Then the PDF and CDF of $\mathcal{X}$ are respectively

$$f_{\mathcal{X}}(x) = \begin{cases} e^{-x} & \text{if } x \geq 0 \\ 0 & \text{if } x < 0. \end{cases} \qquad F_{\mathcal{X}}(x) = \begin{cases} 1 - e^{-x} & \text{if } x \geq 0 \\ 0 & \text{if } x < 0 \end{cases}$$

Let $\mathcal{Y} = \mathcal{X}^2$ and suppose we wanted to find the CDF and PDF of $\mathcal{Y}$. Note that for $y \in \mathbb{R}$, $F_{\mathcal{Y}}(y) = \mathbb{P}(\mathcal{Y} \leq y) = \mathbb{P}(\mathcal{X}^2 \leq y)$, and that for $y < 0$, this CDF is 0 since $\mathcal{X}^2$ is necessarily non-negative. For $y \geq 0$, the CDF is

$$\mathbb{P}(-\sqrt{y} \leq \mathcal{X} \leq \sqrt{y}) = F_{\mathcal{X}}(\sqrt{y}) - F_{\mathcal{X}}(-\sqrt{y}) = 1 - e^{-\sqrt{y}} - 0 = 1 - e^{-\sqrt{y}}.$$

The PDF is but the derivative of this CDF, so that

$$f_{\mathcal{Y}}(y) F'_{\mathcal{Y}}(y) = \frac{d}{dy}(1 - e^{-\sqrt{y}}) = \frac{1}{2\sqrt{y}} e^{-\sqrt{y}} \quad \text{for } y > 0.$$

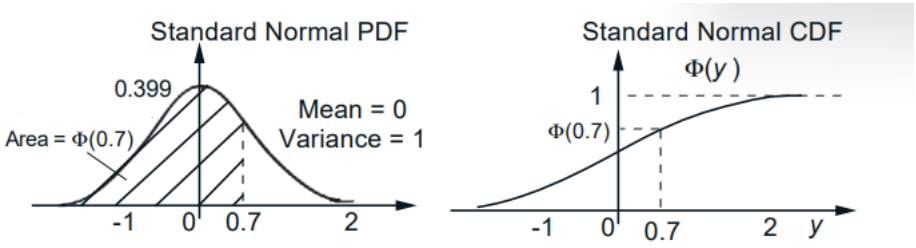
In summary then, we have

$$F_{\mathcal{Y}}(y) = \begin{cases} 1 - e^{-y} & \text{if } y \geq 0 \\ 0 & \text{otherwise.} \end{cases} \qquad f_{\mathcal{Y}}(y) = \begin{cases} \dfrac{1}{2\sqrt{y}} e^{-\sqrt{y}} & \text{if } y > 0 \\ 0 & \text{if } y \leq 0. \end{cases}$$

16.2 The Normal (Gaussian) Distribution

The Standard Normal Distribution. We say that $\mathcal{X}$ is a standard normal random variable by writing $\mathcal{X} \sim \mathcal{N}(0, 1)$. The PDF of the standard normal is

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2} \quad \text{for } x \in \mathbb{R}.$$



Theorem 16.1. *The PDF of the normal distribution is valid, i.e.* $\int_{-\infty}^{\infty} f(x)dx = 1$.

Proof. The following proof is attributed to Gauss. Complete the square thus:

$$\begin{aligned}
\left(\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx\right)^2 &= \left(\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx\right)\left(\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-y^2/2} dy\right) \\
&= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-x^2/2} \cdot \frac{1}{\sqrt{2\pi}} e^{-y^2/2} dxdy \\
&= \frac{1}{2\pi}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{-(x^2+y^2)/2} dxdy \\
&= \frac{1}{2\pi}\int_{0}^{2\pi}\int_{0}^{\infty} e^{-r^2/2} r\,dr\,d\theta \qquad \text{(changing to polar coords.)} \\
&= 1
\end{aligned}$$

(In other words, $\int_{-\infty}^{\infty} e^{-x^2/2} dx = \sqrt{2\pi}$.) ■

The CDF of the standard normal is

$$F(x) = \mathbb{P}(\mathcal{X} \leq x) = \int_{-\infty}^{x} f(y)dy = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-y^2/2} dy.$$

Note that this form of the integral cannot be simplified any further. So think of this form of the CDF as a special function in and of itself.
The mean of the standard normal is computed as

$$\mathbb{E}[\mathcal{X}] = \int_{-\infty}^{\infty} x f(x)dx = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} x e^{-x^2/2} dx = 0,$$

since $\lim_{x\to\infty} e^{-x} = 0$. Consequently, the variance of the standard normal is

$$\mathrm{Var}(\mathcal{X}) = \mathbb{E}[\mathcal{X}^2] = \int_{-\infty}^{\infty} x^2 f(x)dx = \frac{1}{2\pi} \int_{-\infty}^{\infty} x^2 e^{-x^2/2} dx = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx = 1.$$

Thus for $\mathcal{X} \sim \mathcal{N}(0, 1)$, $\mathbb{E}[\mathcal{X}] = 0$ and $\mathrm{Var}(\mathcal{X}) = 1$.

The General Normal Distribution. Suppose we perform a linear scaling of the standard normal. In particular, let $\mu \in \mathbb{R}$ and $\sigma > 0$, and draw $\mathcal{X} \sim \mathcal{N}(0, 1)$. Let $\mathcal{Y} = \mu + \sigma\mathcal{X}$. Then $\mathcal{Y} \sim \mathcal{N}(\mu, \sigma^2)$. The PDF of $\mathcal{Y}$ is

$$f_{\mathcal{Y}}(y) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y-\mu)^2}{2\sigma^2}} \quad \text{for } y \in \mathbb{R}.$$

This (general) normal distribution $\mathcal{N}(\mu, \sigma^2)$ takes parameters μ called its *mean* and σ^2 called its *variance*. A quick jusification of these terms is as follows:

$$\mathbb{E}[\mathcal{Y}] = \mathbb{E}[\mu + \sigma\mathcal{X}] = \mu + \sigma\,\mathbb{E}[\mathcal{X}] = \mu, \text{ and } \mathrm{Var}(\mathcal{Y}) = \mathrm{Var}(\mu + \sigma\mathcal{X}) = \sigma^2 \mathrm{Var}(\mathcal{X}) = \sigma^2.$$

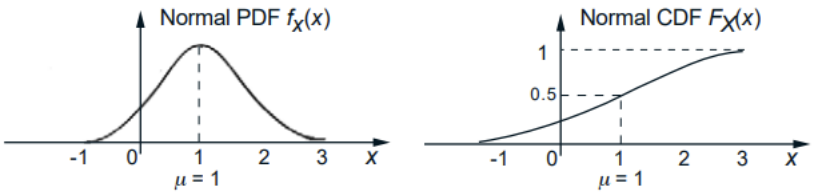
Note that we can derive the PDF of the normal distribution via transformation procedures. Start with the CDF of $\mathcal{Y}$: for $y \in \mathbb{R}$:

$$F_{\mathcal{Y}}(y) = \mathbb{P}(\mathcal{Y} \leq y) = \mathbb{P}(\mu + \sigma\mathcal{X} \leq y) = \mathbb{P}\left(\mathcal{X} \leq \frac{y-\mu}{\sigma}\right) = F_{\mathcal{X}}\left(\frac{y-\mu}{\sigma}\right)$$

All that remains is then to differentiate this term to get the PDF:

$$f_{\mathcal{Y}}(y) = F'_{\mathcal{Y}}(y) = \frac{d}{dy}\left(F_{\mathcal{X}}\left(\frac{y-\mu}{\sigma}\right)\right) = \frac{1}{\sigma} F'_{\mathcal{X}}\left(\frac{y-\mu}{\sigma}\right) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{y-\mu}{\sigma})^2},$$

as desired. The plot of the general normal PDF is as follows:



17 Joint probability density functions (17/11)

17.1 Continuous joint distribution

Definition 17.1. *A pair of random variables $(\mathcal{X}, \mathcal{Y})$ has a continuous joint distribution if there exists a non-negative function $f_{\mathcal{X},\mathcal{Y}} : \mathbb{R}^2 \to \mathbb{R}$ (called the joint PDF) such that*

$$\mathbb{P}((\mathcal{X}, \mathcal{Y}) \in A) = \iint\limits_{(x,y)\in A} f_{\mathcal{X},\mathcal{Y}}(x, y)dxdy,$$

for every subset $A \subset \mathbb{R}^2$.

If $a < b$ and $c < d$, then A is a rectangular region $[a, b] \times [c, d]$, we have

$$\mathbb{P}(a \leq \mathcal{X} \leq b, c \leq \mathcal{Y} \leq d) = \int_a^b \int_c^d f_{\mathcal{X},\mathcal{Y}}(x, y)dydx = \int_c^d \int_a^b f_{\mathcal{X},\mathcal{Y}}(x, y)dxdy.$$

If we set $A = \mathbb{R}^2$, then we get the standard normalisation for the joint PDF as follows:

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f_{\mathcal{X},\mathcal{Y}}(x, y)dxdy = 1.$$

What does the joint PDF mean? Suppose we have δ very small, and consider the probability contained within a very small triangle. Then

$$\mathbb{P}(a \leq \mathcal{X} \leq a + \delta,\ c \leq \mathcal{Y} \leq c + \delta) = \int_a^{a+\delta} \int_c^{c+\delta} f_{\mathcal{X},\mathcal{Y}}(x, y)dydx \approx \delta^2 \cdot f_{\mathcal{X},\mathcal{Y}}(a, c).$$

Thus we might interpret $f_{\mathcal{X},\mathcal{Y}}(a, c)$ as the probability per unit area near (a, c).
When we integrate over a region, we typically want to consider the region over which the values we are integrating is nonzero—the region where f "lives". To formalise this notion, we introduce the *support* S of a function f, defined as the region over which it is nonzero:

$$S \coloneqq \{(x, y) \in \mathbb{R}^2 \mid f(x, y) \neq 0\}.$$

The expectation of some function of these random variables also depends on the support S:

$$\mathbb{E}[g(\mathcal{X}, \mathcal{Y})] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x, y) f_{\mathcal{X},\mathcal{Y}}(x, y)dxdy.$$

Example 17.1. Suppose we have random variables $(\mathcal{X}, \mathcal{Y})$ with joint PDF

$$f(x, y) = \begin{cases} 2x & \text{if } 0 \leq x \leq 1, 0 \leq y \leq 1 \\ 0 & \text{otherwise.} \end{cases}$$

We can check that this is a valid joint PDF:

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x, y)dxdy = \int_0^1 \int_0^1 2xdxdy = \int_0^1 dy = 1.$$

Say we wanted to compute $\mathbb{P}(\mathcal{X} \le 1/2, \mathcal{Y} \ge 1/4)$. The region we are integrating over is the intersection of $A = \{(x, y) \in \mathbb{R}^2 \mid x \le 1/2, y \ge 1/4\}$ and the support of f. This region is

$$\mathbb{P}(\mathcal{X} \le 1/2, \mathcal{Y} \ge 1/4) = \int_{-\infty}^{\frac{1}{2}} \int_{\frac{1}{4}}^{\infty} f(x, y)\ dydx = \int_{-\infty}^{\frac{1}{2}} \int_{\frac{1}{4}}^{\infty} 2x\ dydx = \frac{3}{16}.$$

We could go further to compute $\mathbb{P}(\mathcal{Y} \le 2\mathcal{X})$. We need to first find the region in the support *and* where $y \le 2x$. We could compute this using two approaches, depending on how we choose the order of integration.

$$\mathbb{P}(\mathcal{Y} \le 2\mathcal{X}) = \int_0^1 \int_{\frac{y}{2}}^1 2x\ dxdy = \int_0^1 \left(1 - \frac{y^2}{4}\right)\ dy = \frac{11}{12},$$

or

$$\mathbb{P}(\mathcal{Y} \le 2\mathcal{X}) = \int_0^{\frac{1}{2}} \int_0^{2x} 2x\ dydx + \int_{\frac{1}{2}}^1 \int_0^1 2x\ dydx = \int_0^{\frac{1}{2}} 4x^2\ dx + \int_{\frac{1}{2}}^1 2x\ dx = \frac{11}{12}.$$

Finally, we compute $\mathbb{E}[\mathcal{Y}/\mathcal{X}]$:

$$\mathbb{E}[\mathcal{Y}/\mathcal{X}] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \frac{y}{x} \cdot f(x, y)\ dxdy = \int_0^1 \int_0^1 2y\ dxdy = 1.$$

17.2 Marginal PDFs

Let $(\mathcal{X}, \mathcal{Y})$ be random variables with joint PDF $f_{\mathcal{X},\mathcal{Y}}$. The marginal—individual—PDFs of $\mathcal{X}$ and $\mathcal{Y}$ are then

$$f_{\mathcal{X}}(x) = \int_{-\infty}^{\infty} f_{\mathcal{X},\mathcal{Y}}(x, y)\ dy \qquad \text{for each } x \in \mathbb{R}$$
$$f_{\mathcal{Y}}(y) = \int_{-\infty}^{\infty} f_{\mathcal{X},\mathcal{Y}}(x, y)\ dx \qquad \text{for each } y \in \mathbb{R}$$

Definition 17.2 (Independence)**.** *We say that $\mathcal{X}$ and $\mathcal{Y}$ are independent if $f_{\mathcal{X},\mathcal{Y}}(x, y) = f_{\mathcal{X}}(x) f_{\mathcal{Y}}(y)$ for all $x, y \in \mathbb{R}$.*

Consequences of independence. The following statements follow immediately from the definition of independence:

- For random variables $(\mathcal{X}, \mathcal{Y})$, $\mathbb{P}(a \le \mathcal{X} \le b, c \le \mathcal{Y} \le d) = \mathbb{P}(a \le \mathcal{X} \le b)\mathbb{P}(c \le \mathcal{Y} \le d)$.
- For two real-valued functions g, h, $\mathbb{E}[g(\mathcal{X})h(\mathcal{Y})] = \mathbb{E}[g(\mathcal{X})]\,\mathbb{E}[h(\mathcal{Y})]$.
- For random variables $(\mathcal{X}, \mathcal{Y})$, $\mathrm{Var}(\mathcal{X} + \mathcal{Y}) = \mathrm{Var}(\mathcal{X}) + \mathrm{Var}(\mathcal{Y})$.

Example 17.2. Recall Example 17.1. The marginals can be computed from the joint

distribution as

$$f_{\mathcal{X}}(x) = \begin{cases} 2x & \text{if } 0 \le x \le 1, \\ 0 & \text{otherwise,} \end{cases} \qquad f_{\mathcal{Y}}(y) = \begin{cases} 1 & \text{if } 0 \le y \le 1, \\ 0 & \text{otherwise.} \end{cases}$$

Note that $\mathcal{Y} \sim \mathsf{Unif}[0,1]$, and that $\mathcal{X}$ and $\mathcal{Y}$ are independent.

In general, given $f_{\mathcal{X},\mathcal{Y}}(x,y)$, to check independence, we need to check that both of the following are true:

(a) We can express the joint PDF as a combination of the function of $f_{\mathcal{X}}$ on the support and the function of $f_{\mathcal{Y}}$ on the support.

(b) The support an axis-aligned rectangle.

Example 17.3. The joint PDF defined by, for $c \in \mathbb{R}$,

$$f_{\mathcal{X},\mathcal{Y}}(x,y) = \begin{cases} \frac{1}{c}\cos(x+y) & 0 \le x \le \pi/4, 0 \le y \le \pi/4 \\ 0 & \text{otherwise} \end{cases}$$

is not independent because although it satisfies (b), it doesn't satisfy (a).

Example 17.4. Let the random variables $(\mathcal{X},\mathcal{Y})$ have joint PDF

$$f_{\mathcal{X},\mathcal{Y}}(x,y) = \begin{cases} 8xy & \text{if } 0 \le x \le y \le 1, \\ 0 & \text{otherwise.} \end{cases}$$

Note that $\mathcal{X}$ and $\mathcal{Y}$ are not independent because the support is *not* rectangular (confirm by drawing a picture). Additionally, this PDF is valid:

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_{\mathcal{X},\mathcal{Y}}(x,y) = \int_0^1 \int_0^y 8xy \; dxdy = \int_0^1 4y^3 \; dxdy = 1.$$

Say we wanted to compute $\mathbb{P}(\mathcal{X}+\mathcal{Y} \le 1)$. This is

$$\mathbb{P}(\mathcal{X}+\mathcal{Y}=1) = \int_0^{\frac{1}{2}} \int_x^{1-x} 8xy \; dydx = \int_0^{\frac{1}{2}} 4x - 8x^2 \; dx = \frac{1}{6}.$$

The marginals of both the random variables are $\mathcal{X}$ and $\mathcal{Y}$ are

$$f_{\mathcal{X}}(x) = \int_{-\infty}^{\infty} f_{\mathcal{X},\mathcal{Y}}(x,y) \; dy = \int_x^1 8xy \; dy = 4x(1-x^2) \text{ for } 0 \le x \le 1$$

and

$$f_{\mathcal{Y}}(y) = \int_{-\infty}^{\infty} f_{\mathcal{X},\mathcal{Y}}(x,y) \; dx = \int_0^y 8xy \; dx = 4y^3 \text{ for } 0 \le y \le 1.$$

We see again that $\mathcal{X}$ and $\mathcal{Y}$ are not independent, because

$$f_{\mathcal{X}}(x)f_{\mathcal{Y}}(y) = \begin{cases} 16xy^3(1-x^2) & \text{if } 0 \le x,y \le 1, \\ 0 & \text{otherwise,} \end{cases}$$

is *not* equal to $f_{\mathcal{X},\mathcal{Y}}(x,y)$ for all (x,y).

18 Joint PDFs, conditioning, and independence (22/11)

Recap. Let $(\mathcal{X}, \mathcal{Y})$ be continuous random variables with joint PDF $f_{\mathcal{X},\mathcal{Y}}$. The following follow:

- $f_{\mathcal{X},\mathcal{Y}}$ is non-negative, and the total double integral over all possible $f_{\mathcal{X},\mathcal{Y}}$ values is 1.
- The support of $f_{\mathcal{X},\mathcal{Y}}$ is the region where it is nonzero, and $\mathbb{P}((\mathcal{X},\mathcal{Y}) \in A) = \iint_A f_{\mathcal{X},\mathcal{Y}}(x,y)\,dxdy$.
- The expectation is $\mathbb{E}[g(\mathcal{X},\mathcal{Y})] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(x,y) f_{\mathcal{X},\mathcal{Y}}(x,y)\,dxdy$.
- The marginal PMF of $\mathcal{X}$ (and, mutatis mutandis, of $\mathcal{Y}$) is $f_{\mathcal{X}}(x) = \int_{-\infty}^{\infty} f_{\mathcal{X},\mathcal{Y}}(x,y)\,dy$.
- $\mathcal{X}$ and $\mathcal{Y}$ are *independent* if $f_{\mathcal{X},\mathcal{Y}}(x,y) = f_{\mathcal{X}}(x) f_{\mathcal{Y}}(y)$ for all x, y.

Example 18.1. Let $(\mathcal{X}, \mathcal{Y})$ be a uniformly random point in the unit disk. We can obtain the joint PDF as

$$f_{\mathcal{X},\mathcal{Y}}(x,y) = \begin{cases} 1/\pi & \text{if } x^2+y^2 \le 1 \\ 0 & \text{otherwise.} \end{cases}$$

This is a valid PDF, because

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_{\mathcal{X},\mathcal{Y}}(x,y)\,dxdy = \iint_{x^2+y^2\le 1} \frac{1}{\pi}\,dxdy = \frac{1}{\pi}\cdot\pi = 1.$$

We can find the marginal of $\mathcal{Y}$. Note that $x^2+y^2 \le 1 \iff -\sqrt{1-y^2} \le x \le \sqrt{1-y^2}$ and that $f_{\mathcal{Y}} = 0$ for $y < -1$ or $y > 1$. For $-1 \le y \le 1$, we have that

$$f_{\mathcal{Y}}(y) = \int_{-\infty}^{\infty} f_{\mathcal{X},\mathcal{Y}}(x,y)\,dx = \int_{-\sqrt{1-y^2}}^{\sqrt{1-y^2}} \frac{1}{\pi}\,dx = \frac{2}{\pi}\sqrt{1-y^2},$$

and 0 otherwise.

Example 18.2. Let $\mathcal{X} \sim \mathsf{Unif}[0,1]$ and $\mathcal{X} \sim \mathsf{Exp}(\lambda)$, where $\lambda > 0$. Assume $\mathcal{X}$ and $\mathcal{Y}$ are independent. The marginals are

$$f_{\mathcal{X}}(x) = \begin{cases} 1 & \text{if } 0 \le x \le 1 \\ 0 & \text{otherwise} \end{cases} \qquad f_{\mathcal{Y}}(y) = \begin{cases} \lambda e^{-\lambda y} & \text{if } y \ge 0 \\ 0 & \text{otherwise.} \end{cases}$$

Since $\mathcal{X}$ and $\mathcal{Y}$ are independent, we have that the joint PDF is

$$f_{\mathcal{X},\mathcal{Y}}(x,y) = \begin{cases} \lambda e^{-\lambda y} & \text{if } 0 \le x \le 1, y \ge 0 \\ 0 & \text{otherwise.} \end{cases}$$

We can compute the expectation of the product of these random variables as

$$\mathbb{E}[\mathcal{X}\mathcal{Y}] = \mathbb{E}[\mathcal{X}]\,\mathbb{E}[\mathcal{Y}] = \frac{1}{2\lambda}.$$

Suppose we wanted to find $\mathbb{P}(\mathcal{X} \leq \mathcal{Y})$. This is exactly

$$\mathbb{P}(\mathcal{X} \leq \mathcal{Y}) = \int_0^1 \int_x^\infty \lambda e^{-\lambda y}\, dydx = \int_0^1 e^{-\lambda x}\, dx = (1 - e^{-\lambda})/\lambda$$

Define now a new random variable $\mathcal{Z} = \mathcal{Y}/\mathcal{X}$, and suppose we wanted to find the PDF of $\mathcal{Z}$. As usual, we start with the CDF $F_{\mathcal{Z}}(z) = \mathbb{P}(\mathcal{Z} \leq z)$ for every $z \in \mathbb{R}$, and differentiate to get the PDF $f_{\mathcal{Z}} = F_{\mathcal{Z}}$. The range of possible values is $\mathcal{Z} \geq 0$, thus $F_{\mathcal{Z}}(z) = 0$ for $z \leq 0$. For fixed $r > 0$, we have

$$F_{\mathcal{Z}}(z) = \mathbb{P}(\mathcal{Z} \leq z) = \mathbb{P}(\mathcal{Y}/\mathcal{X} \leq z) = \mathbb{P}(\mathcal{Y} \leq z\mathcal{X}) = \int_0^1 \int_0^{zx} \lambda e^{-\lambda y}\, dydx = 1 - (1 - e^{-\lambda z})/\lambda z.$$

Thus the CDF of $\mathcal{Z}$ is

$$F_{\mathcal{Z}}(z) = \begin{cases} 0 & \text{if } z \leq 0 \\ 1 - \frac{1}{\lambda z}(1 - e^{-\lambda z}) & \text{if } z > 0. \end{cases}$$

The PDF is then, for $z > 0$,

$$f_{\mathcal{Z}}(z) = \frac{d}{dz} F_{\mathcal{Z}}(z) = \frac{d}{dz}\left(1 - \frac{1}{\lambda z}(1 - e^{-\lambda z})\right) = \frac{1 - e^{-\lambda z}}{\lambda z^2} - \frac{e^{-\lambda z}}{z}$$

18.1 Conditioning with continuous random variables

Definition 18.1 (Conditional PDF)**.** *For continuous random variables $(\mathcal{X}, \mathcal{Y})$ with joint PDF $f_{\mathcal{X},\mathcal{Y}}$ and marginals $f_{\mathcal{X}}$ and $f_{\mathcal{Y}}$, the conditional probability density function of $\mathcal{X}$ given $\mathcal{Y}$ is*

$$f_{\mathcal{X}|\mathcal{Y}}(x \mid y) = \begin{cases} \dfrac{f_{\mathcal{X},\mathcal{Y}}(x, y)}{f_{\mathcal{Y}}(y)} & \text{if } f_{\mathcal{Y}}(y) > 0 \\ 0 & \text{otherwise.} \end{cases}$$

Note. The conditional PDF is not a probability: $f_{\mathcal{X}|\mathcal{Y}}(x \mid y) \neq \mathbb{P}(\mathcal{X} = x \mid \mathcal{Y} = y)$. However, we can show that (and this is left as an exercise) $f_{\mathcal{X}|\mathcal{Y}}(x \mid y) = \lim_{\varepsilon \to 0^+} \left(\frac{1}{\varepsilon}\mathbb{P}(x \leq \mathcal{X} \leq x + \varepsilon \mid y \leq \mathcal{Y} \leq y + \varepsilon)\right)$.[2]
How do we use this? For $a < b$ and $y \in \mathbb{R}$ with $f_{\mathcal{Y}}(y) \neq 0$, we have that

$$\mathbb{P}(a \leq \mathcal{X} \leq b \mid \mathcal{Y} = y) = \int_a^b f_{\mathcal{X}|\mathcal{Y}}(x \mid y)\, dx.$$

However, we must *only* interpret this as a definition of $\mathbb{P}(a \leq \mathcal{X} \leq b \mid \mathcal{Y} = y)$. The event $\{\mathcal{Y} = y\}$ has zero probability, so we cannot *really* condition on it. We should really define the conditional PDF as the probability conditioned on an infinitesimally tiny region, i.e.

$$\int_a^b f_{\mathcal{X}|\mathcal{Y}}(x \mid y) = \lim_{\varepsilon \to 0^+} \left(\frac{1}{\varepsilon}\mathbb{P}(x \leq \mathcal{X} \leq x + \varepsilon \mid y \leq \mathcal{Y} \leq y + \varepsilon)\right).$$

[2] Compare this to $f_{\mathcal{X}}(x) = \lim_{\varepsilon \to 0^+} \left(\frac{1}{\varepsilon}\mathbb{P}(x \leq \mathcal{X} \leq x + \varepsilon)\right)$.

Conditional expectation. Analogous to earlier definitions of the expectation, the conditional expectation is

$$\mathbb{E}[g(\mathcal{X}) \mid \mathcal{Y} = y] = \int_{-\infty}^{\infty} g(x) f_{\mathcal{X}|\mathcal{Y}}(x \mid y)\, dx \qquad \mathbb{E}[g(\mathcal{X}, \mathcal{Y}) \mid \mathcal{Y} = y] = \int_{-\infty}^{\infty} g(x, y) f_{\mathcal{X}|\mathcal{Y}}(x \mid y)\, dx.$$

Example 18.3. Revisit Example 18.1, where the joint PDF of $(\mathcal{X}, \mathcal{Y})$ and marginal PDF of $\mathcal{Y}$ are

$$f_{\mathcal{X},\mathcal{Y}}(x, y) = \begin{cases} 1/\pi & \text{if } x^2 + y^2 \le 1 \\ 0 & \text{otherwise.} \end{cases} \qquad f_{\mathcal{Y}}(y) = \begin{cases} \frac{2}{\pi}\sqrt{1 - y^2} & \text{if } -1 \le y \le 1 \\ 0 & \text{otherwise.} \end{cases}$$

Suppose we wanted to find the conditional probability density function of $\mathcal{X}$ given $\mathcal{Y}$. For the disk $x^2 + y^2 \le 1$, we have that

$$f_{\mathcal{X}|\mathcal{Y}}(x \mid y) = \frac{f_{\mathcal{X},\mathcal{Y}}(x, y)}{f_{\mathcal{Y}}(y)} = \frac{1/\pi}{2/\pi\sqrt{1 - y^2}} = \frac{1}{2\sqrt{1 - y^2}},$$

and $f_{\mathcal{X}|\mathcal{Y}}(x \mid y) = 0$ if $x^2 + y^2 > 1$. Thus for a *fixed* $-1 < y < 1$, the conditional PDF is

$$f_{\mathcal{X}|\mathcal{Y}}(x \mid y) = \begin{cases} \dfrac{1}{2\sqrt{1 - y^2}} & \text{if } -\sqrt{1 - y^2} \le x \le \sqrt{1 - y^2} \\ 0 & \text{otherwise.} \end{cases}$$

Note that this is precisely $\mathsf{Unif}\left[-\sqrt{1 - y^2}, \sqrt{1 - y^2}\right]$! This observation simplifies expectation and variance calculations quite a bit. The conditional expectation is

$$\mathbb{E}[\mathcal{X} \mid \mathcal{Y} = y] = \mathbb{E}\left[\mathsf{Unif}\left[-\sqrt{1 - y^2}, \sqrt{1 - y^2}\right]\right] = 0,$$

since the expectation is merely the midpoint of the uniform interval. The conditional variance is

$$\mathrm{Var}(\mathcal{X} \mid \mathcal{Y} = y) = \mathrm{Var}\left(\mathsf{Unif}\left[-\sqrt{1 - y^2}, \sqrt{1 - y^2}\right]\right) = \frac{\left(2\sqrt{1 - y^2}\right)^2}{12} = \frac{1 - y^2}{3}.$$

As an aside, note that the random variables $\mathcal{X}$ and $\mathcal{Y}$ are *not* independent, since the range of possible $\mathcal{X}$ values depends on which $\mathcal{Y}$ value we are given. (Independence would have meant $f_{\mathcal{X}|\mathcal{Y}}(x \mid y) = f_{\mathcal{X}}(x)$.)

Law of conditional variances. For random variables $\mathcal{X}$ and $\mathcal{Y}$,

$$\mathrm{Var}(\mathcal{X}) = \mathbb{E}[\mathrm{Var}(\mathcal{X} \mid \mathcal{Y})] + \mathrm{Var}(\mathbb{E}[\mathcal{X} \mid \mathcal{Y}]).$$

Bayes' Rule in the Continuous Setting. In many situations, we have a model of an underlying unobserved phenomenon, represented by a random variable $\mathcal{X}$ with PDF $f_{\mathcal{X}}$, and we make noisy measurements Y. These measurements provide information about $\mathcal{X}$ and are modelled in terms of a conditional PDF $f_{\mathcal{Y}|\mathcal{X}}$. For example, if $\mathcal{Y}$ is the same as $\mathcal{X}$, but corrupted by zero-mean normally distributed noise, one would let the conditional PDF $f_{\mathcal{Y}|\mathcal{X}}(y \mid x)$ of $\mathcal{Y}$, given that $\{\mathcal{X} = x\}$, be normal with mean equal to x. Once the experimental value of $\mathcal{Y}$ is measured, what information does this provide on the unknown value of $\mathcal{X}$?

Note that the information provided by the event $\{\mathcal{Y} = y\}$ is described by the conditional PDF $f_{\mathcal{X}|\mathcal{Y}}(x \mid y)$. It thus suffices to evaluate the latter PDF. By a quick calculation, we get that

$$f_{\mathcal{X}|\mathcal{Y}}(x \mid y) = \frac{f_{\mathcal{X}}(x) f_{\mathcal{Y}|\mathcal{X}}(y \mid x)}{f_{\mathcal{Y}}(y)} = \frac{f_{\mathcal{X}}(x) f_{\mathcal{Y}|\mathcal{X}}(y \mid x)}{\displaystyle\int f_{\mathcal{X}}(t) f_{\mathcal{Y}|\mathcal{X}}(y \mid t)\, dt}$$

19 IID random variables; Moment generating functions (29/11)

19.1 Independent and identically distributed RVs

Definition 19.1. *We say that random variables $\mathcal{X}_1, \mathcal{X}_2, \mathcal{X}_3, \ldots$ are iid (independent and identically distributed) if*

(1) They are independent*.*

- *If the random variables are discrete, then*
$$\mathbb{P}(\mathcal{X}_1 = x_1, \mathcal{X}_2 = x_2, \ldots, \mathcal{X}_n = x_n) = \mathbb{P}(\mathcal{X}_1 = x_1)\mathbb{P}(\mathcal{X}_2 = x_2) \cdots \mathbb{P}(\mathcal{X}_n = x_n).$$
- *If the random variables are continuous, then*
$$f_{\mathcal{X}_1, \mathcal{X}_2, \ldots, \mathcal{X}_n}(x_1, x_2, \ldots, x_n) = f_{\mathcal{X}_1}(x_1) f_{\mathcal{X}_2}(x_2) \cdots f_{\mathcal{X}_n}(x_n).$$

(2) They are identically distributed*, meaning that each $\mathcal{X}_i$ has the same marginal distribution.*

- *If the random variables are discrete, then*
$$\mathbb{P}(\mathcal{X}_1 = x_1) = \mathbb{P}(\mathcal{X}_2 = x_2) = \ldots = \mathbb{P}(\mathcal{X}_n = x_n) \quad \textit{for each } x \in \mathbb{R}..$$
- *If the random variables are continuous, then*
$$f_{\mathcal{X}_1}(x) = f_{\mathcal{X}_2}(x) = \ldots = f_{\mathcal{X}_n}(x) \quad \textit{for each } x \in \mathbb{R}.$$

Example 19.1. Flip n coins. Let

$$\mathcal{X}_i = \begin{cases} 1 & \text{if the } i\text{th flip is heads} \\ 0 & \text{otherwise.} \end{cases}$$

Then $\mathcal{X}_1, \mathcal{X}_2, \ldots, \mathcal{X}_n$ are iid $\sim \mathsf{Bernoulli}(\frac{1}{2})$. Suppose we wanted to find the distribution of the maximum of n continuous random variables $\mathcal{X}_1, \mathcal{X}_2, \ldots, \mathcal{X}_n$. Let $F_{\mathcal{X}}$ be the CDF for each $\mathcal{X}_i$, $f_{\mathcal{X}}$ be the PDF for each $\mathcal{X}_i$, and $\mathcal{Y} = \max(\mathcal{X}_1, \mathcal{X}_2, \ldots, \mathcal{X}_n)$. We want to find the PDF of $\mathcal{Y}$. For $y \in \mathbb{R}$,

$$\begin{aligned} F_{\mathcal{Y}}(y) = \mathbb{P}(\mathcal{Y} \le y) &= \mathbb{P}(\max(\mathcal{X}_1, \mathcal{X}_2, \ldots, \mathcal{X}_n) \le y) = \mathbb{P}(\mathcal{X}_1 \le y, \mathcal{X}_2 \le y, \ldots, \mathcal{X}_n \le y) \\ &= \mathbb{P}(\mathcal{X}_1 \le y)\mathbb{P}(\mathcal{X}_2 \le y) \ldots \mathbb{P}(\mathcal{X}_n \le y) = F_{\mathcal{X}}(y) \cdot F_{\mathcal{X}}(y) \cdots F_{\mathcal{X}}(y) = (F_{\mathcal{X}}(y))^n \end{aligned}$$

We can then obtain the PDF of $\mathcal{Y}$ by differentiating:

$$f_{\mathcal{Y}}(y) = \frac{d}{dy} F_{\mathcal{Y}}(y) = \frac{d}{dy} (F_{\mathcal{X}}(y))^n = n(F_{\mathcal{X}}(y))^{n-1} \cdot \frac{d}{dy}(F_{\mathcal{X}}(y)) = n f_{\mathcal{X}}(y) F_{\mathcal{X}}(y)^{n-1}.$$

Example 19.2. Consider the exact same problem in Example 19.1, but with the random variable $\mathcal{Z} = \min(\mathcal{X}_1, \mathcal{X}_2, \ldots, \mathcal{X}_n)$. For $z \in \mathbb{R}$ given, we have

$$F_{\mathcal{Z}} = \mathbb{P}(\mathcal{Z} \leq z) = \mathbb{P}(\max(\mathcal{X}_1, \mathcal{X}_2, \ldots, \mathcal{X}_n) \leq z) = \mathbb{P}(\mathcal{X}_1 \leq z \text{ or } \mathcal{X}_2 \leq z \text{ or } \ldots \text{ or } \mathcal{X}_n \leq z)$$

However, we cannot simplify this further, because the inequalities are not disjoint. We work instead with the complement:

$$\begin{aligned} 1 - F_{\mathcal{Z}}(z) = 1 - \mathbb{P}(\mathcal{Z} \leq z) = \mathbb{P}(\mathcal{Z} > z) = \mathbb{P}(\min(\mathcal{X}_1, \mathcal{X}_2, \ldots, \mathcal{X}_n) > z) \\ = \mathbb{P}(\mathcal{X}_1 > z)\mathbb{P}(\mathcal{X}_2 > z) \cdots \mathbb{P}(\mathcal{X}_n > z) = (1 - F_{\mathcal{X}}(z))(1 - F_{\mathcal{X}}(z)) \cdots (1 - F_{\mathcal{X}}(z)) \\ = (1 - F_{\mathcal{X}}(z))^n \end{aligned}$$

Thus we get $1 - F_{\mathcal{Z}}(z) = (1 - F_{\mathcal{X}}(z))^n \implies F_{\mathcal{Z}}(z) = 1 - (1 - F_{\mathcal{X}}(z))^n$. We can then differentiate this to get the PDF:

$$f_{\mathcal{Z}}(z) = \frac{d}{dz} F_{\mathcal{Z}}(z) = \frac{d}{dz} \left(1 - (1 - F_{\mathcal{X}}(z))^n\right) = n f_{\mathcal{X}}(z)(1 - F_{\mathcal{X}}(z))^{n-1}.$$

Example 19.3. Let $\mathcal{X}_1, \mathcal{X}_2, \ldots, \mathcal{X}_n$ be iid $\sim \mathsf{Exp}(\lambda)$ for $\lambda > 0$ given. Here $\mathcal{X}_i$ could model the number of minutes until the next i train arrives, where λ is the arrival rate (in particular, the number of trains per minute on average). Then the random variable $\mathcal{Z} = \min(\mathcal{X}_1, \mathcal{X}_2, \ldots, \mathcal{X}_n)$ models the time until the next train of any kind arrives.
Recall that the PDF, CDF and tail CDF of the $\mathsf{Exp}(\lambda)$ random variable are

$$f_{\mathcal{X}}(x) = \begin{cases} \lambda e^{-\lambda x} & \text{if } x \geq 0 \\ 0 & \text{if } x < 0, \end{cases} \quad F_{\mathcal{X}}(x) = \begin{cases} 1 - e^{-\lambda x} & \text{if } x \geq 0 \\ 0 & \text{if } x < 0, \end{cases} \quad 1 - F_{\mathcal{X}}(x) = \begin{cases} e^{-\lambda x} & \text{if } x \geq 0 \\ 1 & \text{if } x < 0. \end{cases}$$

So using the same method developed in Example 19.2, we find that

$$1 - F_{\mathcal{Z}}(z) = (1 - F_{\mathcal{X}}(x))^n = \begin{cases} e^{-\lambda n x} & \text{if } x \geq 0 \\ 1 & \text{if } x < n \end{cases}$$

Rearranging and differentiating to get the PDF, the CDF and PDF of $\mathcal{Z}$ are as follows:

$$F_{\mathcal{Z}}(z) = \begin{cases} 1 - e^{-\lambda n z} & \text{if } z \geq 0 \\ 0 & \text{if } z < 0. \end{cases} \qquad f_{\mathcal{Z}}(z) = \begin{cases} \lambda n e^{-\lambda n r} & \text{if } z \geq 0 \\ 0 & \text{if } r < 0 \end{cases}$$

Note that $\mathcal{Z} \sim \mathsf{Exp}(\lambda n)$!
Intuition for this. Suppose there are λ trains arriving each minute of each train type, and that there are n train types. Then there are λn trains total per minute.

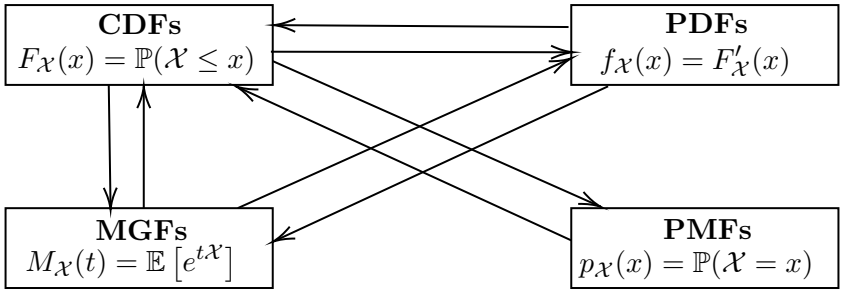
Example 19.4. Let $\mathcal{X}_i \sim \mathsf{Exp}(\lambda_i)$ be independent random variables with $\lambda_i > 0$ distinct. Then

$$\min(\mathcal{X}_1, \mathcal{X}_2, \ldots, \mathcal{X}_n) \sim \mathsf{Exp}(\lambda_1 + \lambda_2 + \ldots + \lambda_n).$$

This is not hard to show; the proof is left as an exercise for the reader.

19.2 Moment generating functions

Moment generating functions (MGFs), also known as *transformers*, are yet another function that "characterises" a random variable. We have already seen a few such functions:



Definition 19.2. *The moment generating function of a random variable $\mathcal{X}$ is the function $M_{\mathcal{X}} : \mathbb{R} \to \mathbb{R}$ is defined by*

$$M_{\mathcal{X}}(t) = \mathbb{E}[e^{t\mathcal{X}}] \quad \text{for } t \in \mathbb{R}.$$

Example 19.5. If $\mathcal{X}$ is a continuous random variable with probability density function $f_{\mathcal{X}}(x)$,

$$M_{\mathcal{X}}(t) = \int_{-\infty}^{\infty} e^{tx} f_{\mathcal{X}}(x)\, dx.$$

Example 19.6. If $\mathcal{X}$ is a discrete random variable with probability mass function $p_{\mathcal{X}}(x)$,

$$M_{\mathcal{X}}(t) = \sum_{k} e^{tk} \cdot \mathbb{P}(\mathcal{X} = k).$$

Note. $M_{\mathcal{X}}(0) = \mathbb{E}[e^0] = 1$, and $M_{\mathcal{X}}(t) > 0$ for all t. Also $M''_{\mathcal{X}}(t) > 0$—that is, the function $M_{\mathcal{X}}$ is convex—but this is harder to show.

Definition 19.3. *For integers $n \geq 0$, the nth moment of a random variable $\mathcal{X}$ is equal to $\mathbb{E}[\mathcal{X}^n]$.*

Theorem 19.4. *The nth moment of a random variable is the derivative of its moment generating function with respect to t, evaluated at $t = 0$, that is, $\dfrac{d^n M_{\mathcal{X}}}{dt^n}(0) = \mathbb{E}[\mathcal{X}^n]$.*

Proof sketch, discrete case, $n = 1$. We have that

$$\frac{d}{dt} M_{\mathcal{X}}(t) = \frac{d}{dt} \sum_{k} e^{tk} \cdot \mathbb{P}(\mathcal{X} = k) = \sum_{k} \frac{d}{dt} e^{tk} \cdot \mathbb{P}(\mathcal{X} = k) = \sum_{k} k e^{tk} \cdot \mathbb{P}(\mathcal{X} = k),$$

so that

$$\frac{dM_{\mathcal{X}}}{dt}(0) = \sum_{k} k \mathbb{P}(\mathcal{X} = k) = \mathbb{E}[\mathcal{X}],$$

like we would expect. ■

Example 19.7. Let $\mathcal{X} \sim \mathsf{Bernoulli}(p)$. The moment generating function is then

$$M_{\mathcal{X}}(t) = \mathbb{E}[e^{t\mathcal{X}}] = e^{t\cdot 1}\mathbb{P}(\mathcal{X}=1) + e^{t\cdot 0}\mathbb{P}(\mathcal{X}=0) = pe^t + 1 - p.$$

Note that because the $1-p$ term is constant, $\dfrac{d^n M_{\mathcal{X}}}{dt^n}(t) = pe^t$. Therefore, $\mathbb{E}[\mathcal{X}^n] = \dfrac{d^n M_{\mathcal{X}}}{dt^n}(0) = pe^0 = p$.

Example 19.8. Let $\mathcal{X} \sim \mathsf{Poisson}(\lambda)$, with $\lambda > 0$. Then the moment generating function is

$$M_{\mathcal{X}}(t) = \mathbb{E}[e^{t\mathcal{X}}] = \sum_{k=0}^{\infty} e^{tk}\cdot\mathbb{P}(\mathcal{X}=k) = \sum_{k=0}^{\infty} e^{tk}\left(e^{-\lambda}\frac{\lambda^k}{k!}\right) = e^{-\lambda}\sum_{k=0}^{\infty}\frac{(\lambda e^t)^k}{k!} = e^{-\lambda}e^{\lambda e^t} = e^{\lambda(e^t-1)}.$$

With this MGF, we can find the moment functions:

$$M'_{\mathcal{X}}(t) = e^{\lambda(e^t-1)}\cdot\lambda e^t = \lambda e^{t+\lambda(e^t-1)} \implies \mathbb{E}[\mathcal{X}] = M'_{\mathcal{X}}(0) = \lambda e^{0+\lambda(1-1)} = \lambda$$

$$M''_{\mathcal{X}}(t) = e^{t+\lambda(e^t-1)}\cdot(1+\lambda e^t) \implies \mathbb{E}[\mathcal{X}^2] = M''_{\mathcal{X}}(0) = \lambda(1+\lambda) = \lambda + \lambda^2$$

Therefore $\mathrm{Var}(\mathcal{X}) = \mathbb{E}[\mathcal{X}^2] - \mathbb{E}[\mathcal{X}]^2 = \lambda + \lambda^2 - \lambda^2 = \lambda$. Note that this is just like we saw earlier, but easier to compute.

20 MGFs and the Law of Large Numbers (01/12)

Recap. The moment generating function (MGF) of a random variable $\mathcal{X}$ is the function

$$M_{\mathcal{X}}(t) = \mathbb{E}[e^{t\mathcal{X}}] \quad \text{for } t \in \mathbb{R}.$$

The nth moment $\mathbb{E}[\mathcal{X}^n]$ is the nth derivative of the moment generating function evaluated at $t = 0$.
Note. Expectations can be infinite! (A famous example is the St. Petersburg paradox—see the Appendix for more.)

Example 20.1. Let $\mathcal{X} \sim \mathsf{Exp}(\lambda)$ with $\lambda > 0$. Recall that the PDF is

$$f_{\mathcal{X}}(x) = \begin{cases} \lambda e^{-\lambda x} & \text{if } x \geq 0 \\ 0 & \text{if } x < 0, \end{cases}$$

and the moment generating function is, for $t \in \mathbb{R}$,

$$M_{\mathcal{X}}(t) = \mathbb{E}[e^{t\mathcal{X}}] = \int_{-\infty}^{\infty} e^{t\mathcal{X}} f_{\mathcal{X}}(x)\,dx = \int_0^{\infty} e^{tx}\lambda e^{-\lambda x}\,dx = \lambda\int_0^{\infty} e^{(t-\lambda)x}\,dx = I.$$

We now proceed casewise.
Case 1. If $t = \lambda$, then $I = \lambda\displaystyle\int_0^{\infty} dx = \infty$.
Case 2. If $t > \lambda$, then $I = \infty$, obviously.
Case 3. If $t < \lambda$, $I = \lambda\displaystyle\int_0^{\infty} e^{(t-\lambda)x}\,dx = \frac{\lambda}{t-\lambda}(0 - e^0) = \frac{\lambda}{\lambda - t}$. Therefore, in summary, we

have

$$M_{\mathcal{X}}(t) = \begin{cases} \dfrac{\lambda}{\lambda - t} & \text{if } t < \lambda \\ \infty & \text{if } t \geq \lambda. \end{cases}$$

For $t < \lambda$, the first few moments of $\mathcal{X}$ are

$$\frac{dM_{\mathcal{X}}}{dt} = \frac{\lambda}{(\lambda - t)^2}, \quad \frac{d^2 M_{\mathcal{X}}}{dt^2} = \frac{2\lambda}{(\lambda - t)^3}, \quad \frac{d^3 M_{\mathcal{X}}}{dt^3} = \frac{3 \cdot 2 \cdot \lambda}{(\lambda - t)^4}, \quad \dots, \quad \frac{d^n M_{\mathcal{X}}}{dt^n} = \frac{n!\lambda}{(\lambda - t)^{n+1}}.$$

We can plug in $t = 0$ and obtain the nth moment of $\mathcal{X}$ as

$$\mathbb{E}[\mathcal{X}^n] = \frac{d^n M_{\mathcal{X}}}{dt^n}(0) = \frac{n!\lambda}{\lambda^{n+1}} = \frac{n!}{\lambda^n}$$

Thus $\mathbb{E}[\mathcal{X}] = 1/\lambda$, $\mathbb{E}[\mathcal{X}^2] = 2/\lambda^2$, and $\mathrm{Var}(\mathcal{X}) = \mathbb{E}[\mathcal{X}^2] - \mathbb{E}[\mathcal{X}]^2 = 2/\lambda^2 - 1/\lambda^2 = 1/\lambda^2$.

Example 20.2. Let $\mathcal{X} \sim \mathcal{N}(0, 1)$. Recall the PDF of the standard normal and the general normal is

$$f_{\mathcal{X}}(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}, \quad f_{\mathcal{X}}(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \quad \text{for } x \in \mathbb{R}.$$

Say we wanted to compute the MGF of the standard normal. Then for $t \in \mathbb{R}$,

$$M_{\mathcal{X}}(t) = \mathbb{E}[e^{t\mathcal{X}}] = \int_{-\infty}^{\infty} e^{tx} \cdot \frac{1}{\sqrt{2\pi}} e^{-x^2/2}\, dx.$$

With normals, we typically try to transform this integral until its PDF takes the form $\mathcal{N}(\mu, \sigma^2)$. To do this, we complete the square:

$$tx - \frac{1}{2}x^2 = -\frac{1}{2}(x - t)^2 + \frac{1}{2}t^2 \implies e^{tx - \frac{1}{2}x^2} = e^{t^2/2} e^{-\frac{1}{2}(x-t)^2}.$$

Therefore,

$$M_{\mathcal{X}}(t) = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{t^2/2} e^{-\frac{1}{2}(x-t)^2}\, dx = e^{t^2/2} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(x-t)^2}\, dx \overset{\text{(i)}}{=} e^{t^2/2},$$

where the equality (i) holds because the integral on the left hand side of (i) is exactly the total integral of the PDF of $\mathcal{N}(t, 1)$, so it equals 1 (as shown earlier). The first two moments of the standard normal are

$$M'_{\mathcal{X}}(t) = te^{t^2/2} \implies \mathbb{E}[\mathcal{X}] = M'_{\mathcal{X}}(0) = 0 \text{ and } M''_{\mathcal{X}}(t) = (t^2 + 1)e^{t^2/2}$$
$$\implies \mathbb{E}[\mathcal{X}^2] = M''_{\mathcal{X}}(0) = 1 = \mathrm{Var}(\mathcal{X}).$$

The following is a very important property of moment-generating functions.

Theorem 20.1 (Inversion Theorem)**.** *The moment generating function $M_{\mathcal{X}}(t)$ completely determines the distribution of the random variable $\mathcal{X}$. In particular, if $M_{\mathcal{X}}(t) = M_{\mathcal{Y}}(t) \iff \mathbb{E}[e^{t\mathcal{X}}] = \mathbb{E}[e^{t\mathcal{Y}}]$ for every $t \in \mathbb{R}$ (and if $M_{\mathcal{X}}(t) \ll \infty$ for all t near 0), then $\mathcal{X}$ and $\mathcal{Y}$ have the same distribution.*

Example 20.3. If we are told that a random variable $\mathcal{X}$ satisfies

$$\mathbb{E}[e^{t\mathcal{X}}] = e^{t^2/2} \text{ for all } t \in \mathbb{R},$$

then the inversion theorem implies that $\mathcal{X} \sim \mathcal{N}(0,1)$.

20.1 Law of large numbers (LLN)

Let $\mathcal{X}_1, \mathcal{X}_2, \mathcal{X}_3, \ldots$ be non-constant iid random variables. Let

$$m = \mathbb{E}[\mathcal{X}_1] = \mathbb{E}[\mathcal{X}_2] = \ldots \text{ and } \sigma^2 = \text{Var}(\mathcal{X}_1) = \text{Var}(\mathcal{X}_2) = \ldots > 0$$

Definition 20.2 (Sample average)**.** *The sample average of n terms is defined by the transformed random variable*[3] $\overline{\mathcal{X}_n} = \frac{1}{n}\sum_{i=1}^{n} \mathcal{X}_i$

Theorem 20.3 (Strong LLN)**.** *In the above setting,* $\mathbb{P}\left(\lim_{n\to\infty} \overline{\mathcal{X}_n} = m\right) = 1$.

What this theorem means is that as the number of samples n goes to infinity, the sample average converging to m (the true mean) is certain.

Example 20.4. Flip a lot of coins. The fraction of flips which are heads approaches 1/2.

Theorem 20.4 (Weak LLN)**.** *The weak law of large numbers states that for any prespecified error threshold $\varepsilon > 0$,* $\lim_{n\to\infty} \mathbb{P}\left(|\overline{\mathcal{X}_n} - m| > \varepsilon\right) = 0$.

What this theorem means is that as n becomes very large, the probability that $\overline{\mathcal{X}_n}$ deviates from m by at least ε becomes zero.

A few important facts. Recall that $m = \mathbb{E}[\mathcal{X}_i]$, $\sigma^2 = \text{Var}(\mathcal{X}_i)$, and $\overline{\mathcal{X}_n} = \frac{1}{n}\sum_{i=1}^{n} \mathcal{X}_i$. The following are true:

1. (Unbiased estimator.) $\mathbb{E}[\overline{\mathcal{X}_n}] = m$.

   Proof.

   $$\mathbb{E}[\overline{\mathcal{X}_n}] = \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} \mathcal{X}_i\right] \overset{\text{(i)}}{=} \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}[\mathcal{X}_i] = \frac{1}{n}\sum_{i=1}^{n} m = m,$$

   where (i) follows by linearity of expectation. ■

[3] It is typically very hard to compute the distribution of this transformed random variable.

2. $\mathrm{Var}(\overline{\mathcal{X}_n}) = \sigma^2/n$. Intuitively this should make sense, since for large n, $\mathrm{Var}(\overline{\mathcal{X}_n}) \to 0$ and constants have zero variance, so $\overline{\mathcal{X}_n}$ should be *almost* constant. In addition to this reasoning, a proof is provided:

   Proof.

$$\mathrm{Var}(\overline{\mathcal{X}_n}) = \mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^{n}\mathcal{X}_i\right) \overset{\text{(i)}}{=} \frac{1}{n^2}\mathrm{Var}\left(\sum_{i=1}^{n}\mathcal{X}_i\right) \overset{\text{(ii)}}{=} \frac{1}{n^2}\sum_{i=1}^{n}\mathrm{Var}\left(\mathcal{X}_i\right) = \frac{1}{n^2}\cdot\sum_{i=1}^{n}\sigma^2 = \frac{\sigma^2}{n},$$

   where (i) follows since $\mathrm{Var}(a\mathcal{X}) = a^2\mathrm{Var}(\mathcal{X})$, and (ii) follows since variance is linear when the random variables are independent (because then their covariance is 0). ■

3. (Markov's Inequality.) If $\mathcal{Y}$ is a nonnegative random variable and $c > 0$ is a constant, then $\mathbb{P}(\mathcal{Y} > c) \leq \frac{1}{c}\,\mathbb{E}[\mathcal{Y}]$. This inequality is useful, because it bounds a probability in terms of an expectation.

   Proof. Define an indicator function

$$f(x) = \begin{cases} 1 & \text{if } x > c \\ 0 & \text{if } x \leq c \end{cases}$$

   Then

$$\mathbb{E}[f(\mathcal{Y})] = \mathbb{E}[\mathbf{1}_{\{\mathcal{Y}>c\}}] = 1\cdot\mathbb{P}(f(\mathcal{Y})=1) + 0\cdot\mathbb{P}(f(\mathcal{Y})=0) = \mathbb{P}(\mathcal{Y}>c).$$

   Note that regardless of the (non-negative) value x has, $f(x) \leq x/c$. Thus

$$\mathbb{P}(\mathcal{Y} > c) = \mathbb{E}[f(\mathcal{Y})] \leq \mathbb{E}[\mathcal{Y}/c] = \frac{1}{c}\,\mathbb{E}[\mathcal{Y}],$$

   as desired. ■

4. (Chebyshev's Inequality.) If $\mathcal{Y}$ is a non-negative random variable with mean m and variance σ^2, then for $a > 0$ a constant, $\mathbb{P}(|\mathcal{X} - m| \geq a) \leq \frac{\sigma^2}{a^2}$. This inequality asserts that if the variance of a random variable is small, then the probability that it takes a value far from its mean is also small. The proof is not hard—consider the non-negative random variable $|\overline{\mathcal{X}_n} - m|^2$ and apply the Markov inequality with $c = a^2$.

Proof of weak LLN. Let $\varepsilon > 0$. Use Markov's inequality with $\mathcal{Y} = |\overline{\mathcal{X}_n} - m|^2$, and $c = \varepsilon^2$. Then

$$\begin{aligned}
\mathbb{P}\left(|\overline{\mathcal{X}_n} - m| > \varepsilon\right) &= \mathbb{P}\left(|\overline{\mathcal{X}_n} - m|^2 > \varepsilon^2\right) \\
&\leq \frac{1}{\varepsilon^2}\,\mathbb{E}[|\overline{\mathcal{X}_n} - m|^2] && \text{by fact (3), Markov's} \\
&= \frac{1}{\varepsilon^2}\,\mathbb{E}[(\overline{\mathcal{X}_n} - \mathbb{E}[\overline{\mathcal{X}_n}])^2] && \text{by fact (1)} \\
&= \frac{1}{\varepsilon^2}\mathrm{Var}(\overline{\mathcal{X}_n}) \\
&= \sigma^2/n\varepsilon^2 && \text{by fact (2),}
\end{aligned}$$

which goes to 0 as $n \to \infty$.

21 Central limit theorem, CLT (06/12)

Recap. Let $\mathcal{X}_1, \mathcal{X}_2, \ldots$ be iid random variables. Let m be the mean and σ^2 the variance. Assume $0 < \sigma^2 < \infty$.

We define the sample average as $\overline{\mathcal{X}_n} = \frac{1}{n}\sum_{i=1}^{n} \mathcal{X}_i$. The LLN states that if n is large, then $\overline{\mathcal{X}_n} \approx m$.

Take $\mathbb{E}[\overline{\mathcal{X}_n}] = m$ and $\operatorname{Var}(\overline{\mathcal{X}_n}) = \sigma^2/n$. Then for any error threshold $\varepsilon > 0$, $\mathbb{P}(|\overline{\mathcal{X}_n} - m| > \varepsilon) \leq \sigma^2/n\varepsilon^2$.

Informally, what is the shape of the "fluctuations" $\overline{\mathcal{X}_n} - m$? We can find that the answer to this question is roughly $\overline{\mathcal{X}_n} \approx m + \frac{\sigma}{\sqrt{n}}\mathcal{Z}$ where $\mathcal{Z} \sim \mathcal{N}(0,1)$. We can re-arrange this to get the form $\mathcal{Z} = \frac{\sqrt{n}}{\sigma}(\overline{\mathcal{X}_n} - m) \approx \mathcal{N}(0,1)$.

Theorem 21.1 (Central limit theorem)**.** *Let $\mathcal{X}_1, \mathcal{X}_2, \ldots$ be a sequence of independent and identically distributed random variables with common mean m and variance σ^2, and define*

$$\mathcal{Z}_n = \frac{\sqrt{n}}{\sigma}(\overline{\mathcal{X}_n} - m) = \frac{\mathcal{X}_1 + \mathcal{X}_2 + \ldots + \mathcal{X}_n - nm}{\sqrt{n\sigma^2}}.$$

Then the CDF of $\mathcal{Z}_n$ converges pointwise to the standard normal CDF

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

In particular,

$$\lim_{n\to\infty} \mathbb{P}(\overline{\mathcal{Z}_n} \leq x) = \mathbb{P}(\mathcal{Z} \leq x) \quad \text{for } \mathcal{Z} \sim \mathcal{N}(0,1).$$

Comments on the CLT. The CLT is brilliant. Here are some immediately apparent facts from the central limit theorem.

- The PDF of the distribution that $\overline{\mathcal{Z}_n}$ converges to is the PDF of the standard normal, namely $\frac{1}{\sqrt{2\pi}} e^{-x^2/2}$.
- (Universality.) We always get $\mathcal{Z} \sim \mathcal{N}(0,1)$, no matter what the underlying $\mathcal{X}_i$ distribution is![4]
- (Standard error.) $\overline{\mathcal{Z}_n}$ can be rewritten as $\overline{\mathcal{Z}_n} = \dfrac{\overline{\mathcal{X}_n} - \mathbb{E}[\mathcal{X}_n]}{\sigma_{\overline{\mathcal{X}_n}}}$

  Proof. Write $m = \mathbb{E}[\overline{\mathcal{X}_n}]$ and $\sigma_{\overline{\mathcal{X}_n}} = \sqrt{\operatorname{Var}(\overline{\mathcal{X}_n})} = \sigma/\sqrt{n}$. Then

  $$\overline{\mathcal{Z}_n} = \frac{\sqrt{n}}{\sigma}(\overline{\mathcal{X}_n} - m) = \frac{1}{\sigma_{\overline{\mathcal{X}_n}}}\left(\overline{\mathcal{X}_n} - \mathbb{E}[\overline{\mathcal{X}_n}]\right),$$

  as expected. ∎

[4] This fact can be demonstrated experimentally by the use of the Galton board.

- (Mean and variance of $\overline{\mathcal{Z}_n}$.) We have that

$$\mathbb{E}[\overline{\mathcal{Z}_n}] = \frac{\mathbb{E}[\overline{\mathcal{X}_n}] - \mathbb{E}[\overline{\mathcal{X}_n}]}{\sigma_{\overline{\mathcal{X}_n}}} = 0,$$

  by linearity of expectation, and that

$$\mathrm{Var}(\overline{\mathcal{Z}_n}) = \frac{1}{\sigma_{\overline{\mathcal{X}_n}}}\mathrm{Var}(\overline{\mathcal{X}_n}) = \frac{\mathrm{Var}(\overline{\mathcal{X}_n})}{\mathrm{Var}(\overline{\mathcal{X}_n})} = 1$$

- (Property of the normal distribution.) If $\mathcal{Z} \sim \mathcal{N}(0,1)$ and $\mu \in \mathbb{R}, \sigma > 0$ are constants, then by the transformation described earlier (section 16), $\mu + \sigma\mathcal{Z} \sim \mathcal{N}(\mu, \sigma^2)$. Thus the central limit theorem implies that

$$\begin{aligned}\overline{\mathcal{X}_n} &\approx m + \frac{\sigma}{\sqrt{n}}\mathcal{N}(0,1) \\ &\sim \mathcal{N}(m, \sigma^2/n).\end{aligned}$$

21.1 Practical uses of the CLT

Note the distinction between the following: $\overline{\mathcal{Z}_n} = \frac{\sqrt{n}}{\sigma}(\overline{\mathcal{X}_n} - m)$, and $\mathcal{Z} \sim \mathcal{N}(0,1)$.

To calculate approximate probabilities involving $\overline{\mathcal{X}_n} = \frac{1}{n}\sum_{i=1}^n \mathcal{X}_i$ or even $\sum_{i=1}^n \mathcal{X}_i$, we use the CLT approximation $\overline{\mathcal{X}_n} \approx m + (\sigma/\sqrt{n})\mathcal{Z}$ with $\mathcal{Z} \sim \mathcal{N}(0,1)$.

Example 21.1. Substituting the above approximation, we get that

$$\mathbb{P}(1 \le \overline{\mathcal{X}_n} \le 5) \approx \mathbb{P}(1 \le m + (\sigma/\sqrt{n})\mathcal{Z} \le 5) = \mathbb{P}(\frac{\sqrt{n}}{\sigma}(1-m) \le \mathcal{Z} \le \frac{\sqrt{n}}{\sigma}(5-m)).$$

This is useful because $\overline{\mathcal{X}_n}$ is usually too complicated to compute with, and $\mathcal{Z} \sim \mathcal{N}(0,1)$ is much easier to compute, despite the CDF of $\mathcal{Z}$ not having a closed form. Recall that this CDF of $\mathcal{Z}$ is

$$\mathbb{P}(\mathcal{Z} \le x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-t^2/2}\, dt.$$

This form of the integral is easily evaluated on a computer using numerical integration techniques. However, a few values are known: $\mathbb{P}(|\mathcal{Z}| \le 1) = \mathbb{P}(-1 \le \mathcal{Z} \le 1) \approx 0.68$, $\mathbb{P}(|\mathcal{Z}| \le 2) \approx 0.95$, and $\mathbb{P}(|\mathcal{Z}| \le 3) \approx 0.998$. We can also derive some of these values from already known values. For example, $\mathbb{P}(|\mathcal{Z}| > 2) = 1 - \mathbb{P}(|\mathcal{Z}| \le 2) \approx 1 - 0.95 \approx 0.05$. Also, by symmetry, $\mathbb{P}(\mathcal{Z} > 2) = \frac{1}{2}\mathbb{P}(|\mathcal{Z}| > 2) = 0.025$.

21.2 Applying the central limit theorem

The central limit theorem is often helpful for answering the question *How many samples do we need to achieve desired accuracy?*

Example 21.2. Flip n coins. We know that about a half of these should be heads. We want to flip enough coins to be 95% sure that the fraction of heads is within 0.05 of the expected 1/2. Suppose now we wanted to find the number of flips we need.
Let $\mathcal{X}_1, \mathcal{X}_2, \ldots$ be iid $\mathsf{Bernoulli}(\frac{1}{2})$, where $\mathcal{X}_i = 1$ means that the ith flip shows up heads. So $m = \mathbb{E}[\mathcal{X}_1] = 1/2$ and $\sigma^2 = \text{Var}(\mathcal{X}_1) = 1/4 \implies \sigma = 1/2$. Then the fraction of heads out of n flips is

$$\overline{\mathcal{X}_n} = \frac{1}{n}\sum_{i=1}^{n} \mathcal{X}_i.$$

We desire the n such that $\mathbb{P}\left(\left|\overline{\mathcal{X}_n} - \frac{1}{2}\right| \le 0.05\right) \ge 0.95$. By CLT then,

$$\mathbb{P}\left(\left|\overline{\mathcal{X}_n} - \frac{1}{2}\right| \le \frac{1}{20}\right) \approx \mathbb{P}\left(\left|\frac{1}{2} + \frac{1}{2\sqrt{n}} - \frac{1}{2}\right| \le \frac{1}{20}\right) = \mathbb{P}\left(|\mathcal{Z}| \le \frac{\sqrt{n}}{10}\right).$$

We want n large enough to make this ≥ 0.95, and we know that $\mathbb{P}(|\mathcal{Z}| \le 2) \approx 0.95$. So we just need $\sqrt{n}/10 \ge 2$, or $n \ge 400$.

Binomial approximation of the normal. It is true that $\mathsf{Binomial}(n, 1/2) \approx \mathcal{N}(n/2, n/4)$.
Why? If $\mathcal{X}_1, \mathcal{X}_2, \ldots$ are iid $\mathsf{Bernoulli}(\frac{1}{2})$, then $\sum_{i=1}^{n} \mathcal{X}_i \sim \mathsf{Binomial}(n, 1/2)$. By the CLT, $\frac{1}{n}\sum_{i=1}^{n} \mathcal{X}_i = \overline{\mathcal{X}_n} = m + (\sigma/\sqrt{n})\mathcal{Z} = 1/2 + (1/2\sqrt{n})\mathcal{Z}$, so if $\mathcal{Z} \sim \mathcal{N}(0,1)$

$$\sum_{i=1}^{n} \mathcal{X}_i \approx \frac{n}{2} + \frac{n}{2\sqrt{n}}\mathcal{Z} = \frac{n}{2} + \frac{\sqrt{n}}{2}\mathcal{Z} \sim \mathcal{N}(n/2, n/4),$$

since $\mu + \sigma\mathcal{Z} \sim \mathcal{N}(\mu, \sigma^2)$.

Example 21.3. Suppose I have a stock portfolio. Each day $i = 1, 2, 3, \ldots$, my stock portfolio earns not necessarily non-negative $\$\,\mathcal{X}_i$. Assume that the $\mathcal{X}_i$ are iid with mean $\$\,1$ and variance $\$\,16$. *How many days do I need to be 97.5% certain that I make a profit?*
My "net earnings" or "profit" after n days is $\sum_{i=1}^{n} \mathcal{X}_i$, and I want this to be positive, namely $\mathbb{P}\left(\sum_{i=1}^{n} \mathcal{X}_i > 0\right) \ge 0.975$. The question is then the same as *How large does n need to be?* Note that

$$\sum_{i=1}^{n} \mathcal{X}_i = n\left(\frac{1}{n}\sum_{i=1}^{n} \mathcal{X}_i\right) = n\overline{\mathcal{X}_n} \implies \mathbb{P}\left(\sum_{i=1}^{n} \mathcal{X}_i > 0\right) = \mathbb{P}(n\overline{\mathcal{X}_n} > 0) = \mathbb{P}(\overline{\mathcal{X}_n} > 0).$$

Note that since the variance is 16, the standard deviation is 4, so by CLT,

$$\mathbb{P}\left(\sum_{i=1}^{n} \mathcal{X}_i > 0\right) \approx \mathbb{P}\left(m + \frac{\sigma}{\sqrt{n}}\mathcal{Z} > 0\right) = \mathbb{P}\left(1 + \frac{4}{\sqrt{n}}\mathcal{Z} > 0\right) = \mathbb{P}\left(\mathcal{Z} > -\frac{\sqrt{n}}{4}\right).$$

We want this to be ≈ 0.975. We also know that $\mathbb{P}(\mathcal{Z} > -2) \approx 0.975$. Thus we should take

$$-\frac{\sqrt{n}}{4} \approx -2 \implies n^2 \approx 8^2 = 64 \text{ days}.$$

The De Moivre–Laplace Approximation. *If S_n is a binomial random variable with parameters n and p, n is large, and k, ℓ are nonnegative integers, then*

$$\mathbb{P}(k \leq S_n \leq \ell) \approx \Phi\left(\frac{\ell + \frac{1}{2} - np}{\sqrt{np(1-p)}}\right) - \Phi\left(\frac{k - \frac{1}{2} - np}{\sqrt{np(1-p)}}\right),$$

where Φ is the cumulative distribution function of the standard normal.

Appendix

Selected Recitation Problems

1. Suppose a bond A has a 50% default probability and bond B has a 30% default probability. What is the range of probability that at least one bond defaults? Give a range for their correlation, defined by

$$\text{Corr}(A,B) = \frac{\mathbb{P}(A \cup B) - \mathbb{P}(A)\,\mathbb{P}(B)}{\sqrt{\mathbb{P}(A)(1-\mathbb{P}(A))}\sqrt{\mathbb{P}(B)(1-\mathbb{P}(B))}}.$$

2. Let A and B be events with probabilities $\mathbb{P}(A) = 3/4$ and $\mathbb{P}(B) = 1/3$. Show that $1/12 \leq \mathbb{P}(A \cap B) \leq 1/3$, and give examples to demonstrate that both extremes are possible.

3. Suppose that a probability measure on the sample space $\Omega = \{1, 2, \ldots, n\}$ is such that the probability of the event $\{1, 2, \ldots, k\}$ is proportional to k^2. In other words, there exists $\alpha \in \mathbb{R}$ verifying $\mathbb{P}(\{1, 2, \ldots, k\}) = \alpha k^2$, for all $k \leq n$. Determine $\mathbb{P}(\{k\}$ for every $k \leq n$.

4. A point is chosen at random from a unit square $ABCD$. Find the probability that it is in triangle ABD given that it is in triangle ABC.

5. A professor frequently forgets his keys. For all n, let E_n denote the event that the professor forgets his keys on day n. Let $P_n = \mathbb{P}(E_n)$ and $Q_n = \mathbb{P}(E_n^c)$.
   Suppose that $P_1 = a$ is given and that if the professor forgets his keys on day n, then he forgets them the following day with probability 1/10; if he doesn't forget his keys on day n, then he forgets them the following day with probability 4/10.

   (i) Show that $P_{n+1} = \frac{1}{10}(P_n + 4Q_n)$.

   (ii) Write P_{n+1} in terms of P_n.

   (iii) Give the expression P_n in terms of n.

6. How might one generate even odds using a *biased* coin?

7. There is one amoeba in a pond. Every minute, it can die, stay the same, split into two, or split into three with equal probability. All offspring, if any, will behave the same. What is the probability that the population will die out?

8. Let $\Omega = \{1, 2, \ldots, p\}$, where p is a prime number. Suppose that in this setup, $\mathbb{P}(A) = |A|/p$ for every event A. Show that, if A and B are independent events then at least one of A and B is either $\emptyset$ or Ω.

9. We keep tossing an unfair coin. At each flip, the probability of the outcome being heads is 2/3. The flips are supposed to be independent, and $\mathcal{X}$ is the random variable equal to the number of necessary flips to get, for the first time, two consecutive heads. For $n \geq 1$, let $p_n = \mathbb{P}(\{\mathcal{X} = n\})$.

   (i) Write explicitly the events $\{\mathcal{X} = 2\}$, $\{\mathcal{X} = 3\}$, $\{\mathcal{X} = 4\}$, and determine the values of p_2, p_3, p_4.

   (ii) Show that $p_n = \frac{1}{9}(2p_{n-2} + 3p_{n-1})$, for $n \geq 4$.

(iii) Give the expression of p_n for all n. By calculating the derivative of $\sum_{n=0}^{\infty} q^n$, give the expression of $\sum_{n=0}^{\infty} nq^n$ for $-1 < q < 1$. Calculate the expectation $\mathbb{E}[\mathcal{X}]$.

10. We flip a fair coin n times. What is the probability of getting strictly more heads than tails?

11. Let $\mathcal{X}$ and $\mathcal{Y}$ be two Bernoulli independent random variables with parameters p and q respectively. Determine the distribution of the random variable $\mathcal{Z} = \max(\mathcal{X}, \mathcal{Y})$.

12. Two archers shoot independently at n targets. For each shot, the first archer hits the target with probability p, and the second archer with probability q. Let $\mathcal{N}$ be the random variable corresponding to the number of targets hit at least one time. Find the distribution of $\mathcal{N}$, as well as the number of missed targets.

13. Consider the binomial random variable $\mathcal{X}$ with parameters n and p. Show that as k increases, the PMF $p_{\mathcal{X}}(k)$ first increases monotonically and then decreases monotonically, and that the maximum is observed for k the largest integer less than or equal to $(n+1)p$.

14. Two gamblers play a coin toss game. Gambler A has $n+1$ fair coins. Gambler B has n fair coins. What is the probability that A will have strictly more heads than B if both gamblers flip all their coins?

15. In how many ways can you invest $\$20,000$ into five funds in increments of $\$1,000$? For example, one way to do it is ($\$0$; $\$4000$; $\$1000$; $\$2000$; $\$13000$).

16. You take out candies one by one from a jar that has 10 red candies, 20 blue candies, and 30 green candies in it. What is the probability that there are at least one blue candy and one green candy left in the jar when you have taken out all the red candies?

17. Let $a \leq b$ be two positive integers, and $\mathcal{X}$ a random variable that takes value—with equal probability—the powers of 2 inside the interval $[2^a, 2^b]$. Find $\mathbb{E}[\mathcal{X}]$.

18. (St. Petersburg Paradox)**.** Consider the following game. You flip a coin until the first tail appears. If the tail appears on the nth flip, you receive 2^n dollars. What is the expected gain? How much would you be willing to pay to play this game?

19. (Hypergeometric)**.** Let $\mathcal{X}$ be a random variable following a hypergeometric distribution with parameter (n, N_1, N_0). In particular, the PMF is given by

$$p_{\mathcal{X}}(k) = \frac{\binom{N_0}{k}\binom{N_1 - N_0}{n-k}}{\binom{N_1}{n}},$$

with the convention that $\binom{n}{k} = 0$ whenever $k < 0$ or $k > n$. Show that this is a valid PMF[5], and compute the mean of this distribution.

[5]Hint: Suppose a committee consists of m men and n women. In how many ways can a subcommittee of r members be formed?

20. (Negative binomial). Let $\mathcal{X}$ be a random variable following a negative binomial distribution with parameter (r, p). In particular, the PMF of $\mathcal{X}$ is given by

$$p_{\mathcal{X}}(k) = \binom{k+r-1}{r-1}(1-p)^k p^r \qquad k \geq 0.$$

Show that this is a valid PMF, and compute the mean of this distribution.

21. You roll three dice. What is the probability that the highest of the three numbers will be exactly 4?

22. Two games are offeres to you. In Game 1, you roll a die once and you are paid $\$1,000,000$ times the number of dots on the upturned face of the die. In Game 2, you roll a die one million times, and for each roll you are paid \$1 times the number of dots on the upturned face of the die. Which game do you prefer?

23. There are N distinct types of coupons in cereal boxes and each type, independent of prior selections, is equally likely in a box. Each box has only one coupon.

    - If a girl wants to collect a complete set of coupons with at least one of each type, how many coupons on average are needed to make such a complete set?
    - If the girl has collected n coupons, what is the expected number of distinct coupon types?

24. We independently and uniformly draw two subsets A and B from $\{1, 2, \ldots, n\}$.

    (i) Find the PMF of the random variable $\mathcal{X} = |A|$, where $|A|$ denotes the cardinality of A.

    (ii) Find the expectation and variance of $\mathcal{Y} = |A \cap B|$.

    (iii) Find the expectation and variance of $\mathcal{Z} = |A \cup B|$.

    (iv) Compute the covariance of Y and Z.

25. Let $(\mathcal{X}_1, \mathcal{X}_2, \ldots, \mathcal{X}_n)$ be independent random variables and set $\mathcal{X} = \sum_{k=1}^{n} \mathcal{X}_k$.

    (i) We choose $(p_1, p_2, \ldots, p_n)$ such that each $\mathcal{X}_i$ follows a Bernoulli distribution with parameter p_i and such that $\mathbb{E}[\mathcal{X}] = \mu > 0$. How might one choose $(p_1, p_2, \ldots, p_n)$ under these constraints to *maximise* $\mathrm{Var}(\mathcal{X})$?

    (ii) We choose $(p_1, p_2, \ldots, p_n)$ such that each $\mathcal{X}_i$ follows a geometric distribution with parameter p_i and such that $\mathbb{E}[\mathcal{X}] = \mu > 0$. How might one choose $(p_1, p_2, \ldots, p_n)$ under these constraints to *minimise* $\mathrm{Var}(\mathcal{X})$?

26. Suppose a receptor receives a number of bits according to a Poisson distribution with parameter λ. The bits can (independently) be either 1 with probability p or 0 with probability $1 - p$. What distribution corresponds to the number of 1's received?

27. N points are drawn randomly on the circumference of a circle. What is the probability that they are all within a semicircle?

28. What is the expected number of cards that need to be turned over in a regular 52-card deck in order to see the first ace?

29. You just bought one share of stock A and want to hedge it by shorting stock B. How many shares of B should you short to minimise the variance of the hedged position?
We will denote σ_A for the standard deviation of stock A's return, σ_B for stock B's return, and ρ their correlation coefficient. Note that the return a portfolio with x shares of stock A and y shares of stock B can be written as $xr_A + yr_B$ where r_A (resp. r_B) is the return of stock A (resp. B).

30. Let $\mathcal{X}$ be a discrete random variable such that $\mathcal{X}(\Omega) = \{x_1, x_2, \ldots, x_n\}$, and set $\mathbb{P}(\mathcal{X} = k) = p_k$ for $k = 1, 2, \ldots, b$. We define the *entropy* of $\mathcal{X}$ as follows:

$$H(\mathcal{X}) = -\sum_{k=1}^{n} p_k \log(p_k).$$

(i) Verify that $H(\mathcal{X}) \geq 0$.

(ii) Let $q_1, q_2, \ldots, q_n$ be positive real numbers such that $\sum_{k=1}^{n} q_k = 1$. Show that:

$$H(\mathcal{X}) \leq -\sum_{k=1}^{n} p_k \log(q_k),$$

with equality attained if and only if $p_k = q_k$ for all $k = 1, 2, \ldots, n$.[6] What can you say about the uniform distribution over $\{1, 2, \ldots, n\}$?

(iii) Let $\mathcal{X}, \mathcal{Y}$ be two discrete random variables and denote by $p_{\mathcal{X},\mathcal{Y}}$ the joint PMF. Define:

$$I(\mathcal{X}, \mathcal{Y}) = \sum_{x,y} p_{\mathcal{X},\mathcal{Y}}(x, y) \log \frac{p_{\mathcal{X},\mathcal{Y}}(x, y)}{p_{\mathcal{X}}(x) p_{\mathcal{Y}}(y)}.$$

Show that $I(\mathcal{X}, \mathcal{Y}) \geq 0$ and that $I(\mathcal{X}, \mathcal{Y}) = 0$ if and only if $\mathcal{X}$ and $\mathcal{Y}$ are independent.

(iv) Let $\mathcal{X}, \mathcal{Y}$ be two discrete random variables and denote by $p_{\mathcal{X},\mathcal{Y}}$ the joint PMF. Define naturally:

$$H(\mathcal{X}, \mathcal{Y}) = -\sum_{x,y} p_{\mathcal{X},\mathcal{Y}}(x, y) \log p_{\mathcal{X},\mathcal{Y}}(x, y).$$

Show that:

$$I(\mathcal{X}, \mathcal{Y}) = H(\mathcal{X}) + H(\mathcal{Y}) - H(\mathcal{X}, \mathcal{Y}).$$

(v) Define the *conditional entropy* of $\mathcal{X}$ given $\mathcal{Y}$ by:

$$H(\mathcal{X} \mid \mathcal{Y}) = -\sum_{x,y} p_{\mathcal{Y}}(y) p_{\mathcal{X}|\mathcal{Y}}(x \mid y) \log p_{\mathcal{X}|\mathcal{Y}}(x \mid y).$$

Show that:

$$I(\mathcal{X}, \mathcal{Y}) = H(\mathcal{X}) - H(\mathcal{X} \mid \mathcal{Y}).$$

31. Consider the function

$$f(x) = \begin{cases} 1 - |x| & \text{if } |x| \leq 1, \\ 0 & \text{otherwise.} \end{cases}$$

[6] Hint: If $\mathcal{X}$ is a random variable and f is a convex function, then $f(\mathbb{E}[\mathcal{X}]) \leq \mathbb{E}[f(\mathcal{X})]$ with equality iff f is linear or $\mathcal{X}$ is constant.

(i) Verify that f is a probability density function.

(ii) Compute the associated cumulative distribution function F.

(iii) Compute the probabilities $\mathbb{P}(\mathcal{X} > 0)$ and $\mathbb{P}(\mathcal{X} > 0 \mid |\mathcal{X}| < 1/2)$.

32. Pick $\mathcal{X}$ at random between 0 and 5. Let $\mathcal{Y}$ be the volume of a sphere with radius $\mathcal{X}$. Then $\mathcal{Y} = \frac{4}{3}\pi\mathcal{X}^3$, but for convenience take $\mathcal{Y} = \mathcal{X}^3$. Find the density of $\mathcal{Y}$.

33. Let f be the following function defined on $\mathbb{R}$:

$$f(x) = \begin{cases} \dfrac{a}{x\sqrt{x}} & \text{for } x \geq 1, \\ 0 & \text{otherwise.} \end{cases}$$

(i) Specify the parameter a such that f is a probability density function for some random variable $\mathcal{X}$.

(ii) Compute the cumulative distribution function $F_{\mathcal{X}}$ of $\mathcal{X}$.

(iii) Does $\mathcal{X}$ have a finite expectation?

34. Suppose that you can easily sample from a uniform distribution over $[0, 1]$. You want to simulate a random variable $\mathcal{X}$ with cumulative distribution function F. Suppose that F is strictly increasing. How would you do it?

35. (Moments of the exponential)**.** Suppose that $\mathcal{X}$ follows an exponential distribution with parameter $\lambda > 0$. Compute $\mathbb{E}[\mathcal{X}^n]$ for each $n \in \mathbb{N}$.

36. (Moments of the standard normal)**.** Suppose that $\mathcal{X}$ follows an standard normal distribution with parameter $\lambda > 0$. Compute $\mathbb{E}[\mathcal{X}^n]$ for each $n \in \mathbb{N}$.

37. Any linear combination of normal random variables is also a normal random variable.

- Suppose $\mathcal{X} \sim N(\mu, \sigma^2)$ and $\mathcal{Y} \sim N(\nu, \tau^2)$ are independent random variables. Compute the mean and the variance of any linear combination of $\mathcal{X}$ and $\mathcal{Y}$.
- Suppose that you can only simulate *independent* standard normals. How would you simulate two standard normals with correlation $\rho \in (-1, 1)$?
- Suppose that $\mathcal{X} \sim N(0, \sigma^2)$. You can accept that $\mathbb{P}(-2\sigma \leq \mathcal{X} \leq 2\sigma) \approx 0.95$ (this is a fact that proves occasionally helpful once committed to memory). Let $n \in \mathbb{N}$ and define $\mathcal{Z}_n = \frac{1}{n}\sum_{k=1}^{n} \mathcal{X}_k$ where $\mathcal{X}_1, \mathcal{X}_2, \ldots, \mathcal{X}_n$ are independent random variables with the same distribution as $\mathcal{X}$. Find a real x such that $\mathbb{P}(-x \leq \mathcal{Z}_n \leq x) \approx 0.95$.

38. Express $f(x) = \displaystyle\int_x^\infty e^{-at^2/2+bt}\mathrm{d}t$ in terms of the cumulative distribution function of the standard normal distribution.

39. Let $\mathcal{X}_1$ and $\mathcal{X}_2$ be two independent random variables with uniform distribution between 0 and 1. Let $\mathcal{Y} = \min(\mathcal{X}_1, \mathcal{X}_2)$ and $\mathcal{Z} = \max(\mathcal{X}_1, \mathcal{X}_2)$. Compute $\mathbb{P}(\mathcal{Y} \geq y \mid \mathcal{Z} \leq z)$ for any $y, z \in [0, 1]$ and compute the correlation $\rho(\mathcal{Y}, \mathcal{Z})$.

40. (a) Let $n \in \mathbb{N}$. Suppose $\mathcal{X}_1, \mathcal{X}_2, \ldots, \mathcal{X}_n$ be n independent random variables with uniform distribution between 0 and 1. What are the cumulative distribution function, probability density function, and expected value of the random variables $\mathcal{Y}_n = \min(\mathcal{X}_1, \mathcal{X}_2, \ldots, \mathcal{X}_n)$ and $\mathcal{Z}_n = \max(\mathcal{X}_1, \mathcal{X}_2, \ldots, \mathcal{X}_n)$?

(b) (Application.) Suppose 500 infinitely small ants are randomly put on a 1-foot string (with independent uniform distribution for each ant between 0 and 1). Each ant randomly moves toward one end of the string at a speed of 1 foot/minute until it falls off at one end on the string. When two ants collide, they both change directions and keep moving at 1 foot/minute. What is the expected time it takes for all ants to fall off the string?

41. (Buffon needle.) Suppose we have a floor made of parallel strips of wood, each the same width, and we drop a needle on the floor. What is the probability that the needle will lie on a line between two strips? *(Hint: Consider two cases. The short needle case where the length of the needle is smaller than the width between strips, and the long needle case where the length of the needle is longer.)*

42. Pick $\mathcal{X}$ at random between 0 and 1. If $\mathcal{X} = x$, toss a coin five times where $\mathbb{P}(\text{heads}) = x$. Let N be the number of heads.

    (a) Compute $\mathbb{E}[\mathcal{X}]$.

    (b) Calculate $\mathbb{P}(\mathcal{X} \leq 0.5 \mid N = 0)$.

    (c) Calculate the posterior distribution $\mathbb{E}[\mathcal{X} \mid N = n]$.

43. Suppose at the first stage of an experiment, the result is $\mathcal{X}$ where

$$f_{\mathcal{X}}(x) = xe^{-x} \text{ for } x > 0.$$

    Then, if $\mathcal{X} = x$, at the second stage $\mathcal{Y}$ is chosen at random between 0 and x. In other words,

$$f(y \mid x) = \frac{1}{x} \text{ for } x \geq 0 \text{ and } 0 \leq y \leq x.$$

    Find $\mathbb{P}(\mathcal{Y} \leq 2)$.

44. (Least squares.) A sequence $(\mathcal{X}_n)$ of random variables is said to converge to a number c **in the mean square**, if $\lim_{n\to\infty} \mathbb{E}[(\mathcal{X}_n - c)^2] = 0$.

    (a) Show that convergence in the mean square implies convergence in probability; that is, for all $\varepsilon > 0$, $\lim_{n\to\infty} \mathbb{P}(|\mathcal{X}_n - c| > \varepsilon) = 0$.

    (b) Give an example that shows that convergence in probability does not show convergence in the mean square.

45. A factory produces $\mathcal{X}_n$ gadgets on day n where the $\mathcal{X}_n$ are independent and identically distributed random variables, with mean 5 and variance 9.

    (a) Find an approximation to the probability that the total number of gadgets produced in 100 days is less than 440.

    (b) Find (approximately) the largest value of n such that

$$\mathbb{P}(\mathcal{X}_1 + \mathcal{X}_2 + \ldots + \mathcal{X}_n \geq 200 + 5n) \leq 0.05.$$

    (c) Let N be the first day on which the total number of gadgets produced exceeds 1000. Calculate an approximation to the probability that $N \leq 220$.

46. Let $(\mathcal{X}_1, \mathcal{Y}_1), (\mathcal{X}_2, \mathcal{Y}_2), \ldots$ be independent random variables, uniformly distributed in the unit interval [0, 1], and let

$$\mathcal{W} \coloneqq \frac{(\mathcal{X}_1 + \ldots + \mathcal{X}_{16}) - (\mathcal{Y}_1 + \ldots + \mathcal{Y}_{16})}{16}.$$

Find a numerical approximation to the quantity $\mathbb{P}(|\mathcal{W} - \mathbb{E}[\mathcal{W}]| < 0.001)$.

47. Let $\mathcal{Y} = \mathcal{X}_1 + \mathcal{X}_2 + \ldots + \mathcal{X}_n$ where the $\mathcal{X}_i$ are independent such that $\mathcal{X}_i$ has a normal distribution with mean μ_i and variance σ_i^2. Find the distribution of $\mathcal{Y}$.

48. Let $\mathcal{X}$ be a continuous random variable with probability density function $f_{\mathcal{X}}$ and moment generating function $M_{\mathcal{X}}$ defined on an interval $(-h, h)$, for some $h > 0$. Show that

$$\mathbb{P}(\mathcal{X} \geq a) \leq e^{-at} M_{\mathcal{X}}(t) \quad \text{for } 0 < t < h.$$

49. Given $t < 1/5$, let $\mathcal{X}$ and $\mathcal{Y}$ be two independent random variables with respective moment generating functions

$$M_{\mathcal{X}}(t) = \frac{1}{1 - 5t}, \quad M_{\mathcal{Y}}(t) = \frac{1}{(1 - 5t)^2}.$$

Find $\mathbb{E}[\mathcal{X} + \mathcal{Y}]^2$.

50. Consider a sequence of $(\mathcal{Y}_n)$ of non-negative random variables and suppose that

$$\sum_{n=1}^{\infty} \mathbb{E}[\mathcal{Y}_n] < \infty.$$

Show that $(\mathcal{Y}_n)$ converges to 0 in probability.

51. Find the moment generating function of $\mathcal{X}$ whose density is $x \mapsto xe^{-x}$.

52. (Gamma distribution.) Let $\alpha, \beta > 0$ be two positive reals. We say that a random variable $\mathcal{X}$ follows a Gamma distribution with parameters (α, β) if $\mathcal{X}$ has the PDF

$$f_{\alpha,\beta}(x) = \frac{\beta^\alpha x^{\alpha-1} e^{-\beta x}}{\Gamma(\alpha)}, \quad \text{with } x > 0,$$

where $\Gamma : (0, \infty) \to \mathbb{R}$ is defined by

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

(a) Show that the function Γ is well-defined for $x > 0$, express $\Gamma(x+1)$ in terms of $\Gamma(x)$ and deduce a formula for $\Gamma(n)$ for any $n \in \mathbb{N}$.

(b) Compute the moment generating function of a random variable $\mathcal{X} \sim \mathsf{Gamma}(\alpha, \beta)$. Deduce its mean and variance.

(c) (Application.) Let $\mathcal{X}_1, \mathcal{X}_2, \ldots, \mathcal{X}_n$ be independent random variables with exponential distributions of parameter θ. Give the distribution of $\mathcal{Z}_n = \mathcal{X}_1 + \mathcal{X}_2 + \ldots + \mathcal{X}_n$.