Data Analysis

Lecture Notes

Typeset by Max Melching
Based on a lecture given by Prof. Dr. Maria Alessandra Papa

Outline and Information

For the latest version of this file, see https://github.com/MaxMelching/physics_notes.

Important Note: of course, I tried my best to avoid mistakes. However, I can not guarantee that all statements in these notes are actually correct!

Motivation: often there is stochastic element to data, e.g. noise, so we need statistics → often use same, simple examples, e.g. coin flips

Goal of this course: learn to ask the right questions

Book recommendation: "Introduction to probability" by Bertsekas & Tsitsikas

Contents

1 Theory **1**
1.1 Foundations 1
1.2 Conditional Probability 4
1.3 Independence of Events 10
1.4 Counting 13

2 Random Variables **18**
2.1 What is a RV? 18
2.2 Multiple RVs 23
2.3 Functions of RVs 25
2.4 Continuous RVs 26
2.5 Bayes' Theorem 38
2.6 Distributions of Functions of RVs 41
2.7 Iterated Expectations and Conditional Variances 45
2.8 Limit Theorems 47

1 Theory

1.1 Foundations

Besides an intuitive understanding of experiments, it is also important to have a theoretical description to be able to make reliable predictions. The most important object needed for this is a set that contains outcomes of the experiment. Once this is defined, one can make sense of functions that act on this set and e.g. assign probabilities to outcomes.

Definition 1.1: Sample Space

The sample space of an experiment is a set, whose elements are all possible outcomes of the experiment. These elements have three important properties:

(i) They are mutually exclusive, i.e. if one outcome occurs, then none of the others can occur (having heads means we cannot have tails)

(ii) They are collectively exhaustive, all possible outcomes are contained in it

(iii) One has to be careful with choosing the right granularity, i.e. which variables to incorporate into the definition of outcomes (there are many that can be taken into account, but not all should be; this is personal judgement)

Example 1.2: Construction of Sample Spaces

Consider rolling a four-faced die twice. One way to describe this is using coordinates (x, y) to encode the result of the first (x) and second roll (y), which leads to a 4×4 grid (table 1.1). Another useful description is sequential-based, which means we model the outcomes of each roll as branches in a tree diagram (figure 1.2).

In contrast, when a dart is thrown onto a board, there is a continuous range of possible outcomes collected in the sample space $\Omega = \{(x, y) : 0 \leq (x, y) \leq 1\}$ (potentially, some rescaling of the board is needed for these boundaries to be valid; figure 1.1). In this case, probabilities can not be assigned to points, only to areas (a bit like mass).

Logically, some properties have to hold for the notion of probability to make sense. That includes a finite total probability (specific value is conventional, 1 is customary) or that negative probabilities must not occur. These properties can be formulated as axioms, which build the mathematical foundation of probability theory.

Property 1.3: (Kolmogorov) Axioms of Probability Theory

(i) $P(A) \geq 0$ for all events $A \subset \Omega$

(ii) $P(\Omega) = 1$ (*closure rule*)

(iii) $P(A \cup B) = P(A) + P(B) - P(A \cap B)$ (*total probability rule*)

y				
4				
3				
2				
1				
	1	2	3	4 x

Table 1.1: Tabular representation of the sample space the for double roll of a four-faced die. Each entry can be filled with the corresponding probability.

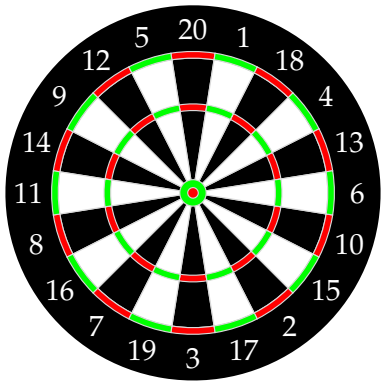


Figure 1.1: Dart board. One can assign probabilities to each of the singles, doubles, triples. The code for this picture is copied from: https://de.overleaf.com/latex/templates/dartboard/bhpfmdvjsjmk

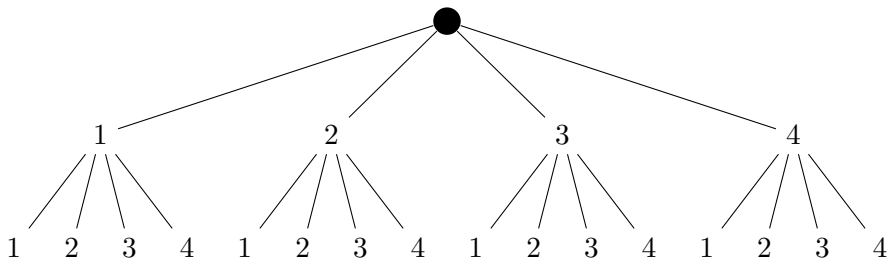


Figure 1.2: Sequential/tree diagram representation of the sample space for the double roll of a four-faced die. On each path, one could denote corresponding probability (or in each node, with the path holding the corresponding number).

A direct corollary of these is $P(A) \leq 1, \forall A$. Property (iii) reflects that the probability of A or B occurring is nothing but the probability of A occurring plus the probability of B occurring, but one has to subtract the probability of A and B occurring simultaneously (their *joint probability*) to avoid double counting. It also motivates the analogy that probability behaves like mass.

These properties put constraints on probability values and explain how we may infer probabilities of combined events, but they say nothing about how to actually assign them.

Example 1.4: Assigning Probabilities

In the example of throwing a fair, four-faced die twice, the probability of each outcome $(x, y),\ 1 \leq x, y \leq 4$ is simply $1/16$ (the sample space is uniform). From that, one can assign more abstract probabilities like $P(\{x = 1\})$ or $P(\{\min(x, y) = 2\})$ by looking at how many outcomes lead to this event being realized ($\equiv$ summing, which is justified by total probability rule) and then dividing by the total number of outcomes.

The reason that one has to divide by this total number is that probabilities have to be normalized such that the closure rule is fulfilled. Since $\Omega = A \cup A^c\ (A \cap A^c = \emptyset)$ for each event A, the axioms of probability tell us that

$$P(A) + P(A^c) = P(\Omega) = 1 \quad \Leftrightarrow \quad P(A^c) = 1 - P(A)\,. \tag{1.1}$$

Calling n the number of outcomes that realize A (implies the number of outcomes in A^c is $N - n$ where N is the number of all outcomes in Ω), this can not be fulfilled for $P(A) = n, P(A^c) = N - n$. Instead, one has to divide by N, yielding

$$P(A) + P(A^c) = \frac{n}{N} + \frac{N - n}{N} = \frac{N}{N} = 1$$

as desired.

For the events mentioned in the previous paragraph, this means

$$P(\{x = 1\}) = \frac{1}{16} + \frac{1}{16} + \frac{1}{16} + \frac{1}{16} = \frac{4}{16} = \frac{1}{4}, \qquad P(\{\min(x, y) = 2\}) = \frac{5}{16}$$

as one can see by looking at table 1.2.

For continuous variables on the other hand, it is only possible to compute areas and thus assign probabilities like $P(x + y \leq 1/2)$ (see 2.4 for more details).

The number of elements that a set contains is also called the *cardinality* $|\cdot|$. Using this notation, the results of the example can be summarized in the formula

$$P(A) = \frac{|A|}{|\Omega|} \tag{1.2}$$

which is a very useful way to compute probabilities in uniform sample spaces.

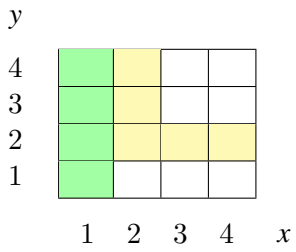


Table 1.2: Visual representation of the events $A = \{x = 1\}$ (green) and $A = \{\min(x, y) = 2\}$ (yellow). The x-axis represents the first roll, the y-axis the second.

1.2 Conditional Probability

A belief in the likelihood of some event defines the way we assign probabilities of it occurring, i.e. the *model* we use. However, as soon as new information such as data from measurements comes in, this initial belief/assignment has to (or at least should) be updated. A very easy example is that a priori, we would assign a probability of $1/6$ to each face of a die, but that changes if we get to know that it is rigged. To incorporate new information into probabilities, *conditional probabilities* $P(A|B)$ can be used.

To define them we use the fact that new information, e.g. that an event B has been measured, should affect the way one assigns probabilities to other events A (that potentially have some finite overlap with B, i.e. $A \cap B \neq \emptyset$).[1] To distinguish "old" and "new" probabilities, P and P' will be used. Since it is known that B was measured, $P'(B) \overset{!}{=} 1$ regardless of the value of $P(B)$. As a consequence, all events C with $C \cap B = \emptyset$ are ruled out because $P'(C) = 0$. Basically, this means there is a new sample space, namely $\Omega' = \Omega \cap B = B$. Assigning probabilities in this new, conditioned universe works just like before (assuming a uniform sample space),

$$P'(A) = \frac{|A'|}{|\Omega'|} = \frac{|A \cap \Omega'|}{|\Omega'|} .$$

Using $A' = A \cap \Omega'$ ensures only outcomes in the conditioned universe Ω' occur, just like Ω' was defined as $\Omega \cap B$.[2] This definition allows to find an expression for $P'(A)$ using P (also see figure 1.3):

$$P(A|B) := P'(A) = \frac{|A \cap \Omega'|}{|\Omega|} \frac{|\Omega|}{|\Omega'|} = \frac{|A \cap B|}{|\Omega|} \frac{1}{|B|/|\Omega|} = \frac{P(A \cap B)}{P(B)} . \tag{1.3}$$

The intuition behind this formula is that the conditioning forbids certain outcomes, which leads to a violation of the closure rule. To solve this issue, one rescales the probabilities of all outcomes that are left (i.e. that lie in B, which is why each event has an additional $\cap B$) such that the closure rule is fulfilled again. The same idea also applies to non-uniform sample spaces: events are restricted to their intersection with B and then the probabilities are normalized by dividing by the probability of the "new sample space" B.

[1] It is important to emphasize that events remain unchanged under conditioning, only probabilities change.
[2] Implicitly, this is done before when using P as well, but always omitted since $A \cap \Omega = A$, $\forall A$.

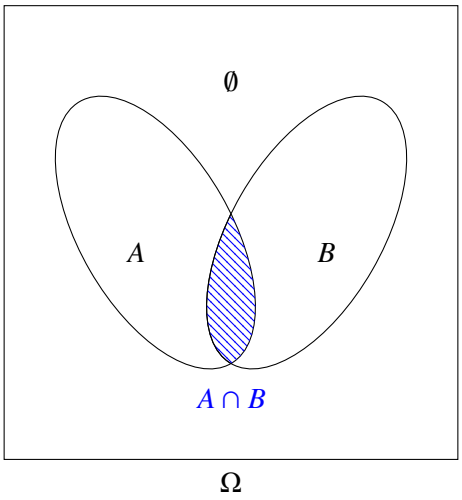


Figure 1.3: Visualization of idea behind conditional probabilities. The square is used to represent the whole sample space Ω and two ellipses to represent events A, B with $A \cap B \neq \emptyset$ and $\Omega \backslash (A \cap B) = \emptyset \Leftrightarrow \Omega = A \cup B$.

For another viewpoint on this definition, let us rearrange it to read $P(A \cap B) = P(B)P(A|B)$. One can interpret this very intuitively: in order for A and B to occur, either B has to occur first and A after that (i.e. knowing B has already occurred) or vice versa (equivalent since $P(A \cap B) = P(B \cap A) = P(A)P(B|A)$). This idea applies to non-uniform sample spaces, too.

Example 1.5: Four-faced Die

Consider the case of two ordered rolls of a four-faced die, which is assumed to be fair (so the sample space is uniform with each event having probability $1/16$). The corresponding sample space has already been constructed in table 1.1.

Let $M = \max(x, y)$, $m = \min(x, y)$ (recall: x is the index for the first throw, y for the second). Knowing that $m = 2$, what is the probability that M takes a certain value?

The first way to compute these probabilities is to visualize the conditioned universe Ω' as well as the event A and then use $P(A|\{m = 2\}) = \frac{|A'|}{|\Omega'|}$ (figure 1.3), yielding

$$P(\{M = 1\}|\{m = 2\}) = \frac{0}{5} = 0\,, \qquad P(\{M = 2\}|\{m = 2\}) = \frac{1}{5}\,.$$

On the other hand, we can use the definition and not work in the conditioned universe explicitly (to check that the definition works). This approach confirms that

$$P(\{M = 1\}|\{m = 2\}) = \frac{P(\{M = 1\} \cap \{m = 2\})}{P(\{m = 2\})} = \frac{0/16}{5/16} = 0$$
$$P(\{M = 2\}|\{m = 2\}) = \frac{P(\{M = 2\} \cap \{m = 2\})}{P(\{m = 2\})} = \frac{1/16}{5/16} = \frac{1}{5}\,.$$

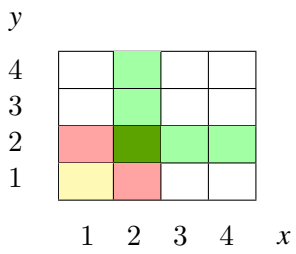


Table 1.3: Visual representation of the conditioned universe where $m = \min(x, y) = 2$ (green) and the events $A = \{M = \max(x, y) = 1\}$ (yellow), $A = \{M = \max(x, y) = 2\}$ (red). The x-axis represent the first roll (index x), the y-axis the second (index y). **Remark**: cell $(2, 2)$ is red *and* green, which leads to color it has.

An interesting property is that conditioning on a uniform sample space results in a uniform sample space again, after all it is essentially a renormalization (can also observed in the previous example). Reasoning in terms of conditional probabilities turns out to be very important and widely applicable.

Say, for example, there is a feature that potentially exists in data obtained from a detector. Using F to denote that the feature is present (F^c that it is not) and D to denote that the detector shows that F is present (D^c that it does not show it)[3], the following conditional probabilities have different, important interpretations:

Definition 1.6: True and False Positives/Negatives

- $P(D|F)$: *detection efficiency/probability*, measures how often the detector correctly shows that the feature is there (true positive)
- $P(D^c|F)$: *false dismissal rate/probability*, measures how often the detector wrongly shows that the feature is not there (i.e. detector shows it is not there while it is there; false negative). Also called *type I error*
- $P(D|F^c)$: *false alarm rate/probability*, measures how often the detector wrongly shows the feature is there (i.e. detector shows there while it is not there; false positive). Also called *type II error*.
- $P(D^c|F^c)$: *rejection efficiency/probability* (?don't remember correct name; maybe true dismissal rate?), measures how often the detector correctly shows that the feature is not there (true negative)

It should be obvious that high detection, rejection efficiencies and low false dismissal, false alarm rates are desirable for real-world applications. Keeping the false alarm rate as small as possible should always be prioritized, though.

[3] This is regardless of whether the feature really is present and only about the output of the observation.

Example 1.7: Radar

To see how all the definitions can be used and how they differ from each other, one can look at the detection of airplanes A using data obtained from a radar R (airplane is feature F, radar is detector D). From looking up to the sky, one can estimate that in about 5% of the times an airplane flies by. Our goal is to estimate how reliably this can be detected using a radar.

Consider now a device with the following specs: 99% detection efficiency (if airplane flies by, radar registers) and 10% false alarm rate (radar registers something despite no airplane flying by). The interesting question is how good these specs actually are, i.e. how sure one can be that a detection claim (e.g. clicking) of the radar was caused by an airplane. From the numbers, we would assess that the statements are very certain since the detection efficiency is 99% (only 1% of the cases are false dismissals, where the radar does not click despite an airplane being present).

However, instead of just relying on intuition, it is better to compute the probability $P(A|R)$ that assesses how likely radar clicks $\Rightarrow$ airplane present. By definition,

$$P(A|R) = \frac{P(A \cap R)}{P(R)} = \frac{P(A)P(R|A)}{P(R)}\,.$$

$P(R|A)$ is just the detection efficiency (known), $P(A)$ is our guess/belief how often an airplane is present when looking up in the first place (estimated to 5%) and

$$P(R) = P(R \cap A) + P(R \cap A^c) = P(R|A)P(A) + P(R|A^c)P(A^c)\,.$$

Now we can finally compute

$$P(A|R) = \frac{99\% \cdot 5\%}{99\% \cdot 5\% + 10\% \cdot 95\%} = 34.26\%\,,$$

which is surprisingly small, considering the detection efficiency of 99%. This is due to the false alarm rate being relatively high, which becomes relevant because there is no airplane for the majority of time (95%), so the majority of clicks will be caused by false alarms. To be more precise,

$$P(A^c|R) = \frac{P(R|A^c)P(A^c)}{P(R|A)P(A) + P(R|A^c)P(A^c)} = 65.74\%$$

of the clicks happen with no airplane being present, which is nothing but $1 - P(A|R)$. This is expected and in fact has to be fulfilled because the total probability rule tells us

$$1 = P(\Omega) = P(\Omega|R) = P(A \cup A^c|R) = P(A|R) + P(A^c|R)\,.$$

This example is a first application of *Bayes' theorem*, which is very widely used in science for all kinds of inference. It allows to reverse conditioning, i.e. get the probability $P(A|E)$ of a scenario A being true if an effect E is observed from the probability $P(E|A)$ of the scenario

causing this effect (the latter being a causal *model*, something that is e.g. noted in spec sheets of instruments). That can be done by interpreting measurements E in terms of a *likelihood* $P(E|A)$ that A was the underlying scenario.[4] This interpretation can then be used to update the initial probability (better to think of it in terms of a belief) of some causal model being true from $P(A)$ to $P(A|E)$, i.e. one can infer something about A from E by incorporating it systematically into the results. To reflect these roles, the names *prior probability* for $P(A)$ and *posterior probability* for $P(A|E)$ are used.[5]

Very helpful tools when working with Bayes' theorem are the total probability and multiplication rule. Both of them have already been mentioned, but their most general form has not been given yet (as it requires conditional probabilities).

Property 1.8: Total Probability Rule, Multiplication Rule

For disjoint events $A_1, \dots, A_n$ forming a partition of the sample space Ω (each outcome is included only once in the partition) and an arbitrary event B

$$P(B) = P(B \cap A_1) + \cdots + P(B \cap A_n) = P(A_1)P(B|A_1) + \cdots + P(A_n)P(B|A_n)\,. \tag{1.4}$$

Furthermore,

$$P\left(\bigcap_{i=1}^{n} A_i\right) = P(A_1)P(A_2|A_1)P(A_3|A_1 \cap A_2)\dots P(A_n|\bigcap_{i=1}^{n-1} A_i)\,. \tag{1.5}$$

These properties allow formulating slightly more general versions of Bayes' theorem, e.g. for multiple scenarios A_i forming a partition of the sample space:

$$P(A_i|E) = \frac{P(A_i \cap E)}{P(E)} = \frac{P(A_i)P(E|A_i)}{\sum_j P(A_j)P(E|A_j)}\,. \tag{1.6}$$

One of the difficulties in real-world inference is assigning the priors $P(A_i)$. This is especially problematic if the data does not contain much helpful information regarding the problem. In this case, the posterior will be prior-dominated (i.e. a change in the prior also changes the posterior substantially), which means it plays an even more important role. This is less relevant if the data contains a lot of information because the likelihood will dominate the posterior values. In the example of radar and airplane, changing the prior from 5% to 10% causes the posterior to go from 34.26% to 52.38%, which means there is a clear sensitivity to it. More informative data would lead to a likelihood with a sharper peak, i.e. the difference between detection efficiency $P(R|A)$ and false alarm rate $P(R|A^c)$ would be bigger (as already stated, the likelihood is a function of the scenario we condition on, *not* the input data, so no closure rule needs like $P(R|A) + P(R|A^c) = 1$ has to be fulfilled).

[4] The likelihood is a function of A, not E, and not a probability distribution with respect to this parameter. Instead, it is a distribution with respect to the data, which is evaluated in the point/measurement E and has a parameter A that is varied.

[5] The "probability" part of the name is often omitted.

Example 1.9: Monty-Hall Problem

A very famous problem in probability theory is the Monty-Hall problem from a TV-show in the US. The basic setting consists of three doors, two of which hide a goat and one of which hides a car. Each participant of the show can keep whatever is behind the door he chooses to open. However, the door is not opened immediately after the participant makes his first choice. Instead, the host (who knows where the car is) opens a door that reveals a goat (and that was not chosen by the participant). After that, there is a possibility to switch doors or keep the chosen one. The relevant question in this context is: what is the best strategy, switching or keeping?

It might seem unintuitive, but the best way to go is switching. A priori, each door has probability $1/3$ of hiding the car. However, the host has additional information that he reveals when opening the door. This is due to the constraints that he must not open the participant's door, but also not the door hiding the car. As a consequence, by opening one of the other doors its probability transfers to the other non-opened door, so all of a sudden this door has a probability of $2/3$ to hide the car (while the participant's door remains at $1/3$). Therefore, switching is best strategy.

One can also derive this mathematically, e.g. from the corresponding sequential diagram (figure 1.4). Without loss of generality, one can assume that the participant chooses door 1 (which is denoted as P_1). This is because the car is in a random spot anyway, so randomly choosing the participant's door as well is not required (this choice removes one level from the sequential diagram, which makes it much smaller and less confusing). First of all, there are three spots C_i where the car can be and all have equal probability $1/3$ (and they are independent of the participant's choice, $P(C_i|P_1) = P(C_i)$). In the next step/level, the host chooses a door D_i. This choice is based on the participant's choice P_1 as well as the location of the car C_i, i.e. conditioned on them. As one can see, in many cases the host has no real choice other than selecting one specific door because he would open the car's door otherwise. From the probabilities on each path one can compute the relevant probabilities of success according to:

$$\begin{aligned} P(\text{car} \wedge \text{stay}) &= P(C_1 \wedge D_2) + P(C_1 \wedge D_3) = P(D_2|C_1)P(C_1) + P(D_3|C_1)P(C_1) \\ &= \frac{1}{2} \cdot \frac{1}{3} + \frac{1}{2} \cdot \frac{1}{3} = \frac{1}{3} \\ P(\text{car} \wedge \text{switch}) &= P(C_2 \wedge D_3) + P(C_3 \wedge D_2) = P(D_3|C_2)P(C_2) + P(D_2|C_3)P(C_3) \\ &= 1 \cdot \frac{1}{3} + 1 \cdot \frac{1}{3} = \frac{2}{3}. \end{aligned}$$

This is a very illustrative example of how new information changes problems and that sticking to certain basic rules is very powerful despite being seemingly simple.

Fun fact: there was a long, vivid discussion among mathematicians on whether it is $1/3$ vs. $2/3$ or $1/2$ vs. $1/2$. Many of them were not convinced by this logical argument, but instead by Monte-Carlo simulations.

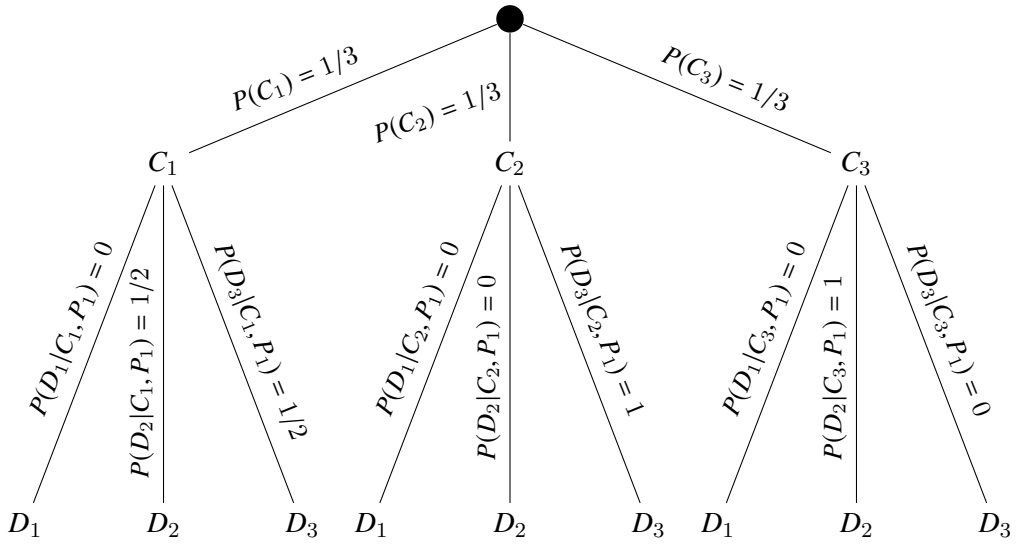


Figure 1.4: Sequential diagram for the Monty-Hall problem. Without loss of generality, it is assumed that the participant chooses door 1 (P_1). Based on that, the first level models car location C_i and the second level which door D_i is chosen by the host.

1.3 Independence of Events

Making sense of the notion of independence can be done intuitively, without using any math. Two events A, B are independent if A happening does not change beliefs about the likelihood of B happening or, in the language of conditional probabilities,

$$P(B) = P(B|A) = \frac{P(A \cap B)}{P(A)}\,. \tag{1.7}$$

However, this only works if $P(A) > 0$. Of course though, we want the definition to cover all possible cases. That means we are looking for a better way to phrase things, the general idea does not change. It turns out that we can just use a rearranged version of (1.7).

Definition 1.10: Independence of Events

Two events A, B are *independent* if and only if

$$P(A \cap B) = P(A)P(B)\,. \tag{1.8}$$

An interesting corollary of this definition is that events A with $P(A) = 0$ as well as Ω are independent of all other events (should be intuitive) because

$$P(A \cap B) = P(A)P(B|A) = 0 = P(A)P(B),\ \forall B$$
$$P(\Omega \cap B) = P(B) = 1 \cdot P(B) = P(\Omega)P(B),\ \forall B\,.$$

Despite a very intuitive definition, one has to be a bit careful about this notion. One reason is

that disjoint events A, B (where we assume $A, B \neq \emptyset \implies P(A), P(B) > 0$) are not independent since $P(A \cap B) = P(\emptyset) = 0 \neq P(A)P(B)$. In fact, this is a poster-child of non-independent events as measuring one of them excludes the possibility of the other event occurring. On the other hand, if A, B do have a finite overlap $A \cap B$, it does not necessarily mean that they are independent. To be really sure, one has to verify it via calculation each time.

Another source of confusion is that the independence of events is not really related to the events themselves, but rather to their probabilities (therefore, the wording is kind of unfortunate). At first, it might seem unintuitive that this makes a difference and indeed, it is subtle. Nonetheless, it is important and we can see that this is really the case by looking at *conditional independence*. In a conditional universe, demanding $P(A \cap B) = P(A)P(B)$ takes the form $P(A \cap B|C) = P(A|C)P(B|C)$. It is now very easy to construct an example where A, B are independent in the non-conditioned universe, but not independent in the conditioned one. For A, B with $A \cap B, A \cap C, B \cap C \neq \emptyset$ it is still possible that $A \cap B \cap C = \emptyset$, which implies $P(A \cap B|C) = 0 \neq P(A|C)P(B|C)$. Hence, A, B are not independent when we look at probabilities conditioned on C. This example shows that for the same events A, B, them being independent depends on the probabilities $P(A), P(B)$ vs. $P(A|C), P(B|C)$ rather than just A, B as events. An illustration of the situation is provided in figure 1.5.

Example 1.11: Unfair Coins

Suppose we flip two unfair coins C_1, C_2 with $P(H|C_1) = 0.9,\ P(H|C_2) = 0.1$. We will now calculate the possibility that the result of two consecutive coin flips is head, not knowing which of the coins is flipped, i.e. $P(HH)$. To do that, we will apply the total probability rule (1.4) using the disjoint events C_1, C_2:

$$P(HH) = P(C_1)P(HH|C_1) + P(C_2)P(HH|C_2)\,.$$

The prior probability to get each coin is $P(C_1) = 0.5 = P(C_2)$. To calculate the conditional probabilities $P(HH|C_1),\ P(HH|C_2)$, we can simply change perspective and think in the conditioned universe where we know which coin is flipped. In this case, the flips are clearly independent and we can simply multiply the respective probabilities to obtain

$$P(HH|C_1) = P(H|C_1)^2 = 0.9^2 = 0.81, \qquad P(HH|C_1) = P(H|C_2)^2 = 0.1^2 = 0.01\,.$$

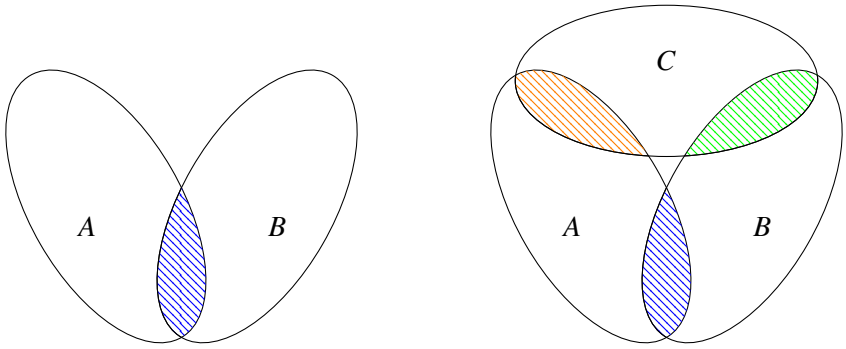
Combining these results yields

$$P(HH) = \frac{1}{2} \cdot 0.81 + \frac{1}{2} \cdot 0.01 = 0.41\,,$$

which is higher than $P(HH) = 0.25$ in the case of fair coins.

A very interesting observation can be made when examining the results of $P(H)^2$, which is the probability of getting two heads when treating the coin flips as independent events. Using the total probability rule again, we can calculate

$$P(H) = P(C_1)P(H|C_1) + P(C_2)P(H|C_2) = \frac{1}{2} \cdot 0.9 + \frac{1}{2} \cdot 0.1 = 0.5$$



(**a**) Unconditioned A, B (**b**) A, B with conditioning event C

Figure 1.5: Visualization of idea behind conditional independence. (a) shows two events with a finite overlap $A \cap B$ (blue) and (b) shows the same situation with a conditioning event C that has finite overlaps $A \cap C$ (orange), $B \cap C$ (green).

which implies

$$P(H)^2 = 0.25 \neq 0.41 = P(HH)\,.$$

This result is very interesting because it tells us that the coin flips are not independent. How does that come about? We can answer this question by thinking in terms of information and extending the experiment to more flips. If, for example, 90 out of 100 flips gave H, it is *much* more likely that we are flipping the first coin rather than the second one.[6] For this reason, it makes a lot of sense that the probability of the 101-st flip to give H differs from the probability of the first flip giving H. The same argument also applies to the second flip, where information coming from the first flip can be used (which causes $P(H)^2 \neq P(HH)$), with the only difference being that the amount of information gained is smaller.

As always, one could obtain the same result from looking at the corresponding sequential diagram (the total one or only the conditioned branch). The bottom line of this example is that independence is about knowledge as well! To really understand this, one also has to remember that probabilities are assigned by us, they are not necessarily intrinsic properties (so they might as well be wrong if we do not understand the problem or if knowledge is missing, e.g. whether the coin we flip is fair or not).

The second condition for independence can be extended very easily to more than two events, where it takes the form

$$P(A_1 \cap \cdots \cap A_n) = P(A_1) \ldots P(A_n)\,. \tag{1.9}$$

Although it might not be obvious, this is equivalent to demanding independence for any subset of events $\{A_i\}$. Pair-wise independence on the other hand is *not* sufficient.

[6] As a sidenote: Bayes' theorem allows us make this statement more quantitative ($A := \{90\,H \text{ in } 100 \text{ flips}\}$), yielding $P(C_1|A) = \frac{P(C_1)P(A|C_1)}{P(C_1)P(A|C_1)+P(C_2)P(A|C_2)} = \frac{0.5\cdot 0.9^{100}}{0.5\cdot 0.9^{100}+0.5\cdot 0.1^{100}} = 0.9999\ldots$, which is indeed *much* more likely than the probability of flipping the second coin, $P(C_2|A) = 1 - P(C_1|A) = 3.76 \cdot 10^{-96}$.

Example 1.12: Coin Flips

To validate the claim that pair-wise independence of events is not a sufficient condition for independence, we can look at the example of two coin flips again (where we assume a fair coin now). Consider the event $A = \{\text{result of the first flip is } H\}$ and $B = \{\text{result of the second flip is } H\}$. Clearly, $P(A) = \frac{1}{2} = P(B)$ and thus

$$P(A) \cdot P(B) = \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{4}\,.$$

As we can see from table 1.4, this coincides with $P(HH)$ since the sample space is uniform and A, B are independent (as expected).

Consider now the additional event $C = \{\text{first } \textit{and} \text{ second flip give the same result}\}$. First of all, $P(C) = \frac{1}{2}$ as two of the four outcomes fulfil the condition (see table 1.4). Also, since there is only one way to get H in both flips,

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

This is pair-wise independence. However,

$$P(A \cap B \cap C) = \frac{1}{4} \neq \frac{1}{8} = \frac{1}{2} \cdot \frac{1}{2} \cdot \frac{1}{2} = P(A) \cdot P(B) \cdot P(C)\,,$$

the joint probability does not factorize in the way that is necessary for independence. The reason behind this is that knowing A, B are happening, we know for sure that C also happens, i.e.

$$P(C|A \cap B) = 1 \quad \Rightarrow \quad P(A \cap B)P(C|A \cap B) = \frac{1}{4} \neq \frac{1}{8} = \frac{1}{4} \cdot \frac{1}{2} = P(A \cap B)P(C)\,.$$

Hence, there is a non-independent connection between A, B, C and this shows that pair-wise independence is not equivalent to independence of more than two events.

1.4 Counting

A very important task in assigning probabilities is counting, which is basically combinatorics. This is especially important in case of a uniform sample space Ω with N elements (i.e. there is a finite number of outcomes, each occurring with equal probability $1/N$). The *cardinality* of a set is the number of elements in it, e.g. $|\Omega| = N$. The reason why counting is important in this case is that for an event A that can be realized in $n = |A|$ ways,

$$P(A) = \sum_{i=1}^{n} \frac{1}{N} = \frac{n}{N} = \frac{|A|}{|\Omega|}\,, \tag{1.10}$$

as mentioned previously. This is one of the reasons why Monte-Carlo methods are used so widely to simulate probabilities and statistical quantities.

flip 2		
T	HT	TT
H	HH	TH
	H	T

flip 1

Table 1.4: Visual representation of the sample space of two consecutive coin flips (where a fair coin is assumed, every outcome has the same probability $\frac{1}{4}$. As the axes labels indicate, the x-axis is used for results of the first flip and the y-axis for results of the second flip.

Sometimes though, sets are defined implicitly, so counting is tricky. Luckily, some very useful shortcuts exist.

Example 1.13: Number Plate

Using combinatorics, one can figure out how many different number plates with three letters and four digits can be constructed. Assuming a 26-letter alphabet, there are

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

possible choices. If no letters and digits may be repeated, this number reduces to

$$26 \cdot 25 \cdot 24 \cdot 10 \cdot 9 \cdot 8 \cdot 7 = 78\,624\,000\,. \tag{1.11}$$

Very generally speaking, the treatment of such problems often involves counting permutations of a set, i.e. different orderings. Denoting the number of elements in a set with n, the idea is to count the number of different ways each spot can be occupied: the first one can contain each of the n objects; having fixed the first spot, there are $n-1$ objects left to choose from for the second spot, etc. The resulting number is called *factorial*

$$n! = n \cdot (n-1) \ldots 1\,. \tag{1.12}$$

A more general case is an arbitrary number of stages/choices (with index i), where each stage has n_i possible outcomes. The total number of outcomes of this multi-stage event is

$$\prod_i n_i\,. \tag{1.13}$$

A similar task is to find the number of subsets that can be created from a set of n elements. The idea is to look at the corresponding decision tree (basically sequential diagram) that has n levels, one for each object. For each of them, starting from the empty set $\emptyset$, one can create a new set by adding it and one by not adding it, so there are two new sets per object. Consequently, 2^n subsets can be created.

Example 1.14: King's Sibling

Suppose a royal family has two kids and we know there will be a king, which means one of the children is a boy. What are the odds that the other child is a girl (remember: it does not matter if girl is older, boy will become king)?

To get the answer, we can look at all possible outcomes:

B B	B G
G B	G G

Notice that not all of them can happen. The red one is excluded because of the knowledge there will be a king (i.e. one child is a boy). Therefore, the answer is $2/3$.

Example 1.15: Die Roll

For a fair six-faced die, the probability of each outcome $1, 2, 3, 4, 5, 6$ is $P(A) = \frac{1}{6}$. For six rolls, each outcome like six times 6 therefore has a probability $P(A) = \frac{1}{6^6} \Leftrightarrow |\Omega| = 6^6$.

If we are now interested in the event $B = \{\text{all six rolls yield different numbers}\}$, we have to count the number of times this can happen. Obviously, the sequence $(1, 2, 3, 4, 5, 6) \in B$. From the definition of B, we know that every other element in it has to be a permutation of this sequence, i.e. $|B| = 6!$ and

$$P(B) = \frac{|B|}{|\Omega|} = \frac{6!}{6^6} = 0.015\,.$$

This is a very small number because we are looking at a very rare, unlikely event.

Besides the question how many permutations of a set there are, we can also deal with similar questions related to combinatorics.

Definition 1.16: Binomial Coefficient

The *binomial coefficient* $\binom{n}{k}$ is the quantity that answers the question "how many *different* ways are there to choose k elements out of a total number of n elements" (which is usually phrased "n choose k"). An explicit expression is

$$\binom{n}{k} = \frac{n!}{k!\,(n-k)!}\,. \tag{1.14}$$

Proof. The basic idea is to solve a simpler problem involving the binomial coefficient. This problem is how many k-element ordered lists one can make, which turns out to be precisely $\frac{n!}{(n-k)!}$. On the other hand, there are $\binom{n}{k}$ ways to choose k different lists out of these n elements. For each of these lists, there are $k!$ different ways to order them, i.e. $\binom{n}{k} k! = \frac{n!}{(n-k)!}$. □

As always when deriving and defining a new quantity, some sanity checks should be done. These include the cases $k = n$, where we expect $1 = \binom{n}{n}$ (one way to choose all elements), and $k = 0$, where we expect $1 = \binom{n}{0}$ as well (since the only choice is $\emptyset$, which still forms an element of the sample space and thus requires to be "chosen"). Both are indeed fulfilled.

Properties of the binomial coefficient can be derived using algebra (messy and lengthy) or using the definition and our intuition about it. An example is that the number of subsets that can be created using n elements is known to be 2^n. But this is also exactly what the sum over the number of all k-element subsets out of n objects represents, so we have "derived"

$$\sum_{k=0}^{n} \binom{n}{k} = 2^n \,. \tag{1.15}$$

A more general version of the binomial coefficient is

$$\frac{N!}{\prod_i^p n_i!} \,. \tag{1.16}$$

It answers the question in how many ways N elements can be divided into p sets with cardinality n_i (note: the n_i have to sum up to N). This definition reduces to $\binom{n}{k}$ for $p = 2$ and cardinalities $k, N - k$, which represent the two sets of chosen, not chosen elements.

Example 1.17: Two Heads in Ten Coin Flips

A more complex scenario is tossing a coin ten times (independently) with someone telling us that heads occurred three times (without us having seen the outcomes). What is the probability that heads occurred in the first two throws?

Mathematically, this is a conditional probability $P(\{S = HH\ldots\}|\{\#\{H \in S\} = 3\})$ (S denotes the sequence of results of the ten coin flips), so we have to count in the conditional universe. The number of different ways to get three H is simply $\binom{10}{3}$. Considering the case where the first two slots are H, there is one H left to distribute over eight slots, which means there are eight possibilities. In the end,

$$P(\{S = HH\ldots\}|\{\#\{H \in S\} = 3\}) = \frac{|\{S = HH\ldots\} \cap \{\#\{H\} \in S = 3\}|}{|\{\#\{H\} \in S = 3\}|} = \frac{8}{\binom{10}{3}} = \frac{1}{15} \,.$$

Example 1.18: Four Aces for Four Players

The task in this example is to compute the probability that each of four players gets an ace (four of which exist in a deck of 52 cards).

First of all, it is important to know in how many different ways one can distribute the aces and this is $4!$ (by definition of the factorial; this is different from previous example because the aces are distinguishable, whereas H tosses are not). The number of ways the

48 remaining cards can be distributed over the remaining twelve spots for each player is

$$\binom{48}{12} \cdot \binom{36}{12} \cdot \binom{24}{12} \cdot \binom{12}{12} = \frac{48!}{(12!)^4} .$$

Similarly, the total number of ways that 13 cards can be distributed over four players is

$$\binom{52}{13} \cdot \binom{39}{13} \cdot \binom{26}{13} \cdot \binom{13}{13} = \frac{52!}{(13!)^4} .$$

Therefore, the probability of $A = \{\text{each player gets an ace}\}$ is

$$P(A) = \frac{|A|}{|\Omega|} = \frac{4! \cdot \frac{48!}{(12!)^4}}{\frac{52!}{(13!)^4}} = \frac{4! \cdot 48! \cdot 13^4}{52!} = \frac{4! \cdot 13^4}{52 \cdot 51 \cdot 50 \cdot 49} = 0.105 .$$

2 Random Variables

2.1 What is a RV?

Until now, we only talked about events and their probabilities. However, for the majority of problems, events are not numbers but rather encode more abstract outcomes like heads or tails for coin flips. A more formal way to assign probabilities would use functions, but they take numbers as an input. Therefore, we need rules mapping outcomes and thus events to numbers. This is what *random variables* (RVs) do, which are defined as functions $X : \Omega \to \mathbb{R}$ that assign a number to each outcome in the sample space. The values a RV X can take will be denoted as x ($\in \mathbb{R}$) and are called *realizations*. A very natural question is how likely a value x of a random variable X is to be realized. One can quantify that by collecting the outcomes that result in x and adding their probabilities, i.e.

$$P_X(x) := P(X = x) := P(\{\omega \in \Omega : \ X(\omega) = x\}) = \sum_{\omega \in \Omega:\ X(\omega)=x} P(\{\omega\})\,. \tag{2.1}$$

P_X is called *probability mass function* (PMF) in case of discrete RVs (that map from and to a finite number of x's) and we will use p_X synonymously to P_X. More generally, this is also what the notion of a *probability distribution* (or simply: distribution) of a RV refers to and it is based on the distribution P that belongs to the sample space. The distribution of a RV also inherits some properties of P:

$$P_X(x) \geq 0 \qquad \sum_x P_X(x) = 1\,. \tag{2.2}$$

One advantage of having defined the notion of a RV is that it enables to assign more general probabilities, in a sense it allows to answer the question "what is the probability that a person from a group weighs 60kg" rather than "what is the probability that person 1 weighs 60kg" (this is much more similar to "how many persons weigh 60kg" converted to a probability). However, it should be noted that since the explicit map used for this assignment X is arbitrary, the results obtained after calculations still require an interpretation.

For a fixed sample space, many RVs exist (taking Ω to be a group of people, RVs would be their height, weight, color of their eyes, ...). In fact, one can build new RVs from existing ones by looking at functions $f : \mathbb{R} \to \mathbb{R}$ that take their results as arguments, i.e. $f \circ X : \Omega \to \mathbb{R}$.

Example 2.1: Four-faced Die

Now we can reformulate the treatment of experiments like the four-faced die that is tossed twice (see example 1.5). Defining the RV $X = \min(\text{first toss}, \text{second toss})$, we can calculate in very convenient notation:

$$P_X(2) = \frac{5}{16}\,,$$

which still corresponds to the green part of table 1.3.

Example 2.2: Coin Flip

We will now treat the very simple experiment of a coin flip with two outcomes H, T and $P(H) = p \Rightarrow P(T) = 1 - p$. We may be interested in the question how often we have to flip the coin in order to get a head H. The corresponding RV X is basically just a counter, it maps a series of flips (which always end at first H, are cut off there) as follows:

$$X(T^{k-1}H) = k\,.$$

For example, $X(H) = 1, X(TH) = 2, X(TTTTH) = 5$.

The corresponding PMF is rather simple to calculate because different values k of X can be realized in only one way. Consecutive coin flips are independent, so

$$P_X(k) = p(1-p)^{k-1}\,.$$

This is the *geometric PMF* because its values form a geometric series, i.e. they fulfil

$$\frac{P_X(k+1)}{P_X(k)} = 1 - p = \text{const}\,.$$

Example 2.3: Coin Flip 2

Instead of looking for the first H, we can also look at the whole series of outcomes of continuous tosses and compute the probability for a specific outcome. Using $P(H) = p$ again, it should be clear that the probability of such a series H^iT^j to occur is

$$P(H^iT^j) = p^i(1-p)^j\,.$$

Since consecutive coin tosses are independent events, it does not make too much sense to assign probabilities to a specific series like H^iT^j, but rather to the number k of H in it. This number forms a RV X and denoting the series length with n, there are $\binom{n}{k}$ series' producing the same realization k of Y (by definition of the binomial coefficient, we are interested in *different* permutations of H^kT^{n-k}, not all). Hence,

$$P_X(k) = \sum_{\{H,T\}^{\times n}:\, \#H=k} p^k(1-p)^{n-k} = \binom{n}{k} p^k(1-p)^{n-k}\,.$$

This is the *binomial distribution*.

Definition 2.4: Expectation

The *expected value/expectation* of a RV is

$$E[X] = \sum_x x P_X(x)\,. \tag{2.3}$$

This is essentially a weighted average of all realizations of a RV and gives us a statement about the realizations that are most likely to occur. There is no analogue of this for events in sample spaces because it does not make sense to speak of half heads. In the interpretation of probability as mass, the expectation is the center of mass.

Example 2.5: Wheel of Fortune

Depending on the experiment, the expectation as a weighted average has different meanings. We now treat the example of a wheel of fortune (figure 2.1), where we get the reward that the thick black line stops on after spinning the wheel. In this case, the probabilities for each reward X are nothing but its fraction of area occupied and thus

$$E[X] = \frac{1}{6} \cdot 1\$ + \frac{1}{2} \cdot 2\$ + \frac{1}{3} \cdot 4\$ = \frac{15}{6}\$ = 2.5\$\,.$$

Since $P_X(x)$ can be interpreted as a frequency of occurrence, $2.5\$$ is the average reward we will get from repeatedly spinning the wheel.

When thinking of the expectation as an average, we can think of many interesting quantities to compute it for. A simple idea, which is complementary to the expectation itself, is the average deviation of realizations from their expectation ≡ average. However, measuring linear deviations turns out to make no sense because

$$E[X - E[X]] = E[X] - E[X] = 0\,.$$

When looking for a quantity that is not trivially vanishing, it makes sense to go to quadratic deviations, which leads to the *variance*

$$\operatorname{var}(X) := E[(X - E[X])^2] = E[X^2] - E[X]^2\,. \tag{2.4}$$

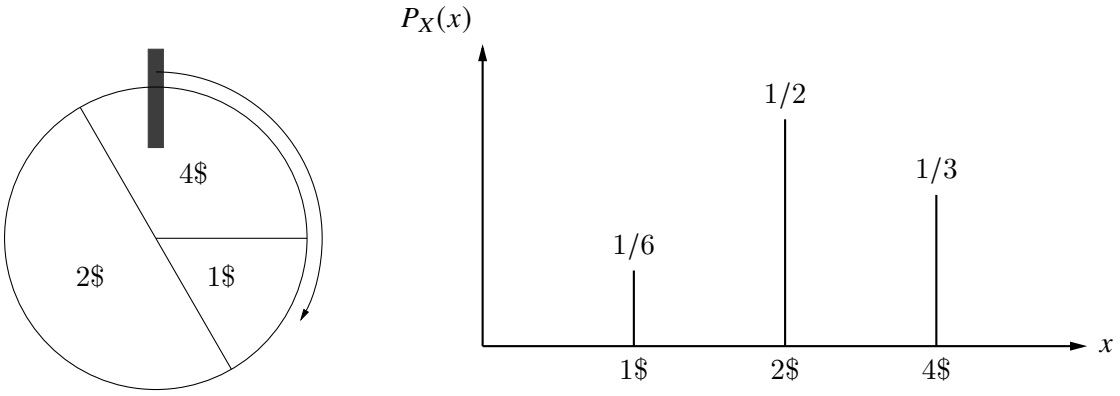
Property 2.6: Linearity of Expectation, Variance

$$E\left[\sum_i a_i X_i + b\right] = \sum_i a_i E[X_i] + b \qquad \operatorname{var}(aX + b) = a^2 \operatorname{var}(X) \tag{2.5}$$

Expected values are widely used as simple quantities to shrink down the information contained in a probability distribution into numbers. Of course, that means we discard information, but it is very helpful as a rough overview (e.g. if the distribution is unknown and only samples are available). For more information, higher order *moments* can be computed.

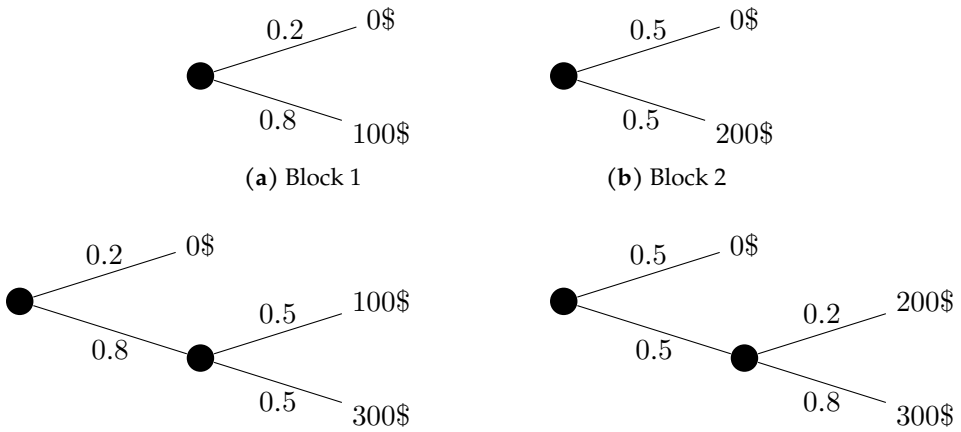
Example 2.7: Maximum Reward Strategy

We will now illustrate how expected values can be used to make decisions and even choose strategies. The setting we choose is visualized in figure 2.2. There are two blocks with different probabilities of success/failure and different rewards, which we go through one after the other. We want to find out which order of blocks maximizes



(**a**) Sketch (**b**) PMF

Figure 2.1: Illustrations for Wheel of Fortune



(**a**) Block 1 (**b**) Block 2

(**c**) Strategy 1 (**d**) Strategy 2

Figure 2.2: Sequential diagram of blocks and strategies that can be built from them

the average reward, i.e. whether to choose strategy 1 or 2. An additional rule is that in case of no reward in the first stage, there is no second stage.

To do that, we use the definition of the expected value and sum up the rewards from all outcomes/paths multiplied with the probability of this path (which is merely the product of the probabilities on this path):

$$\begin{aligned} E_{S1}[\text{reward}] &= 0.2 \cdot 0\$ + 0.8 \cdot 0.5 \cdot 100\$ + 0.8 \cdot 0.5 \cdot 300\$ \\ &= 0.4 \cdot 100\$ + 0.4 \cdot 300\$ = 160\$ \\ E_{S2}[\text{reward}] &= 0.5 \cdot 0\$ + 0.5 \cdot 0.2 \cdot 200\$ + 0.5 \cdot 0.8 \cdot 300\$ \\ &= 0.1 \cdot 200\$ + 0.4 \cdot 300\$ = 140\$\,. \end{aligned}$$

That means we should rather play it safe in the first stage to make sure we get at least some reward and not take the 50% risk of getting no reward.

RVs are still defined on a sample space Ω. That means we should be able to condition them on events A. This is indeed possible because elements of the sample space $\Omega' = A$ can still be mapped after conditioning.[7] From the definition in equation (2.1) we can also make sense of the distribution of such a RV,

$$P_{X|A}(x) := P'(\{\omega' \in \Omega' :\ X(\omega') = x\}) = \sum_{\omega' \in \Omega':\ X(\omega')=x} P'(\{\omega'\}) = \sum_{\omega \in \Omega:\ X(\omega)=x} P(\{\omega\}|A)\,. \quad (2.6)$$

Thus, we can define an expectation using this conditional PMF, the *conditional expectation*

$$E[X|A] = \sum_x x P_{X|A}(x)\,. \quad (2.7)$$

It has the same properties as "regular" expectations $E[X]$.

Property 2.8: Summation Rules

For a partition $\{A_i\}$ of the sample space,

$$P_X(x) = \sum_i P(A_i) P_{X|A_i}(x) \qquad E[X] = \sum_i P(A_i) E[X|A_i]\,. \quad (2.8)$$

These are the total probability rule for RVs and the *total expectation theorem* (a simple corollary of the former). Note that the conditioning is on events, not other RVs.

Example 2.9: Memorylessness

The total expectation theorem may seem a little abstract, but we will now show a very explicit example where it is helpful. In example 2.2 we introduced the geometric PMF

$$P_X(k) = p(1-p)^k\,,$$

[7] All others have probability zero, so they are not relevant for P'.

which describes the probability of getting H in the k-th flip (not at least one, precisely in the k-th). However, we have not computed any properties of it yet, e.g. the expectation

$$E[X] = \sum_{k=1}^{\infty} kp(1-p)^k \,.$$

A convenient way to compute it while avoiding the evaluation of this series, is to once again use the fact that consecutive coin flips are independent. If we know there is no H in the first toss, we can also look at the RV $Y = X - 1 | X > 1$. The distribution of this Y is

$$P_Y(k) = p(1-p)^k = P_X(k)$$

because the probability of a single coin toss does not depend on the number of previous tosses – geometric RVs are *memoryless*. A very illustrative example would be two persons flipping coins, but one starting later. For each flip, they have the same probabilities of getting H, T. Therefore, although they have different counters, they have the same probability to get H in the k-th flip (*their* respective k-th flip). More formally, we can define $A_1 = \{X = 1\}, A_2 = \{X > 1\}$ (first toss is H, T). $A_1 \cup A_2$ includes all outcomes, so

$$E[X] = P(A_1)E[X|A_1] + P(A_2)E[X|A_2] = pE[X|A_1] + (1-p)E[X|A_2] \,.$$

Since $P_{X|A_1}(1) = 1$, the first term equals p. To simplify the second term, we rewrite:

$$\begin{aligned} E[X|A_2] &= E[X|X > 1] = E[X - 1 + 1 | X - 1 > 0] \\ &= E[X - 1 | X - 1 > 0] + 1 = E[X] + 1 \,. \end{aligned}$$

The last equality directly follows from the memorylessness of a geometric RV, as we can see by writing out the expected value:

$$E[X - 1 | X - 1 > 0] = \sum_{k=1}^{\infty} kP_{X-1|X-1>0}(k) = \sum_{k=1}^{\infty} kP_X(k) = E[X] \,.$$

Therefore

$$E[X] = p + (1-p)(E[X] + 1) \quad \Leftrightarrow \quad E[X] = \frac{1}{p} \,,$$

which should be very intuitive since it expresses the average number of tosses to get H for the first time. This will be small if $p = P(H)$ is very high (and vice versa).

2.2 Multiple RVs

We will now generalize our discussions and allow for more than one RV, starting with two of them. For such a pair of RVs X, Y we can also assign probabilities of simultaneous occurrence of realizations x, y. In principle, this is analogous to $P(A \cap B)$ for events A, B, but the *joint*

probability of two RVs will be denoted a bit differently as

$$P_{X,Y}(x,y) := P(X = x \text{ and } Y = y) = \sum_{\omega\in\Omega:\ X(\omega)=x \text{ and } Y(\omega)=y} P(\{\omega\})\,. \tag{2.9}$$

After all, an intersection of realizations is not what we intend to measured here. This joint likelihood is a regular PMF in the sense that $P_{X,Y}(x,y) > 0,\ \forall x,y$ and $\sum_{x,y} P_{X,Y}(x,y) = 1$. A very convenient way to visualize joint PMFs of discrete RVs is a tabular representation like 2.1 that has been used for sample spaces previously. Instead of probabilities P in the sample space, the entries are now values assigned by $P_{X,Y}$.

We will now see how $P_{X,Y}$ is related to the *marginal distributions* P_X, P_Y. To obtain the PMF of only one RV, we can simply remove the information about the other RV by summing over all of its values, i.e.

$$P_X(x) = \sum_y P_{X,Y}(x,y) \qquad P_Y(y) = \sum_x P_{X,Y}(x,y)\,. \tag{2.10}$$

This can be visualized very nicely in the tabular representation (see e.g. table 2.1). Another viewpoint for this marginal PMF is that $y = 2$ may occur for $x = 1,2,3,4$ and if we only care about the probability of y, then we have to consider all these x-values (which is done by summing over them; in table 2.1, this leads to $P_Y(2) = \frac{1}{4}$ as the sum over the green row).

There is also another way to construct a function of only one RV from $P_{X,Y}(x,y)$: by fixing a value of the other RV. This leads to the *conditional PMF* (conditioned on RV, not event)

$$P_{X|Y=y}(x|y) = \frac{P_{X,Y}(x,y)}{P_Y(y)} = \frac{P_{X,Y}(x,y)}{\sum_x P_{X,Y}(x,y)}\,.^{8} \tag{2.11}$$

This definition is analogous to the one for events, but we can also see in a different way that it has to be like that: taking only $P_{X,Y}(x,y)$ for some fixed value $Y = y$ would not produce a "real" PMF because this function does not fulfil the closure rule. To normalize it properly, we divide by all values it can take for this fixed y, i.e. $\sum_x P_{X,Y}(x,y)$. In table 2.1, the conditional PMF $P_{X|Y=2}(x|y=2)$ is colored green (up to normalization, which can be achieved by multiplying the probabilities with a factor $\frac{1}{\sum_x P_{X|Y=2}(x|y=2)} = \frac{1}{5/20} = 4$).

Since $P_{Y|X=x}(y|x)$ is defined in the same manner as $P_{X|Y=y}(x|y)$, we also know how to construct the joint PMF from marginal and conditional PMFs. Some rearranging yields the already well-known relation

$$P_{X,Y}(x,y) = P_X(x)P_{Y|X}(y|x) = P_Y(y)P_{X|Y}(x|y)\,. \tag{2.12}$$

Additionally, as we might guess from conditional probabilities of events, the joint PMF naturally leads to the notion of *independence* of RVs. And just like before, while

$$P_{X|Y=y}(x|y) = P_X(x) \qquad P_{Y|X=x}(y|x) = P_Y(y) \tag{2.13}$$

[8] Of course, this is only defined for $P_Y(y) \neq 0$. However, conditioning does only make sense in this case since we assume y was realized.

y				
4	$\frac{1}{20}$	$\frac{2}{20}$	$\frac{2}{20}$	0
3	$\frac{2}{20}$	$\frac{4}{20}$	$\frac{1}{20}$	$\frac{2}{20}$
2	0	$\frac{1}{20}$	$\frac{3}{20}$	$\frac{1}{20}$
1	0	$\frac{1}{20}$	0	0
	1	2	3	4 x

Table 2.1: PMF $P_{X,Y}(x,y)$ for RVs X,Y that take values $x,y \in 1,2,3,4$. In yellow, we show the values the conditional PMF $P_{X,Y|A}(x,y)$ where $A = \{(x,y) : \ x \leq 2, y \geq 3\}$ can take (although the values are not normalized correctly, 20 has to be replaced with 9). In green, we show $P_{X|Y=2}(x|y=2)$ and also why $P_Y(y) = 1/4$.

may be more intuitive to demand, this is only defined for $P_Y(y) \neq 0, P_X(x) \neq 0$, respectively. A more robust requirement is factorization of the joint PMF into marginals, i.e.

$$P_{X,Y}(x,y) = P_X(x)P_Y(y)\,. \tag{2.14}$$

Using this condition, we can see that the RVs in table 2.1 are not independent, e.g.

$$P_X(3)P_Y(2) = \left(\frac{2}{20} + \frac{1}{20} + \frac{3}{20}\right) \cdot \left(\frac{1}{20} + \frac{3}{20} + \frac{1}{20}\right) = \frac{6}{20} \cdot \frac{5}{20} = \frac{3}{40} \neq \frac{3}{20} = P_{X,Y}(3,2)\,.$$

However, one can verify that conditioning on the event $A = \{(x,y) : \ x \leq 2, y \geq 3\}$ (yellow part of table 2.1) produces independent RVs with joint PMF $P_{X|A,Y|A} = P_{X|A}P_{Y|A}$.

Having developed the formalism for two RVs, we can now generalize it to an arbitrary number k of them. For example, in case of $k = 3$, the condition for independence would read $P_{X,Y,Z}(x,y,z) = P_X(x)P_Y(y)P_Z(z)$, conditional PMFs $P_{X|Y,Z}(x|y,z)$ etc. The most general versions are analogues of the summation/total probability and multiplication rule:

$$P_{X_i}(x_i) = \sum_{x_1,\ldots x_{i-1},x_{i+1},\ldots,x_k} P_{X_1,\ldots,X_k}(x_1,\ldots,x_k) \tag{2.15}$$

$$P_{X_1,\ldots,X_k}(x_1,\ldots,x_k) = P_{X_1}(x_1)P_{X_2|X_1}(x_2|x_1)\ldots P_{X_k|X_1,\ldots,X_{k-1}}(x_k|x_1,\ldots,x_{k-1})\,. \tag{2.16}$$

The former is nothing but the definition of marginal PMFs.

2.3 Functions of RVs

Often, it is possible to measure certain RVs like height or weight, but the quantity we are really interested in is a combination of them, for example the BMI $= \frac{\text{weight}}{\text{height}^2}$. For a mathematical description of this, we have to consider RVs $Z = g(X,Y)$ obtained from two RVs X,Y via a

mapping $g : \mathbb{R}^2 \to \mathbb{R}$. The corresponding PMF is

$$P_Z(z) = \sum_{\omega\in\Omega:\ Z(\omega)=g(X(\omega),Y(\omega))=z} P(\{\omega\}) = \sum_{(x,y):\ g(x,y)=z} P_{X,Y}(x,y)\,. \tag{2.17}$$

Hence the expectation reads

$$E[Z] = \sum_z zP_Z(z) = \sum_{g(x,y)} g(x,y) \sum_{(x,y):\ g(x,y)=z} P_{X,Y}(x,y) = \sum_{x,y} g(x,y)P_{X,Y}(x,y)\,. \tag{2.18}$$

This is the *law of the unconscious statistician*. It implies that in general,

$$E[g(X,Y)] \neq g(E[X],E[Y])\,. \tag{2.19}$$

Exceptions from this are linear functions like $g(X,Y) = aX + bY + c$ and also $g(X,Y) = XY$ for independent X, Y.

Proof. We calculate:

$$\begin{aligned} E[aX + bY + c] &= \sum_{x,y}(ax + by + c)P_{X,Y}(x,y) \\ &= a\sum_{x,y} xP_{X,Y}(x,y) + b\sum_{x,y} yP_{X,Y}(x,y) + c\sum_{x,y} P_{X,Y}(x,y) \\ &= a\sum_x x\sum_y P_{X,Y}(x,y) + b\sum_y y\sum_x P_{X,Y}(x,y) + c \\ &= a\sum_x xP_X(x) + b\sum_y yP_Y(y) + c = aE[X] + bE[Y] + c\,. \end{aligned}$$

$$\begin{aligned} E[XY] &= \sum_{x,y} xyP_{X,Y}(x,y) = \sum_{x,y} xyP_X(x)P_Y(y) \\ &= \sum_x xP_x(x)\sum_y yP_Y(y) = \sum_x xP_X(x)E[Y] = E[X]E[Y]\,. \end{aligned}$$

□

It is even possible to prove the extension (still assuming independent X, Y)

$$E[g(X)h(Y)] = E[g(X)]E[h(Y)]\,. \tag{2.20}$$

We also note that the same relations can be proven for conditional expectations.

2.4 Continuous RVs

Until this point, we have dealt with RVs that map from and to a discrete number of values. However, real world problems often involve sample spaces and RVs with a continuous range of values and thus infinitely many of them. Even many examples we saw (like height or

weight) are of this kind, at least conceptually (in practice, we may sometimes treat them discrete by measuring them to some finite accuracy). If we want to describe them, we have to take another approach than sums for PMFs and expectations. The reason can be seen from the simple example $P(a \leq X \leq b)$: for a discrete RV,

$$P(a \leq X \leq b) = \sum_{x \in [a,b]} P_X(x)\,. \tag{2.21}$$

In the continuous case however, intervals contain an infinite number of points x_i and if these all have a non-zero probability $P_X(x_i) > 0$,

$$P(a \leq X \leq b) = \sum_{x \in [a,b]} P_X(x) = \sum_{i=1}^{\infty} P_X(x_i) = \infty\,. \tag{2.22}$$

This is a (unsolvable!) violation of the closure rule. To fix this problem, we have to move on from assigning probabilities to points and instead assign them to intervals. A tool that naturally assigns numbers to intervals and is also the natural limit of sums is the integral. Consequently, we write the probability of some range of values $S = [a, b] \subset \mathbb{R}$ being realized by a continuous RV (CRV) X as

$$P_X(S) = \int_a^b f_X(x)\,dx = \int_{X^{-1}S} dP = P(\{\omega \in \Omega :\ a \leq X(\omega) \leq b\}) =: P(a \leq X \leq b)\,. \tag{2.23}$$

These are different ways in which probabilities involving CRVs can be computed, either using its distribution P_X or the distribution P on the sample space. $P(a \leq X \leq b)$ is just a general way to denote them. For a single point $S = \{x\}$, $P_X(\{x\}) = 0$ because points are null sets, $\int_x^x f_X(x')\,dx' = 0$. The object that assigns non-zero values to points $x \in \mathbb{R}$ now is the function $f_X(x)$, which is called *probability density function* (PDF) or just density. It may be thought of as a non-normalized probability of points. In the discrete case, density and probability are equal (the PMF P_X assigns probabilities to points x)[9], whereas for CRVs

$$f_X(x) = \frac{P(x \leq X \leq x + dx)}{dx}, \quad dx \ \text{ small}\,. \tag{2.24}$$

In contrast to PMFs, the interval length/distance of points dx has to be taken into account for proper normalization of PDFs. Symbolically, by setting $dx = 1$, we retrieve the claim

$$f_X(x) = \frac{P(x \leq X < x + 1)}{1} = P(X = x) = P_X(x) \tag{2.25}$$

about discrete RVs (it has been assumed that X takes values on $\mathbb{N}$, so $x_{i+1} - x_i = 1$).[10]

We can now introduce certain quantities known from discrete RVs and also state properties that have to hold in order for the axioms of probability theory to be fulfilled.

[9] In the language of measure theory, discrete RVs are just a special case of CRVs occurring in discrete sample spaces, where the counting measure and hence sums are used.

[10] Here, we use a definition/convention where the point $x + dx$ is excluded in order to have this equivalence. This does not change the probability in the CRV case because points are null sets.

Property 2.10: Properties of CRVs

1. Non-negativity: $f_X(x) \geq 0, \forall x$
2. Closure rule: $\int_{-\infty}^{\infty} f_X(x)\,dx = 1$
3. Expectation: $E[X] = \int_{-\infty}^{\infty} x f_X(x)\,dx$
4. Variance: $\mathrm{var}(X) = \sigma_X^2 = \int_{-\infty}^{\infty}(x - E[X])^2 f_X(x)\,dx = E[X^2] - E[X]^2$
5. Law of the unconscious statistician[11]: $E[g(X)] = \int_{-\infty}^{\infty} g(x) f_X(x)\,dx$

Clearly, PMF and PDF have analogous roles. It is even possible to further extend this connection by using *cumulative distribution functions* (CDFs)

$$F_X(x) := P(X \leq x) = \begin{cases} \sum_{x' \leq x} P_X(x'), & X \text{ discrete} \\ \int_{-\infty}^{x} f_X(x')\,dx', & X \text{ continuous} \end{cases} . \tag{2.26}$$

Although the CDF still does not assign values to every point $x \in \mathbb{R}$ for discrete RVs, in a visualization it can look very much like a continuous function (see figure 2.3).

Furthermore, from $f_X(x) \geq 0$ we see that F_X is a monotonically increasing function and it also fulfils (assuming $a \leq b$)

$$P(a \leq X \leq b) = \int_a^b f_X(x)\,dx = \int_{-\infty}^{b} f_X(x)\,dx - \int_{-\infty}^{a} f_X(x)\,dx = F_X(b) - F_X(a)\,. \tag{2.27}$$

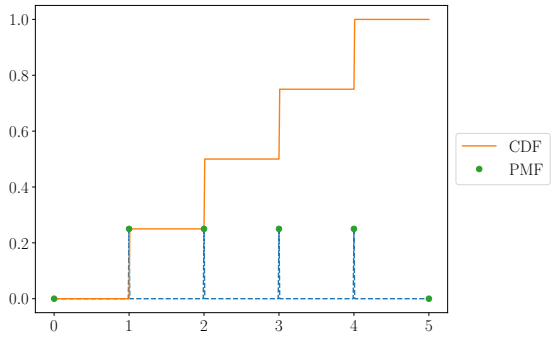
Both properties translate to the discrete case as well because $P_X(x) \geq 0$ and sums can be split accordingly. In the continuous case, combining them with equation (2.24) yields

$$f_X(x) = \frac{F_X(x+dx) - F_X(x)}{dx} = \frac{F_X(x+dx) - F_X(x)}{x + dx - x} \underset{dx \to 0}{=} \frac{dF_X(x)}{dx}\,. \tag{2.28}$$

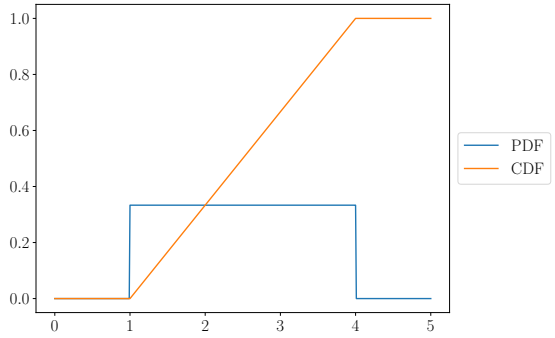
For discrete RVs with $f_X = P_X$, this relationship only holds in points where F_X is defined. Still, it makes sense to write the relationship in this way since P_X is defined only in these points as well. Hence, we have found a quantity that shows no conceptual differences between discrete/continuous RVs and is, in principle, sufficient to describe them.

We will now put into perspective what happened in the previous paragraphs and how we ended up with the notion of a density. Basically, we explained how sums do not work for CRVs, which motivates the replacement $\sum P \to \int dP$. In (2.23), we have written this integral in terms of $f_x dx = P(x \leq X \leq x+dx)$ at first, dP on Ω was only used after that. Mathematically,

[11] This is a theorem of profound importance and it is not as straightforward as it looks at first glance (in particular, it does *not* hold by definition of the expectation). It states $E_Y[Y] = \int y f_Y(y)\,dy = \int g(X) f_X\,dx = E_X[g(X)]$ (where we write $Y = g(X)$), which follows from the transformation rule in measure theory.



(**a**) Uniform PMF and corresponding CDF



(**b**) Uniform PDF and corresponding CDF

Figure 2.3: Comparison of discrete and continuous uniform distributions and cumulative distributions. While the continuous function assigns values to all points in $[0, 5]$, the discrete PMF only assigns values to the integers $0, 1, 2, 3, 4, 5$. The blue dotted line corresponds to a continuous version of that where all other probabilities are set to be zero. Strictly speaking, the discrete CDF would also only assign values to these integers, but to make the similarity more obvious it is plotted on the whole interval $[0, 5]$ (which the orange line is the CDF of the blue, dotted line).

this transition can be made thanks to the concept of pushforward measure $d(X_*P)$ (i.e. using measure theory), by applying the transformation rule and rewriting $d(X_*P) = dF_X = f_X dx$. This is why another function f_X suddenly shows up in addition to P_X. Since the PDF is an object determining probabilities, expectations and therefore basically all other objects of interest, the theory of CRVs mostly deals with finding and manipulating PDFs. To make ourselves more familiar with them, we will now treat important examples.

Example 2.11: Important Distributions

- *Uniform distribution*: the simplest distribution one can think of is a constant one, $f_X(x) = k$ on some interval $[a, b]$. For an explicit expression, we can calculate k:

$$1 = \int_a^b f_X(x)\,dx = \int_a^b k\,dx = k(b-a) \quad \Leftrightarrow \quad f_X(x) = \frac{1}{b-a} =: \mathcal{U}(a,b)\,. \tag{2.29}$$

  This also shows that $-\infty < a < b < \infty$ is a necessary condition, otherwise the closure rule could not be fulfilled. By evaluating the corresponding integrals, we obtain the following properties of a uniform RV X:

$$E[X] = a + \frac{b-a}{2} = \frac{a+b}{2} \qquad \mathrm{var}(X) = \frac{(b-a)^2}{12}\,. \tag{2.30}$$

- *Exponential distribution*: a very practical application of statistics is the detection of radioactive particles that decay after a certain (average) waiting/lag time τ, e.g. using a Geiger-Müller counter. More abstractly, we can describe this problem

as a repeated process with the same statistical properties. From the corresponding differential equation, one can derive the following distribution:

$$f_X(x) = \lambda e^{-\lambda x} \tag{2.31}$$

where $0 \leq x \leq \infty$ is the input parameter of the distribution and $\lambda > 0$.

Properties of an exponential RV X are

$$E[X] = \frac{1}{\lambda} \qquad \text{var}(x) = \frac{1}{\lambda^2}\,. \tag{2.32}$$

Therefore, we can interpret λ as a rate parameter that e.g. encodes when to expect the next radioactive particle in case of the Geiger-Müller counter (it measures number of events per time and is related to the lag time via $\tau = \frac{1}{\lambda}$, so a value of $\lambda = 0.1$ means there is, on average, one event every 10 seconds). Also, this equality tells us that we can *infer* (an estimate of) λ from the mean of repeated measurements of the RV (corresponds to expectation for them).

- *Poisson distribution*: originates from similar idea to exponential, but instead of dealing with waiting time between events and assigning probabilities to them, it deals with the expected/average number of events in a given time interval.

  One can describe this mathematically by dividing the interval of length T into N smaller ones of length $\Delta t = \frac{T}{N}$ and then looking at the limit $N \to \infty$. The probability of observing exactly one event in T is the probability $\lambda\Delta t$ of observing it in a certain time interval Δt multiplied with the probability not to observe it in the other time intervals. Since there are N possible ways this can occur,

$$p_X(1) = N\lambda\Delta t(1-\lambda\Delta t)^{N-1} = \frac{\lambda T}{1-\lambda T/N}(1-\lambda T/N)^N\,.$$

  In the continuum limit $N \to \infty$, this becomes

$$p_X(1) = \lambda T e^{-\lambda T}\,.$$

  Similarly, the probability of observing two events during T is $\binom{N}{2}(\lambda\Delta T)^2(1-\lambda\Delta T)^{N-2}$ or $p_X(2) = \frac{1}{2}(\lambda T)^2 e^{-\lambda T}$ in case of $N \to \infty$. This can be generalized to

$$p_X(n) = \frac{1}{n!}(\lambda T)^n e^{-\lambda T} = \frac{1}{n!}\alpha^n e^{-\alpha}, \quad \alpha = \lambda T > 0,\ n \in \mathbb{N}\,. \tag{2.33}$$

  Properties of a Poissonian RV X are

$$E[X] = \alpha \qquad \text{var}(X) = \alpha\,. \tag{2.34}$$

  Since α is the expectation (makes sense, is rate × time span = number) and its width, it is also a parameter that moves the distribution on the x-axis.

- *Binomial distribution*: describes the outcomes of Bernoulli experiments, i.e. experiments with two outcomes $0, 1$ that have probabilities $p, q = 1 - p$, respectively. When performing N of these experiments, the probability to get outcome 0 the first n times and outcome 1 the next $N - n$ times is $p^n q^{N-n}$. Since there are $\binom{N}{n}$ ways to get n outcomes 0 in different orders, we have derived

$$p_X(n) = \binom{N}{n} p^n q^{N-n} \,. \tag{2.35}$$

  Note: n has to be an integer, this is a discrete distribution.

  Properties of a binomial RV X are

$$E[X] = Np \qquad \text{var}(X) = Npq \,. \tag{2.36}$$

  Again, we can see that N is a parameter that moves the distribution.

- *Gaussian/normal distribution*: also nice derivation, result is

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} =: \mathcal{G}(\mu, \sigma) = \mathcal{N}(\mu, \sigma) \,. \tag{2.37}$$

  Properties of a Gaussian RV X are

$$E[X] = \mu \qquad \text{var}(X) = \sigma^2 \,. \tag{2.38}$$

All of the distributions mentioned here are visualized in figure 2.4.

Although many of these distributions are important in the real world, there is one that clearly stands out and forms some kind of prime example of a PDF. This is the Gaussian distribution and we will now deal with it in more detail.

Example 2.12: Gaussian Distribution

The Gaussian/normal distribution

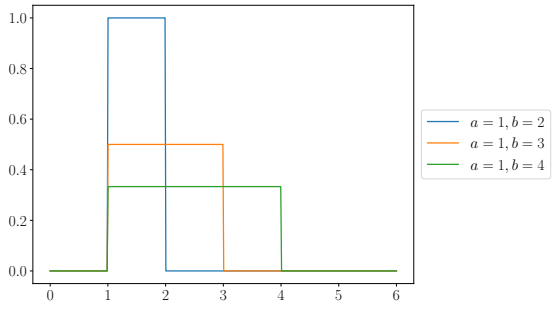
$$\mathcal{G}(\mu, \sigma) = \mathcal{N}(\mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

is particularly interesting because it has many convenient properties. One of them is that it describes a superposition of many random events or rather CRVs (in the limit of infinite trials), such as noise in experiments. A remarkable feature is that this holds independently of the distribution the CRVs have (see 2.8).

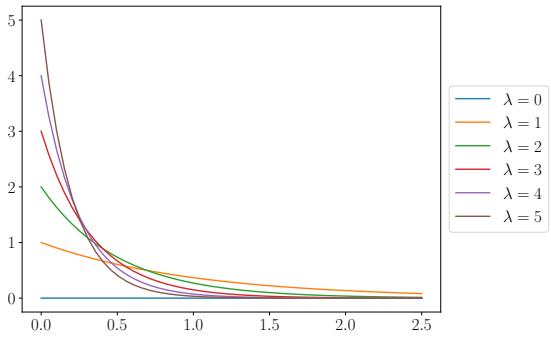
Moreover, the linear combination $Y = aX + b$ of a Gaussian CRV X is still Gaussian with

$$f_Y(y) = \mathcal{G}(a\mu + b, a^2\sigma^2) \,. \tag{2.39}$$

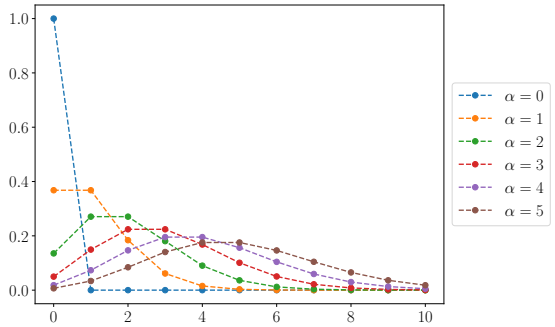
This is because the parameters μ, σ of a Gaussian are precisely its mean and variance, so



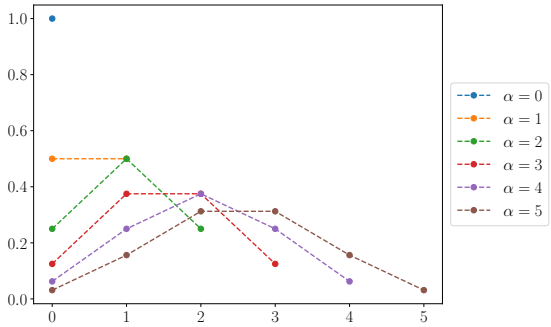
(**a**) Uniform distribution



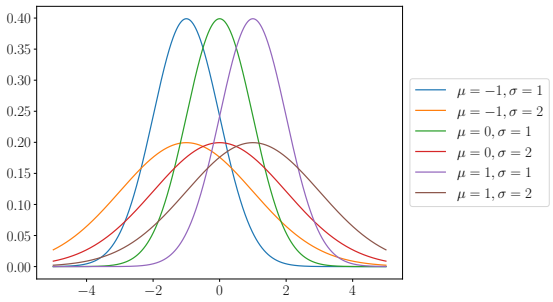
(**b**) Exponential distribution



(**c**) Poisson distribution



(**d**) Binomial distribution



(**e**) Gaussian distibution

Figure 2.4: Visualization of important PDFs

the law of the unconscious statistician (2.45) can be applied directly (gives whole shape in this case). This property is very useful because it allows to perform many calculations involving Gaussians only for the *standard normal distribution*

$$\mathcal{G}(0,1) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} . \tag{2.40}$$

Integrals for this function are mostly standard integrals, which have a general closed-form solution or tabulated values. While this might also be the case for other Gaussians, the calculations often involve much more algebra.

For example, say we want to compute $P(X \leq 3)$ for a Gaussian CRV X. This can be done rather quickly by rewriting $P(X \leq 3) = P\left(X' \leq \frac{3-\mu}{\sigma}\right)$ where $X' = \frac{X-\mu}{\sigma}$ is the "standardized version" of X. Since X' has a standard normal distribution, one can look at the tabulated results of $P\left(X' \leq \frac{3-\mu}{\sigma}\right)$ for the given μ, σ and thereby avoid tedious calculations.

We can now go on to treat other probabilities, which will also involve densities. All of this is possible in a very straightforward manner, so it will be done rather quickly.

Definition 2.13: Quantities for CRVs

- The *joint probability* of two CRVs X, Y is

$$P((x,y) \in S) = \int_S f_{X,Y}(x,y)\, dxdy \tag{2.41}$$

  where $f_{X,Y}$ is the *joint PDF* of X, Y.

- The corresponding *marginal PDFs* are

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)\, dy \qquad f_Y(y) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)\, dx\,, \tag{2.42}$$

  i.e. the joint PDF with one CRV "integrated out".

- The *conditional PDF* is

$$f_{X|Y}(x|y) = \frac{f_{X,Y}(x,y)}{f_Y(y)} \qquad f_{Y|X}(y|x) = \frac{f_{X,Y}(x,y)}{f_X(x)} \tag{2.43}$$

  and the corresponding probability can be obtained via integration.[12]

- Two CRVs are statistically *independent* if and only if

$$f_{X,Y}(x,y) = f_X(x) f_Y(y)\,. \tag{2.44}$$

- The *law of the unconscious statistician* is

$$E[g(X,Y)] = \int g(x,y) f_{X,Y}(x,y)\, dxdy\,. \tag{2.45}$$

  This also shows how to compute expectations like $E[X]$ using joint probabilities.[13]

Just like before, this can be generalized to multiple CRVs and from each PDF, an expectation can be computed (e.g. a conditional expectation $E[X|Y]$).

Although one can think of the marginals as projections/averages of the joint PDF, they do not have the same height (cf. figure 2.5). Intuitively, we can understand this from the fact that the joint PDF is normalized over the whole square $[-5,5]^2$, whereas the marginals are normalized over the interval $[-5,5]$. The same idea is true for $f_{X|Y}(x|y)$. It can be thought of as slices of the joint PDF $f_{X,Y}(x,y)$ for some fixed y, which are then normalized by dividing by all possible values $f_{X,Y}$ can take on this slice when varying x, i.e. $\int_{-\infty}^{\infty} f_{X,Y}(x,y)\, dx = f_Y(y)$.

Direct corollaries of the law of the unconscious statistician are (just like for discrete RVs)

$$\begin{aligned} E[aX + bY + c] &= aE[X] + bE[Y] + c \\ \operatorname{var}(aX + b) &= a^2 \operatorname{var}(X)\,. \end{aligned}$$

However, $\operatorname{var}(X+Y) = \operatorname{var}(X) + \operatorname{var}(Y)$ only holds if X, Y are independent. To see how the general version looks like, we have to deal with the notion of dependence and find a way to quantify it. Intuitively, a systematic relation between RVs should result in simultaneous deviations from their respective mean. Hence, the *covariance*

$$\operatorname{cov}(X,Y) = E[(X - E[X])(Y - E[Y])] = E[XY] - E[X]E[Y] \tag{2.46}$$

is a way to measure relations between RVs (see figure 2.6). Independence is equivalent to

$$\operatorname{cov}(X,Y) = E[XY] - E[X]E[Y] = E[X]E[Y] - E[X]E[Y] = 0\,. \tag{2.47}$$

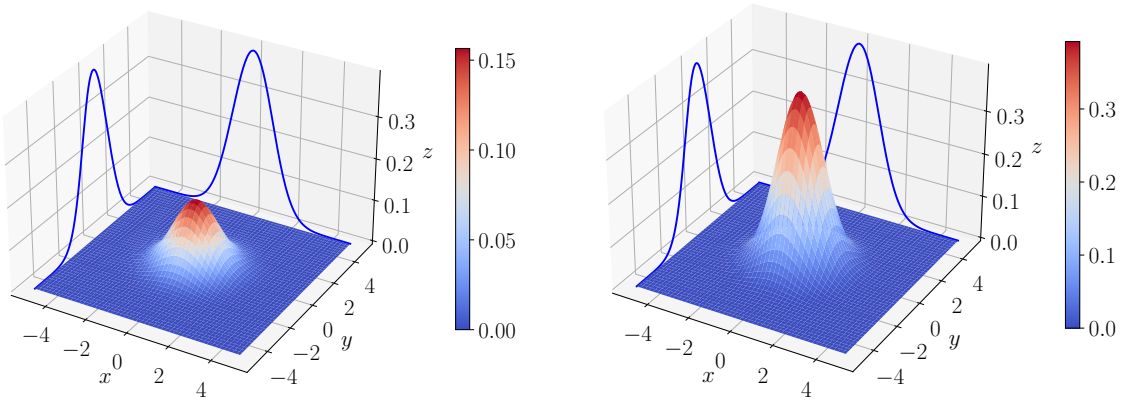
Covariance is also closely related to variance, not only by its name:

$$\operatorname{cov}(X,X) = \operatorname{var}(X)\,. \tag{2.48}$$

Furthermore, it is a linear quantity:

$$\begin{aligned} &\operatorname{cov}\left(\sum_i a_i X_i, \sum_i b_i Y_i\right) = \sum_{i,j} a_i b_j \operatorname{cov}(X_i, Y_j) \\ \Rightarrow \quad &\operatorname{var}\left(\sum_i a_i X_i\right) = \sum_i a_i^2 \operatorname{var}(X_i) + \sum_{i \neq j} a_i a_j \operatorname{cov}(X_i, X_j)\,, \end{aligned} \tag{2.49}$$

[12] Note that we do not demand $Y = y$ anymore because $P(Y = y) = 0$ (does not make sense to do that).
[13] $E[X] = E_{X,Y}[X]$ specifically reduces to $E_X[X]$, i.e. the one using f_X, because of (2.42).



(**a**) Normalized distributions

(**b**) Normalized marginals, scaled joint PDF

Figure 2.5: Visualization of joint PDF (colored) and marginals (blue lines) for independent, Gaussian CRVs. Normalized and non-normalized versions are shown.

which also explains why $\mathrm{var}(X+Y) = \mathrm{var}(X) + \mathrm{var}(Y)$ only holds for independent X, Y.

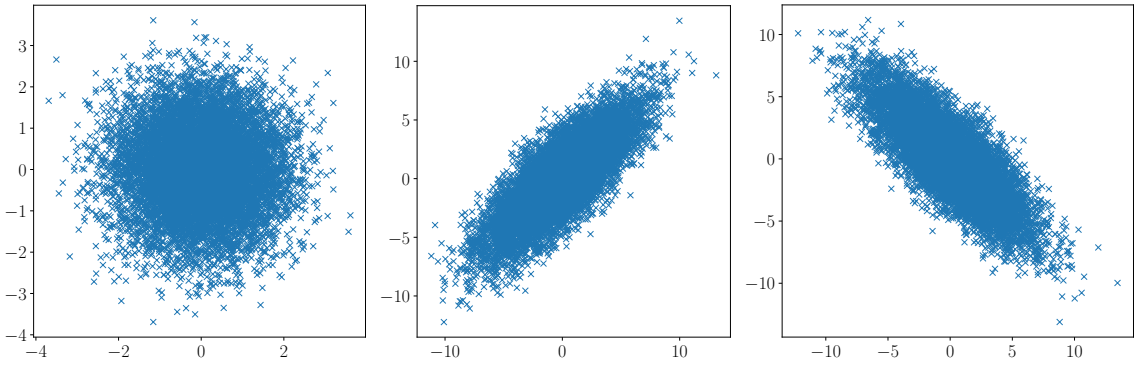
Nonetheless, the covariance also has issues. First, it contains the units of X, Y. If those are different, it may not make much sense to add X, Y in the first place. That also means its value depends on the representation of the data (e.g. meter or inch). Second, if a RV X has very high deviations from its mean, then its covariance with other RVs can be very large, no matter how strong the relation between them is. If these are issues that occur, a more useful and informative quantity is the *correlation (coefficient)*

$$\mathrm{corr}(X,Y) = \frac{\mathrm{cov}(X,Y)}{\sigma_X \sigma_Y} = \mathrm{cov}\left(\frac{X}{\sigma_X}, \frac{Y}{\sigma_Y}\right) = E\left[\frac{X - E[X]}{\sigma_X}\frac{Y - E[Y]}{\sigma_Y}\right]. \tag{2.50}$$

Basically, the idea is to express the RVs as their standardized versions and hence in deviations from the mean and in numbers of standard deviations. That allows to make statement which are independent of units and scales.

Now that many of the tools that probability theory knows are available, we will go through an example that shows how to apply them in detail. The standard procedure is as follows:

1. Setting up the sample space
2. Describing probability laws of the sample space
3. Identifying events of interest
4. Computing quantities of interest for those events (RVs, expectations, ...)



(a) Uncorrelated ($\mathrm{cov}(X,Y) = 0$) (b) Correlated ($\mathrm{cov}(X,Y) > 0$) (c) Anticorrelated ($\mathrm{cov}(X,Y) < 0$)

Figure 2.6: Samples from joint densities $f_{X,Y}$ for different relations between X, Y

Example 2.14: Buffon's Needle

Here we will give a mathematical description of a famous experiment, Buffon's needle. The idea is to randomly drop a needle of length L and see if it intersects with the boundaries of an area spanned by two lines that are separated by $d > L$ (see figure 2.7). We will strictly follow the standard procedure for this example.

1. At first, choose two RVs to describe this 2D scenario: the distance x between needle center and nearest line, as well as the angle θ between needle and lines. From the sketch of the experiment, we see that $0 \leq x \leq d/2$, $0 \leq \theta \leq \pi/2$.

2. Assuming the needle to drop at random, both RVs x, θ are uniformly distributed. Additionally, they are independent because distance and orientation are not related, so

$$f_{X,\Theta}(x,\theta) = f_X(x) f_\Theta(\theta) = \begin{cases} \frac{1}{d/2}\frac{1}{\pi/2}, & 0 \leq x \leq d/2,\ 0 \leq \theta \leq \pi/2 \\ 0, & \text{else} \end{cases} .$$

3. Our main interest (for reasons that we will see shortly) is whether or not the needle intersects one of the lines. Looking at the sketch again, we see that this is the case for

$$x \leq \frac{L}{2}\sin(\theta) .$$

4. To compute the probability of intersection, we use the condition for intersection that has just been derived and incorporate it into the integral boundaries. Since it is not necessary to have certain x- and θ-values at once (only their combination is important), we can let one parameter vary freely and use the condition to get boundaries for the

other parameter. In the end,

$$P(\text{needle intersects line}) = \int_0^{\pi/2} \int_0^{L\sin(\theta)/2} f_{X,\Theta}(x,\theta)\, dx d\theta = \int_0^{\pi/2} \int_0^{L\sin(\theta)/2} \frac{4}{d\pi}\, dx d\theta$$
$$= \frac{2L}{\pi d}[-\cos(\theta)]_0^{\pi/2} = \frac{2L}{\pi d}\,.$$

Now we can see how this experiment might have been useful in the past: it provides a way to approximate π by simply performing a real-world experiment to estimate a probability. A convenient setup choice is $L = d/2$ because in this case,

$$\pi = \frac{1}{P(\text{needle intersects line})} \approx \frac{\text{total number of throws}}{\text{number of intersections}}\,.$$

This method is called *Monte-Carlo integration*. Basically, the idea is to interpret the result of an integral as a probability, which makes the integrand a PDF. We can sample from this PDF (e.g. by repeatedly performing a suited experiment) and then approximate the integral by counting how many samples lie in the boundaries that determine the domain of integration. Using this instead of numerical/analytical approaches can be much more efficient computationally, in particular for very complicated, high-dimensional integrals. This is only one example of many existing Monte-Carlo methods, but the idea of how they work is always the same (exploit properties of randomness).

Example 2.15: Broken Stick

We will now derive mathematically how a stick of length L will break (assuming the stick is "uniform", i.e. made of same material everywhere, of same thickness, etc.). Considering breaking as a random process, the remaining stick length can be described by a uniformly distributed RV X that has realizations $x \in [0, L]$. Clearly,

$$E[X] = \frac{L}{2}$$

from the properties of uniform distributions.

However, nothing prevents us from breaking the stick again and this gives rise to another uniformly distributed RV Y with realizations $y \in [0, x]$. The PDF of X is simply

$$f_X(x) = \begin{cases} \frac{1}{L}, & 0 \le x \le L \\ 0, & \text{else} \end{cases},$$

but for Y, we have to take an intermediate step. The conditional PDF is

$$f_{Y|X}(y|x) = \begin{cases} \frac{1}{x}, & 0 \le y \le x \\ 0, & \text{else} \end{cases}.$$

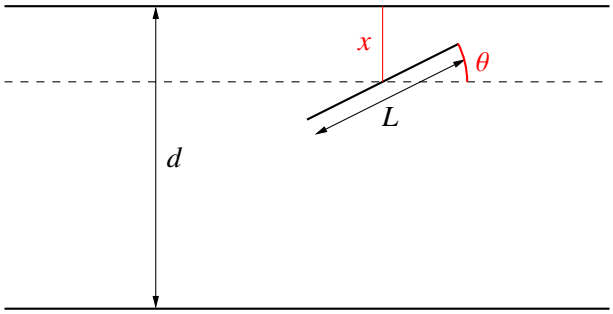


Figure 2.7: Sketch of Buffon's needle experiment

From that, we can also compute the conditional expectation

$$E[Y|X=x] = \int_0^x y f_{Y|X}(y|x)\,dy = \int_0^x y\frac{1}{x}\,dy = \frac{1}{x}\frac{y^2}{2}\Big|_0^x = \frac{x}{2}.$$

This result is totally what we expect, the stick will break at half of its remaining length.

Moreover, knowing $f_{Y|X}(y|x)$ directly gives us the joint PDF

$$f_{X,Y}(x,y) = f_{Y|X}(y|x)f_X(x) = \begin{cases} \frac{1}{L}\frac{1}{x}, & 0 \le x \le L, 0 \le y \le x \\ 0, & \text{else} \end{cases}$$

and thus we can marginalize to finally obtain

$$f_Y(y) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)\,dx = \int_y^L \frac{1}{L}\frac{1}{x}\,dx = \frac{1}{L}\log\left(\frac{L}{y}\right).$$

This enables us to compute the expectation of Y. A rather complicated calculation yields

$$E[Y] = \int_0^L y f_Y(y)\,dy = \frac{L}{4}.$$

This is an intuitive result as well. If sticks are most likely to break at half of their length, the result after breaking twice will be a stick with quarter of the initial length.

2.5 Bayes' Theorem

Now we turn to Bayes' rule again. It has already been mentioned that this is a very useful tool to reverse the order of conditioning, which corresponds to going from a causal model to the probability of this model being true. It is widely used in inference and helps making sense of the world around us – it helps testing models and hypotheses about causal relationships that can not be observed directly. For this reason, probabilities are assigned as our best guess/estimate of the "true" answer.

As an example, we will now deal with the analysis of an apparatus like it is shown in figure

2.8. The idea is that we would like to know which value X takes, but we only have Y, which is potentially corrupted by noise N from the detector (which is why it is a function of X, N, often $Y = X + N$). Therefore, our task/goal is to infer something about X from Y.

Equation (1.6) demonstrated how Bayes' rule applies to events in sample spaces. However, from the joint likelihoods (2.12), (2.43) we see that there is an analogous relationship for discrete and continuous RVs:

$$P_{X|Y}(x|y) = \frac{P_X(x)\,P_{Y|X}(y|x)}{P_Y(y)} \qquad f_{X|Y}(x|y) = \frac{f_X(x)\,f_{Y|X}(y|x)}{f_Y(y)}\,. \tag{2.51}$$

For PMFs, this should be rather intuitive because they are defined via probabilities that are computed in the sample space. From that, we get the continuous version in the usual manner by replacing PMF $\to$ PDF and $\sum \to \int$. Mathematically, this is how one can estimate discrete (continuous) RVs from measurements of discrete (continuous) RVs.

However, this is not the end of the story. It is actually possible to make sense of the case where X is discrete and Y continuous (or vice versa).

Property 2.16: Bayes' Theorem for Mix of Discrete, Continuous RVs

For a continuous RV X and a discrete RV Y

$$f_{X|Y}(x|y) = \frac{f_X(x)\,P_{Y|X}(y|x)}{P_Y(y)} \quad \Leftrightarrow \quad P_{Y|X}(y|x) = \frac{P_Y(y)\,f_{X|Y}(x|y)}{f_X(x)}\,. \tag{2.52}$$

This property allows us to make inference on CRVs using measurements of discrete values. An example from physics would be sending a current into a device that converts it into photons. The reverse is also possible: we can get a continuous stream of values from discrete input, e.g. a current that detects photons (which is the basic idea behind a photodiode).

Conceptually, that is all there is to inference. In practice however, there are always problems to overcome, e.g. characterizing and controlling devices, finding the correct PMFs and PDFs or calculating complicated integrals to get normalization factors.

Example 2.17: Blueprint for Inferences

Bayes' theorem is widely used for inferences and the idea how to apply it will be shown here. Suppose we have some data $\{\tau_i\}$ measuring the ticks of a Geiger-Müller-counter, i.e. we have data on the radioactive decay of an unknown source. To determine this source, we can use the measured waiting times $\{\tau_i\}_{i=1}^N$ to infer (an estimate of) the waiting time/half life τ_s of the source.

Mathematically speaking, we are interested in the posterior $p(\tau|\{\tau_i\})$ and its maximum. The first thing we have to specify is the prior $p(\tau)$. To ensure we do not influence the results inferred from the posterior distribution by putting too much weight (a peak)

onto some value of τ, it is customary to choose a uniform prior

$$p(\tau_s) = \begin{cases} \frac{1}{\tau_{\max} - \tau_{\min}}, & \tau_{\min} \leq \tau_s \leq \tau_{\max} \\ 0, & \text{else} \end{cases}$$

where $\tau_{\min} = \min(\{\tau_i\}), \tau_{\max} = \max(\{\tau_i\})$ (for a sufficient amount of samples, the "true" waiting time should neither be smaller than the smallest measured one nor bigger than the biggest one). The other ingredient needed is the likelihood and we can specify that by using our knowledge that decays follow an exponential distribution (2.31). By making the rate $\lambda = \frac{1}{\tau_s}$ a parameter of the distribution, we can assign a likelihood

$$p(\tau_i|\tau_s) = e^{-\tau_i/\tau_s} \Big/ \tau_s$$

of each value in $\{\tau_i\}$ being measured, assuming τ_s is the parameter of the underlying process. Since each decay is independent of all others, the total likelihood is

$$p(\{\tau_i\}|\tau_s) = \prod_{i=1}^{N} p(\tau_i|\tau_s) = e^{-\frac{1}{\tau_s}\sum_{i=1}^{n}\tau_i} \Big/ \tau_s^N \ .$$

After all, the likelihood is still a probability density, so the same rules regarding independence etc. apply (although it has to be emphasized that it is *not* a PDF with respect to τ_s as it does not necessarily fulfil the closure rule; it is, in principle, a PDF with respect to τ, but we evaluate it in some fixed value τ_i).

This is all we need because these quantities allow the computation of the normalization constant (sometimes called *evidence*) by expressing it as a marginal probability:

$$p(\tau) = \int p(\tau, \tau_s)\, d\tau_s = \int p(\tau_s) p(\tau|\tau_s)\, d\tau_s \ .$$

Now, we are ready to compute posterior values according to

$$p(\tau_s|\{\tau_i\}) = \frac{p(\tau_s)p(\{\tau_i\}|\tau_s)}{\int p(\tau_s)p(\{\tau_i\}|\tau_s)\, d\tau_s} \ .$$

Results for some example data are visualized in figure 2.9. We can see that the posterior changes as more data is collected. However, the changes mainly affect the width of the distribution instead of peak location as more data is taken into account, which means the results remains the same even for increased accuracy. The final estimate is

$$\tau_{s,\max} = \max_{\tau_s} p(\tau_s|\{\tau_i\}) = 4.01 \ .$$

We could also use other quantities for estimation (e.g. mean or median), but the maximum is the most straightforward one.

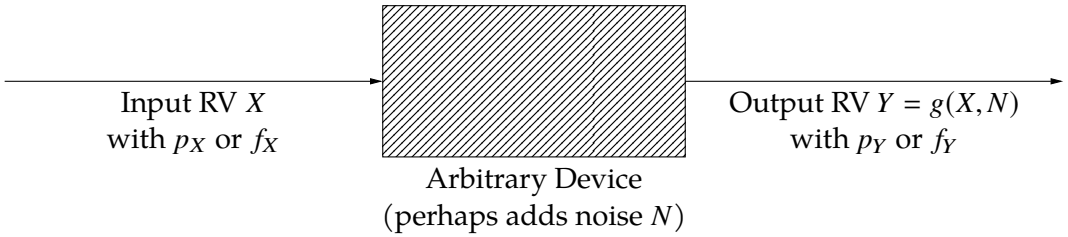


Figure 2.8: Schematic sketch of device that transforms input RV

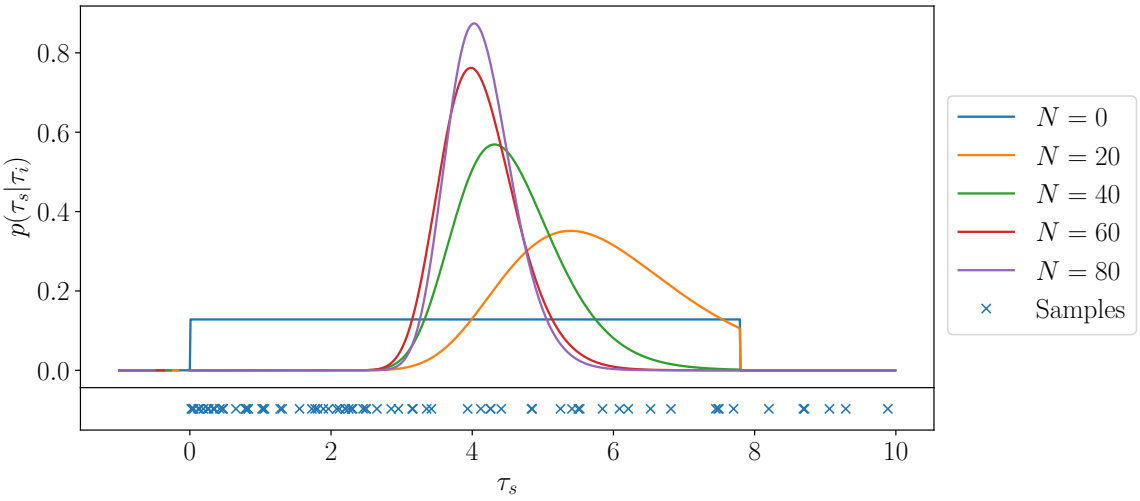


Figure 2.9: Evolution of posterior as the number of data points N is increased. For $N = 0$, the posterior is nothing but the prior.

2.6 Distributions of Functions of RVs

From the law of the unconscious statistician we know that $E[g(X)] \neq g(E[X])$. The same is true for probabilities and densities. However, just like the law of the unconscious statistician still told us how to calculate $E[g(X)]$, there are general methods to calculate the distributions of functions of RVs and we will now introduce them.

It should be intuitively clear that a function g will change the distributions of RVs when going from X to $Y = g(X)$, no matter if they are discrete or continuous. A very easy example to see this in the discrete case is $g(X) = X^2$ with $P_Y(y) = P_X(x_1) + P_X(x_2)$ where $x_{1,2}^2 = y \Leftrightarrow x_{1,2} = \pm\sqrt{y}$. However, working with probabilities is not the general way to go because for CRVs, probabilities of points are zero. As already mentioned when introducing CRVs, the CDF is a quantity that is common between discrete and continuous RVs, so it makes more sense to use this over the PMF/PDF. This approach is equivalent because distributions and densities can be computed from the CDF, e.g. by differentiation.

Before giving a formal description of this cookbook recipe, we will go through a simple example that will help with understanding the general idea.

Example 2.18: Non-linear vs. Linear Function

Consider the uniformly distributed CRV $X \sim \mathcal{U}(0,2)$ (i.e. $f_X(x) = 1/2$). Defining $Y = X^3$, we already know that $y \in [x_{\min}^3, x_{\max}^3] = [0,8]$ because polynomials are monotonic. However, it is a non-linear function and hence, the distribution of Y will be non-uniform. To infer an explicit formula, we write:

$$F_Y(y) = P(Y \leq y) = P(X^3 \leq y) = P(X \leq y^{1/3}) = \int_0^{y^{1/3}} f_X(x)\,dx = \int_0^{y^{1/3}} \frac{1}{2}\,dx = \frac{y^{1/3}}{2}\,.$$

Consequently,

$$f_Y(y) = \frac{dF_Y(y)}{dy} = \frac{y^{-2/3}}{6} = \frac{1}{6y^{2/3}}\,.$$

This density is much higher for smaller y than it is for larger ones. The reason is the relationship $y = x^3$, which has an increasingly steep slope. Therefore, many of the smaller values x get mapped to similar y-values, while higher x are mapped to increasingly distant y, so the density of y-values decreases as x increases.

If the relationship between X, Y was linear, the distribution would not change because

$$F_Y(y) = P(Y \leq y) = P(aX + b \leq y) = P\left(X \leq \frac{y-b}{a}\right) = F_X\left(\frac{y-b}{a}\right).$$

Applying the chain rule once yields

$$f_Y(y) = \frac{dF_Y(y)}{dy} = \frac{dF_X\left(\frac{y-b}{a}\right)}{dy} = \frac{d\frac{y-b}{a}}{dy} \cdot f_X\left(\frac{y-b}{a}\right) = \frac{1}{a} \cdot f_X\left(\frac{y-b}{a}\right).$$

While there is a scaling factor that appears due to the closure rule, the important point is that the distribution shape does not change, $f_Y \sim \mathcal{U}$, unlike for $Y = X^3$.

Looking at the result though, we notice that $a < 0$ would produce $f_Y(y) < 0$, which violates the requirements for a PDF. In this case,[14]

$$F_Y(y) = P(aX + b \leq y) = P\left(x \geq \frac{y-b}{-|a|}\right) = 1 - F_X\left(\frac{y-b}{-|a|}\right)$$

and we obtain the analogous (in fact, more general) result

$$f_Y(y) = \frac{f_X\left(\frac{y-b}{a}\right)}{|a|}\,.$$

This is a mathematical detail and does not change the interpretation of constant shape.

[14] Remembering that dividing by a negative number changes the "direction" of an inequality.

Admittedly, this example was very simple. For other functions, calculations can be much more complicated because it may not be so easy to invert them (at least not globally).

Example 2.19

Dealing with functions of multiple RVs is not so different from dealing with functions of one RV and we will now look at $Z = g(X,Y) = Y/X$. Assuming $f_{X,Y}(x,y) = 1,\ x,y \in [0,1] \Rightarrow z \in [0,\infty]$ and that X,Y are independent, we will now derive $f_Z(z)$. The procedure is the same as before:

$$\begin{aligned} F_Z(z) &= P(Z \le z) = P(Y/X \le z) = P(Y \le xz) = F_Y(xz) \\ &= \int_0^{xz} f_Y(y)\,dy = \int_0^{xz}\int_0^1 f_{X,Y}(x,y)\,dxdy \\ &= \int_0^1 y\Big|_0^{xz} dx = \frac{zx^2}{2}\Big|_0^1 = \frac{z}{2}\,. \end{aligned}$$

However, when looking closer at the steps, this calculation assumes $z \le 1$ (boundaries for y have to be ≤ 1 due to domain; same for resulting probability). For $z > 1$,

$$\begin{aligned} F_Z(z) &= P(Y/X \le z) = P(X/Y \ge 1/z) = P(X \ge y/z) = 1 - F_X(y/z) \\ &= 1 - \int_0^{y/z} f_X(x)\,dx = 1 - \int_0^{y/z}\int_0^1 f_{X,Y}\,dydx \\ &= 1 - \int_0^1 [x]_0^{y/z}\,dy = 1 - \int_0^1 \frac{y}{z}\,dy = 1 - \frac{y^2}{2z}\Big|_0^1 = 1 - \frac{1}{2z}\,. \end{aligned}$$

Now we can compute

$$f_Z(z) = \frac{dF_Z(z)}{dz} = \frac{d}{dz}\begin{cases}\frac{z}{2}, & z \le 1\\ 1-\frac{1}{2z}, & z > 1\end{cases} = \begin{cases}\frac{1}{2}, & z \le 1\\ \frac{1}{2z^2}, & z > 1\end{cases}.$$

That allows calculations of other properties of the CRV Z, such as

$$E[Z] = \int_0^\infty z \cdot f_Z(z)\,dz = \int_0^1 \frac{z}{2}\,dz + \int_1^\infty \frac{1}{2z}\,dz = \frac{z^2}{4}\Big|_0^1 + \frac{\ln(z)}{2}\Big|_1^\infty = \infty\,.$$

This seems odd and thus we might wish for a verification of this result. We can indeed get this by recalling that due to the independence of X,Y,

$$E[Y/X] = E[Y]E[1/X] = \infty$$

since $x = 0$ occurs in the second expectation.

Apparently, although the basic idea is "just" inverting and deriving, determining distributions of functions of RVs can be tricky. Luckily, there is a broad class of problems where this process is much easier because a general formula exists.

Property 2.20: Distributions of Monotonic Functions

For a monotonic function g and $Y = g(X)$,

$$f_Y(y) = \left.\frac{f_X(x)}{\left|\frac{dg(x)}{dx}\right|}\right|_{x=g^{-1}(y)} = f_X(g^{-1}(y))\left|\frac{dg^{-1}(y)}{dy}\right|. \tag{2.53}$$

This is a direct corollary of transformation rules from measure theory and also how one can prove the law of the unconscious statistician. Just like before, the statement should make sense intuitively because for a function with small (high) slope in a certain point, many (few) neighbouring points get mapped close to where this point is mapped and the result is a higher (lower) density. This inverse relationship is expressed in the first equality of (2.53). The same arguing can be done for non-monotonic functions, but the formula does not hold due to difficulties with inverting them.

In the end, a very compact workflow often looks like this: arguing that the function is monotonic (not globally, only on domain that is treated) and then using property 2.20.

Example 2.21: Non-linear vs. Linear Function 2

We can now derive a more general version of previous example 2.18 (and proof the second part of it very quickly). In this example, we looked at $Y = g(X) = X^3 \Leftrightarrow X = Y^{1/3}$ and derived its PDF for $X \sim \mathcal{U}(0,2)$. Now, we do not have to assume an explicit distribution and can instead use (2.53), which yields

$$f_Y(y) = f_X(x)\left/\left|\frac{dx^3}{dx}\right|\right|_{x=y^{1/3}} = \left.\frac{f_X(x)}{3x^2}\right|_{x=y^{1/3}} = \frac{f_X(y^{1/3})}{3y^{2/3}}.$$

For $X \sim \mathcal{U}(0,2) \Rightarrow f_X(x) = 1/2$, we retrieve the result from example 2.18.

Based on these general approaches, one can derive more specialised shortcuts to compute distributions. Suppose we are interested in $Z = X + Y$, assuming discrete, independent RVs for now. The PMF of Z can be obtained from the joint PMF of X, Y according to

$$P_Z(z) = \sum_{x,y:\ x+y=z} P_{X,Y}(x,y) = \sum_{x,y:\ x+y=z} P_X(x)P_Y(y) = \sum_x P_X(x)P_Y(z-x). \tag{2.54}$$

This is the *convolution formula*. For independent[15] CRVs, it takes the form

$$f_Z(z) = \int_{-\infty}^{\infty} f_X(x)f_Y(z-x)\,dx. \tag{2.55}$$

The proof involves CDFs and using property 2.20.

[15] This condition is necessary because we wish to use $f_{Y|X}(y|x) = f_Y(y) = f_Y(z-x)$.

Example 2.22: Convolution of Gaussians

Consider two independent normal RVs X, Y, such that

$$f_{X,Y}(x,y) = \frac{1}{2\pi\sigma_X\sigma_Y} e^{-\frac{(x-\mu_X)^2}{2\sigma_X}} e^{\frac{-(y-\mu_Y)^2}{2\sigma_Y}} .$$

Using the convolution formula, a rather lengthy calculation yields

$$f_Z(z) = \frac{1}{\sqrt{2\pi\left(\sigma_X^2+\sigma_Y^2\right)}} e^{-\frac{1}{2}\frac{(z-(\mu_X+\mu_Y))^2}{\sigma_X^2+\sigma_Y^2}} \tag{2.56}$$

where $Z = X + Y$. This is also a normal RV, $Z \sim \mathcal{G}(\mu_X + \mu_Y, \sigma_X^2 + \sigma_Y^2)$, as expected from the law of the unconscious statistician (for the case of independent CRVs, otherwise variances could not be added).

Put differently, the convolution of two exponentials is an exponential again.

2.7 Iterated Expectations and Conditional Variances

The broken stick example already showed how $E[X|Y = y]$ can be a function of y. By not fixing an explicit value y, we can interpret the result as a RV again and in particular, using the law of the unconscious statistician, take another expected value.

Property 2.23: Law of Iterated Expectations

For two RVs X, Y

$$E_Y[E_X[X|Y]] = E_X[X] . \tag{2.57}$$

The average conditional expectation matches the "regular" expectation. In some cases, using iterated expectations is a very convenient way to calculate expected values, but the reverse statement may also be interesting.

Example 2.24: Broken Stick 2

One example where conditional expectations have been used was the broken stick 2.15. We saw that a stick is most likely to break in its middle, i.e. $E_X[X] = L/2$ where X is the RV for remaining length. Breaking it again can be modeled using another RV Y and clearly, knowing the stick originally broke at some point x, $E_Y[Y|X = x] = x/2$ (also result of a short calculation). But this is not the expectation of Y and to get it, some long calculations were needed. With our new tool however, we can simply use

$$E_Y[Y] = E_X[E_Y[Y|X]] = E_X[X/2] = E_X[X]/2 = L/4,$$

which is the same result.

Going one step further, we can also define the notion of a *conditional variance*

$$\operatorname{var}(X|Y=y) = E_X[(X - E_X[X|Y=y])^2|Y=y]\,. \tag{2.58}$$

In a very similar manner to what was done before, this can be interpreted as a function of y and thus a RV $\operatorname{var}(X|Y) = E_X[(X - E_X[X|Y])^2|Y]$.

Property 2.25: Law of Total Variance

For two RVs X, Y

$$\operatorname{var}(X) = E_Y[\operatorname{var}(X|Y)] + \operatorname{var}\left(E_Y[X|Y]\right). \tag{2.59}$$

Basically, that means the variance of X is the average conditional variance plus some extra term that represents the variance of the average conditioned RV X.

Example 2.26: Quiz among Students

Consider a quiz among students. which are divided into two sections. The quiz score is a RV X, while the section is a RV Y. All information that we are given is that 10 students are assigned to section $y = 1$, while 20 are assigned to section $y = 2$, and that the average score in section 1 is 90, while it is 60 in section 2. In more formal notation, that means:

$$P(Y=1) = \frac{1}{10}, \quad P(Y=2) = \frac{1}{20}, \quad E_X[X|Y=1] = 90, \quad E_X[X|Y=2] = 60\,,$$

which defines the PMF of $Z = Z(Y) = E_X[X|Y]$.

Now, what is the expected score if one student is picked at random? This is equivalent to asking: what is the expectation of Z. To find that out, we compute

$$E_Y[E_X[X|Y]] = E[X|Y=1]\cdot P(Y=1) + E[X|Y=2]\cdot P(Y=2) = \frac{1}{3}\cdot 90 + \frac{2}{3}\cdot 60 = 70\,.$$

The law of iterated expectations implies that this should be equal to $E_X[X]$. We can verify this by rearranging the corresponding sum:

$$E_X[X] = \frac{1}{30}\cdot\sum_i x_i = \frac{1}{3}\cdot\frac{1}{10}\cdot\sum_{i:\,y=1} x_i + \frac{2}{3}\cdot\frac{1}{20}\sum_{j:\,y=2} x_j = \frac{1}{3}\cdot 90 + \frac{2}{3}\cdot 60 = 70\,.$$

The natural additional information besides average/expectation is the uncertainty and hence deviation from the average, which is given by the variance:

$$\begin{aligned}\operatorname{var}(E_X[X|Y]) &= E_Y\left[(E_X[X|Y] - E_Y[E_X[X|Y]])^2\right] = E_Y\left[(E_X[X|Y] - E_X[X])^2\right] \\ &= P(Y=1)\cdot(E[X|Y=1] - E[X])^2 + P(Y=2)\cdot(E[X|Y=2] - E[X])^2 \\ &= \frac{1}{3}\cdot(90-70)^2 + \frac{2}{3}\cdot(60-70)^2 = \frac{400}{3} + \frac{200}{3} = 200\,.\end{aligned}$$

It might be surprising that this result is so much bigger than the average score of 70, but

this simply comes from the quadratic nature of the variance. The corresponding standard deviation would be $\sqrt{\text{var}(E[X|Y])} = 14.14$, a much more reasonable/comparable result. The corresponding 1σ-interval around $E_X[X|Y]$ is given by

$$[70 - 14.14, 70 + 14.14] = [55.86, 84.14]\,.$$

We can see that $E_X[X|Y = 2] = 60$ lies in this interval, while $E_X[X|Y = 1] = 90$ does not, which reflects the fact that more students are in section 2.

2.8 Limit Theorems

In reality, it is seldomly possible to actually measure distributions. Instead, one can measure outcomes/realizations. Those measurements are commonly called *samples* and from their density in certain intervals, one can estimate the underlying PDF (this is the idea of a *histogram*). However, not knowing the distribution imposes a problem: how to get quantities like expectations and variances from a sample? It turns out that there are sample-analogues of most quantities, for example the *sample mean*

$$M_n := \frac{1}{n}\sum_{i=1}^{n} x_i \tag{2.60}$$

for expectations or

$$\frac{1}{n-1}\sum_{i=1}^{n}(x_i - M_n)^2 \tag{2.61}$$

for variances (where n is the sample size).

But that brings up another question: how can we make sure that these quantities make sense and are "correct"? A very reasonable approach is that they should converge to the actual quantities of the distribution in some limit. The appropriate one is $n \to \infty$, where sampling errors become negligible. Hence, probability theory in the way it is often applied nowadays and in fact, the whole interpretation of probability ≡ frequency of occurrence, depends fundamentally on certain limit theorems holding.

In order to understand them, we need some tools at first.

Property 2.27: Markov Inequality

For a RV $X > 0$ and $c > 0$

$$E[X] \geq c\,P(X \geq c) = c\,(1 - F_X(c))\,. \tag{2.62}$$

A simple corollary for the RV $(X - \mu)^2$ is

$$\text{var}(X) \geq c^2\,P\Big((X-\mu)^2 \geq c^2\Big) = c^2\,P(|X - \mu| \geq c), \quad c > 0\,, \tag{2.63}$$

where we abbreviate $\mu = E[X]$ and replaced $c \to c^2$ for convenience. The statement becomes more obvious if we set $c = k\sigma_X$ (in this form, it is called the *Chebyshev Inequality*[16]):

$$P(|X - \mu| \geq k\sigma_X) \leq \frac{1}{k^2}\,. \tag{2.64}$$

Therefore, the Chebyshev inequality tells us that the probability of realizations of X landing k standard deviations away from the mean goes as $1/k^2$. This is a very useful way to characterize outliers, even if the distribution itself is unknown (only mean and standard deviation are needed in order to make such statements).

These are the tools that we need for now. Since our initial goal was to formulate limit theorems, we need a notion of convergence first.

Definition 2.28: Convergence in Probability

A sequence $\{X_n\}_{n\in\mathbb{N}}$ of RVs *converges in probability* to a CRV X if

$$\forall \epsilon > 0\ \exists n_0 :\ P(|X_n - X| \geq \epsilon) \leq \delta, \quad \forall n \geq n_0,\, \delta > 0\,. \tag{2.65}$$

This definition looks very complicated, but the basic idea is that X_n is said to converge to X if the probability of their difference being greater than zero vanishes as n increases, i.e.

$$\lim_{n\to\infty} P(|X_n - X| \geq \epsilon) \to 0\,. \tag{2.66}$$

Note that this is another definition using the CDF, not PMF/PDF.

Example 2.29: Simple Sequence

To gain more understanding, we will now treat a sequence of discrete RVs X_n with two realizations $x_n \in \{0, n\}$ and PMF

$$P_{X_n} = \begin{cases} 1 - \frac{1}{n}, & x_n = 0 \\ \frac{1}{n}, & x_n = n \end{cases}\,.$$

Therefore,

$$\lim_{n\to\infty} P_{X_n}(|X_n| \geq \epsilon) = \lim_{n\to\infty} \frac{1}{n} = 0\,,$$

the sequence converges to 0. We can further compute:

$$E[X_n] = 0 \cdot \left(1 - \frac{1}{n}\right) + n \cdot \frac{1}{n} = 1$$
$$\mathrm{var}(X_n) = (0-1)^2 \cdot \left(1 - \frac{1}{n}\right) + (n-1)^2 \cdot \frac{1}{n} = 1 - \frac{1}{n} + n - 2 + \frac{1}{n} = n - 1\,.$$

[16] Note that $(X - \mu)^2$ is a positive RV for any X, not only $X > 0$. Thus, (2.63) and (2.64) hold for arbitrary X.

These results are surprising, after all we have shown that X_n converges to zero. In contrast, its expectation converges to 1 and the variance diverges. This can be understood from the fact that $X_n = n$ with probability $1/n$. While the probability of this value being realized decreases, its separation to $X_n = 0$ increases and in fact, goes to ∞. Hence, it still has a contribution in expectations and from l'Hospital's rule we see that the limit of this contribution to $E[X_n]$ is

$$\lim_{n\to\infty} \frac{n}{n} = \lim_{n\to\infty} 1 = 1\,.$$

Therefore, this unintuitive behaviour of expectation and variance is caused by the unusual nature of the RV itself, one of its values/realizations diverges.

The notion of convergence in probability allows us to examine the behaviour of linear combinations of RVs, for example of the sample mean

$$M_n = \frac{1}{n}\sum_{i=1}^{n} x_i$$

as introduced in equation (2.60). The samples x_i are realizations of RVs X_i and since they can have different realizations, we can associate with M_n a RV

$$M_n = \frac{1}{n}\sum_{i=1}^{n} X_i\,. \tag{2.67}$$

A realization of M_n is nothing but the mean of a sample, i.e. the mean of a set of realizations $\{x_i\}$. We will now assume that the RVs X_i are independent and identically distributed (*i.i.d.*) with finite mean, variance μ, σ^2. For example, they could be a *random process*, which is a sequence of RVs X_t at different times t (noise can be described as a random process). This assumption of i.i.d. RVs implies

$$E[M_n] = \frac{1}{n}\sum_{i=1}^{n} E[X_i] = \frac{1}{n}\cdot n\cdot \mu = \mu \tag{2.68}$$

$$\mathrm{var}(M_n) = \frac{1}{n^2}\sum_{i=1}^{n} \mathrm{var}(X_i) = \frac{1}{n^2}\cdot n\cdot \sigma^2 = \frac{\sigma^2}{n}\,. \tag{2.69}$$

Therefore, the distribution of M_n has the mean of the RVs it consists of (which is why it is also called an *unbiased estimator*) and a deviation/uncertainty that decreases as the sample size n is increased.[17] Hence, it might be interesting to look at what the Chebyshev inequality tells us about $|M_n - \mu|$:

$$\begin{aligned} &\mathrm{var}(M_n) \geq \epsilon^2\, P(|M_n - \mu| \geq \epsilon),\ \forall \epsilon > 0 \\ \Rightarrow\quad &\lim_{n\to\infty} P(|M_n - \mu| \geq \epsilon) \leq \lim_{n\to\infty} \frac{\mathrm{var}(M_n)}{\epsilon^2} = \lim_{n\to\infty} \frac{\sigma^2}{n\epsilon^2} = 0\,. \end{aligned} \tag{2.70}$$

[17] The latter is only true due to the assumption of i.i.d. RVs. Otherwise, covariances would potentially contribute.

This is a statement of tremendous importance, so we will state it again in a more formal way:

Property 2.30: (Strong) Law of Large Numbers

The sample mean M_n of i.i.d. RVs X_i converges in probability to $E[X_i]$.

This theorem is what justifies the sampling approach to probability theory. It ensures that large sample sizes will lead to estimates M_n approximating the mean of the underlying distribution well. This approach can be very useful to compute quantities from complicated distributions (for which no analytical solutions might exist) or if the distribution is unknown, but it is possible to sample from it.

Example 2.31: Poll Design for Coca Cola

Say Coke is our employer and our boss wants to know which proportion of the population prefers Coke over Pepsi. To find that out, we have to ask people, i.e. make a poll. We can describe this poll as a RV X with realizations

$$x = \begin{cases} 1 \text{ (yes)}, & \text{probability } f \\ 0 \text{ (no)}, & \text{probability } 1-f \end{cases}.$$

Apparently, $f = E[X]$ is what answers our question and hence what we want to know. Since it is not feasible to ask every person in the population, we can only estimate it as

$$f_n = \frac{1}{n}\sum_{i=1}^{n} x_i \,.$$

It is very reasonable to demand that n shall be big enough that f_n is accurate to a certain degree, for example 1%. Of course, due to the stochastic nature of taking a finite sample, it is impossible to guarantee this accuracy. Instead, we can only make statements at a certain level of confidence by using the Markov inequality. Setting the desired confidence level to 95%, we want the probability of $|f_n - f|$ exceeding 1% to be smaller than $5\% = 100\% - 95\%$. This is surely the case if (note again that $E[X] = f$)

$$P(|f - f_n| \geq 0.01) \leq \frac{\sigma_{f_n}^2}{0.01^2} = \frac{\sigma_X^2}{n \cdot 0.01^2} \leq 1 - 0.95 = 0.05 \,.$$

X is a binomial RV (but we only make one "trial" of the corresponding experiment, which is the poll) and thus $\sigma_X^2 = f(1-f) \leq 1/4$. This shall still be ≤ 0.05, so

$$\frac{1}{4 \cdot n \cdot 0.01^2} \leq 0.05 \quad \Leftrightarrow \quad 4 \cdot n \cdot 0.01^2 \geq \frac{1}{0.05} \quad \Rightarrow \quad n \geq 50000 \,.$$

50000 people have to be asked for 95% confidence that f_n estimates f to 1% accuracy.

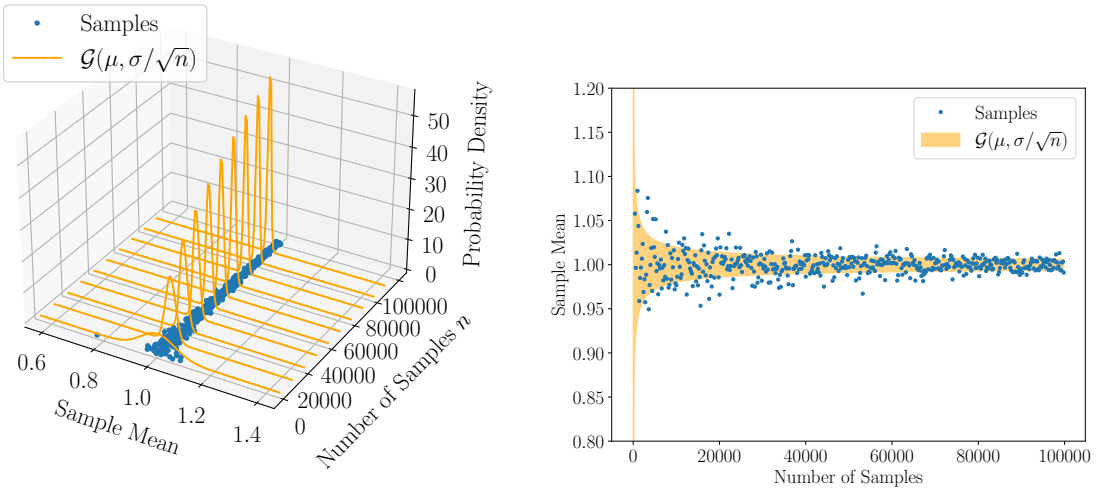
There is another statement that tells us more about samples – this time, however, it is about the whole distribution of their sample mean and not just the average sample mean:

Property 2.32: Central Limit Theorem

For i.i.d. RVs X_i with finite mean, variance μ, σ^2 and their mean $M_n = \frac{1}{n}\sum_i X_i$, the distribution of the RV $\frac{\sqrt{n}(M_n-\mu)}{\sigma}$ converges to $\mathcal{G}(0,1)$.[18] This is equivalent to the distribution of M_n converging to $\mathcal{G}(\mu, \sigma/\sqrt{n})$.

As we had already seen, the mean of M_n is equal to μ and its variance decreases as $1/n$. However, we now know that the corresponding distribution of M_n is Gaussian, regardless of the underlying distribution of the X_i. Similar to the law of large numbers, this is a very interesting statement in case we only have one realization of M_n computed from a fixed set of samples $\{x_i\}$ available, since it tells us something about how likely it is to have a certain deviation from the "true" result μ (depending on the number of samples n that are used). An illustration of this where the X_i are Gaussian RVs themselves is provided in figure 2.10.

[18] This is convergence in distribution, which was not defined here. It describes the behaviour of the CDF. More details are not important here because it is just a mathematical way to state the intuitive understanding.



(**a**) Samples and underlying distribution

(**b**) Samples and projection of Gaussian into xy-plane

Figure 2.10: Sample mean for Gaussian distribution $\mathcal{G}(1, 2)$. The shaded region in (b) represents one standard deviation $\sigma/\sqrt{n} = 2/\sqrt{n}$ around the mean $\mu = 1$.