A crash course in probability

Periklis A. Papakonstantinou

Rutgers Business School

LECTURE NOTES IN Elements of Probability and Statistics

Periklis A. Papakonstantinou
MSIS, Rutgers Business School

Piscataway, January 2017

Acknowledgments

I am grateful to Tanay Talukdar for reconstructing much of the calculations and arguments and for discovering subtle mistakes. I would also like to thank Jia Xu for detailed expository remarks and technical suggestions. Finally, let me thank Hafiz Asif and Andrea Nemes, my teaching assistants and students, for their remarks.

Lecture 1

An interlude of basic probability theory

These notes are not a replacement for any proper textbook on the subject. You are encouraged to review (or cover) material from proper sources, such as the textbooks suggested at the end.

1.1 What is a probability space?

A *probability space* (or *sample space*) is two things:

i. a set Ω, together with

ii. a function $f : \Omega \to [0, 1]$

For simplicity, say that Ω is finite, e.g. it has 10 elements. We only require f to have the property: $f(\text{1st element in }\Omega) + f(\text{2nd element in }\Omega) + \cdots + f(\text{10th element in }\Omega) = 1$. Sometimes, we say "the space Ω" and by this we always mean the pair (Ω, f). We allow ourselves to be sloppy when f is well-understood from the context. Furthermore, in most cases we write Pr instead of f. Using the same symbol "Pr" for measuring probability for all probability spaces may cause confusion. For example, when in a calculation two distinct probability spaces are involved – i.e. the same symbol Pr is used for each of the different spaces. In this case we try to infer things from context. The main reason we use the same symbol Pr to refer to different measure functions is tradition.

For now, we will focus on finite Ω's.

An *event* is just a subset of Ω, e.g. $\mathcal{E} \subseteq \Omega$. We define $\Pr[\mathcal{E}] = \Pr[e_1] + \Pr[e_2] + \cdots + \Pr[e_k]$, where $\mathcal{E} = \{e_1, e_2, \ldots, e_k\}$. Each e_i is called an *elementary event* or *elementary outcome* and corresponds to the event $\{e_i\}$.

Probability theory aims to precisely model our real-world intuition in formal (i.e. unambiguous) terms.

Example 1. *Consider the following statement we wish to evaluate:*

"The probability that the outcome of a roll of a fair die is even"

*Our real-world intuition is that this probability is 50%, which as a fraction is $\frac{1}{2}$. What if we try to write this less informally as $\Pr[\textit{fair die outcome is } 2 \textit{ or } 4 \textit{ or } 6]$? Is this a correct probability expression? No, unless there is a rigorously defined probability space it is wrong (and meaningless) to write $\Pr[\ldots]$ (probability of what? over what? what is the exact thing we wish to measure?). The notation $\Pr$ performs a measurement only **over** a probability space.*

In real life we may say "formal statistical model" instead of "probability space" (same thing). Let us now define the formal model.[1]

Fair die: this means that the space consists of all faces of the die outcomes $\Omega = \{\textit{face } 1, \textit{face } 2, \ldots, \textit{face } 6\}$ and all faces[2] *are equiprobable, i.e. $\Pr[\textit{face } 1] = \frac{1}{6}, \ldots, \Pr[\textit{face } 6] = \frac{1}{6}$. This is our model of the world. The event that the outcome is an even face is $\mathcal{E} = \{\textit{face } 2, \textit{ face } 4, \textit{ face } 6\}$. Then, $\Pr[\mathcal{E}] = \frac{1}{6} + \frac{1}{6} + \frac{1}{6} = \frac{1}{2}$.*

The original intuitive "$\Pr[\text{fair die outcome is } 2 \text{ or } 4 \text{ or } 6] = \frac{1}{2}$" coincides with the detailed formal treatment. It is immediate how to go from informal to formal. When the details are completely understood from context we will trade formality for readability.

[1] The gain in having a formal model is that we can forget about the real-world (real-world is complex). Now, all calculations and inferences can be done unambiguously (any disagreement can *only* be raised *before* the mathematical modeling).

[2] Usually, in a die "face 1" is a dot $\boxed{\bullet}$, "face 2" is $\boxed{\bullet\,\bullet}$, and so on.

Remark: It is very important to remember that *a probability space describes exactly one realization of an experiment*. Given the space $\Omega = \{\text{face } 1, \text{face } 2, \ldots, \text{face } 6\}$ defined as above can we measure in this Ω the probability that when the die is rolled twice and the first time the outcome is face 1 and the second time the outcome is face 2? No, in *this* space Ω the probabilistic question does not even make sense. The elements of the space are outcomes of a single die roll. For example the event $\{\text{face } 1, \text{face } 2\}$ corresponds to the probabilistic event that in a single (same) die roll the outcome is face 1 *or* face 2. If we want to formally measure two rolls of a die then we should have used a more complicated Ω. That is, a *different model* of the world; for example, a *joint model*, i.e. modeling jointly two successive die rolls. In this case every elementary event consists of two outcomes of a die roll. Instead of $\{\text{face } 1, \ldots, \text{face } 6\}$ the new space consists of pairs $\big\{(\text{face } 1, \text{face } 1),\ (\text{face } 1, \text{face } 2),\ \ldots, (\text{face } 6, \text{face } 5),\ (\text{face } 6, \text{face } 6)\big\}$.

Question Given one probability space can we construct other, more interesting spaces?

1.2 Product spaces

Let $(\Omega, \Pr_\Omega)$ be a probability space.[3] Let us now define the product space. This is just a definition (i.e. "definitions" can even be arbitrary notions – no room for disagreement). We define the *product space* Ω^2 as: (i) $\Omega^2 = \Omega \times \Omega$ and (ii) $\Pr_{\Omega^2}[(x, y)] = \Pr_\Omega[x] \Pr_\Omega[y]$, for every $x, y \in \Omega$.

Remark on terminology 2. *Recall that Ω^2 is just one set. That is, Ω^2 is one symbol (similar to Ω) that denotes a single set.*

[3] Note that we change notation a little bit and write $\Pr_\Omega$, instead of the plain Pr, just to put emphasis on the fact that $\Pr_\Omega$ is associated with this specific Ω.

Remark on terminology 3. *We decided to subscript* Pr *with each of the corresponding probability spaces to avoid confusion (one space is Ω^2 whereas the other two, each is a copy of Ω).*

Example 4. *Let $\Omega = \{H, T\}$ be the space of the outcomes when flipping once a fair (unbiased) coin. Then, $\Omega^2 = \{(H, H), (H, T), (T, H), (T, T)\}$ is the set where each elementary outcome has probability $\frac{1}{2} \cdot \frac{1}{2} = \frac{1}{4}$.*

Therefore, the product of *uniform* probability spaces is itself a uniform space. Recall, that "uniform" is a probability space where each elementary event has the same probability.

So far a "product space" appears to be an arbitrary mathematical definition. Arbitrariness is due to the multiplication of probabilities of the original spaces. Why "multiply" the probabilities of $\Pr[H]$ and $\Pr[T]$ when defining the probability of $\Pr[(H, T)]$ and not do something else?[4] There is a very natural connection of product spaces with the notion of "chance" and "probability" in real-life.

What is a product space in practice? It corresponds to an idealized experiment where one flips an unbiased coin once, records its outcome, and "independently" flips another unbiased coin (or the same – doesn't matter) and records its outcome. For example, if the first outcome is "heads" and the second is "tails" this corresponds to the element (H, T) is the above product space. But there is something much deeper about (H, T), which has to do with the fact that the "coin flips are independent". We will see that the theory captures amazingly well our real-world perception. A product space is a *special case* of what we call *statistical independence* (there are many ways in which statistical independence arises and "product spaces" is one such way).

[4] For example, why not adding the probabilities, or why not to multiply $\Pr[H]$ by 2 and $\Pr[T]$ by 42 and then add them up? One problem is that the new set must be a probability space; e.g. after we define the probabilities of the elementary events it should be the case $\Pr[(H, H)] + \Pr[(H, T)] + \Pr[(T, H)] + \Pr[(T, T)] = 1$. But this is not a serious problem at all. We can always add everything up and then normalize each individual elementary event. There is a far more important reason why we decided to define $\Pr[(H, T)]$ as $\Pr[H]\Pr[T]$.

We are not restricted to define the product space over the same Ω. For two probability spaces Ω_1 and Ω_2, define $\Omega' = \Omega_1 \times \Omega_2$ and $\Pr_{\Omega'}[(x, y)] = \Pr_{\Omega_1}[x] \Pr_{\Omega_2}[y]$.

For instance, Ω_1 may correspond to rolling a die and Ω_2 to flipping a coin. Then, Ω' is the *joint model* of the experiment of rolling a die and independently flipping a coin.[5]

1.3 From intuition to definition

Humans have some intuitive idea about what is "independence". It means that the (statistical) realization of one event does not "affect" the (statistical) realization of the other. For example, if I flip "independently" the same unbiased coin twice I expect the outcome of both the first and the second time to be 50-50 heads and tails.

The quantitative problem we have to solve now is to give a *formal* definition of independence. Whichever definition we give, this should formalize precisely (i.e. with numbers regarding probabilistic calculations), the above intuitive idea we have about independence.

Statistical independence

Let $(\Omega, \Pr)$ be a probability space. We say that $\mathcal{E}, \mathcal{E}' \subseteq \Omega$ are *independent* (or *statistically independent*) if $\Pr[\mathcal{E} \cap \mathcal{E}'] = \Pr[\mathcal{E}] \Pr[\mathcal{E}']$.

It is not immediately obvious whether this formalizes the idea that the realization of $\mathcal{E}$ does not affect the realization of $\mathcal{E}'$.

Have we succeeded in transferring our intuition into quantitative reasoning?

1.4 Disjoint versus independent events

Two events $\mathcal{E}, \mathcal{E}' \subseteq \Omega$ are *disjoint* when $\mathcal{E} \cap \mathcal{E}' = \emptyset$. Are disjoint events similar to the previous intuitive idea of independence?

[5] This term, "independently" does not yet make sense. We haven't said what "independence" formally means. We do this below (and then everything will make sense).

It is a *common mistake* to confuse "disjointness" and "independence".

Consider an experiment and two disjoint events $\mathcal{E}, \mathcal{E}'$ expressed after we formalize the experiment in terms of probability spaces. Intuitively, disjoint events give rise to very strong dependence. If $\mathcal{E}$ happens then we know that $\mathcal{E}'$ cannot happen (for sure). In a sense, this is the opposite of being independent (they are "fully dependent").

Statistically disjoint events

How about disjoint events? For disjoint events $\mathcal{E} \cap \mathcal{E}' = \emptyset$ by definition $\Pr[\mathcal{E} \cap \mathcal{E}'] = 0$. Formally speaking, $\mathcal{E}$ and $\mathcal{E}'$ can never be independent because their product has to be zero (i.e. unless one of them is nothing – the empty set). Note that for $\mathcal{E}, \mathcal{E}'$ disjoint events we have $\Pr[\mathcal{E} \cup \mathcal{E}'] = \Pr[\mathcal{E}] + \Pr[\mathcal{E}']$.

You should formally explain using the definition of probability that for disjoint $\mathcal{E}$ and $\mathcal{E}'$ we have $\Pr[\mathcal{E} \cup \mathcal{E}'] = \Pr[\mathcal{E}] + \Pr[\mathcal{E}']$.

We stress out that:

- $\Pr[\mathcal{E} \cup \mathcal{E}'] = \Pr[\mathcal{E}] + \Pr[\mathcal{E}']$ is a *property* of disjoint sets $\mathcal{E}$ and $\mathcal{E}'$. Property means that this is a provable consequence of the definition of probability space.
- In contrast, for independent $\mathcal{E}, \mathcal{E}'$ we have $\Pr[\mathcal{E} \cap \mathcal{E}'] = \Pr[\mathcal{E}] \Pr[\mathcal{E}']$, which was a definition (not some provable consequence).

Remark 5. *It helps to remember that the "AND" (the intersection = common points) of independent events corresponds to a product, and the "OR" (the union = put everything together) of disjoint events to a sum.*

These definitions work extremely well together with reality. Let us consider the experiment "independently flip two unbiased coins". Consider the event $\mathcal{E}$ = "the outcome of the *first* coin in HEADS", and the event $\mathcal{E}'$ = "the outcome of the *second* coin in HEADS".

What is the probability that when we finish flipping both coins both events have happened?

Intuition: The first event $\mathcal{E}$ refers only to the first coin and the event $\mathcal{E}'$ refers only to the second coin. At an intuitive level if the coin flips where "independent" then the outcome of the first coin flip should not affect the outcome of the second.

Formal verification of independence: We have $\mathcal{E} = \{(H,H),(H,T)\}$, and $\mathcal{E}' = \{(H,H),(T,H)\}$. Note that $\Pr[\mathcal{E}] = \Pr[\mathcal{E}'] = \frac{1}{2}$. Therefore, $\Pr[\mathcal{E}]\Pr[\mathcal{E}'] = \frac{1}{4}$. Furthermore, the event $\mathcal{E}'' = \mathcal{E} \cap \mathcal{E}' = \{(H,H)\}$, and thus $\Pr[\mathcal{E}''] = \frac{1}{4}$. Therefore, $\Pr[\mathcal{E} \cap \mathcal{E}'] = \Pr[\mathcal{E}]\Pr[\mathcal{E}']$, which according to our definition of independence means that $\mathcal{E}, \mathcal{E}'$ *are (formally) independent*.

Here is what we did so far. We gave two *definitions*: one for product space and one for independence. Then, we modeled two intuitive events, one that was referring only to the first coin flip and the second only to the second. Finally, we observed that it happened that the definition of product space satisfied the definition of independence for these two events. Therefore, under these formal definitions our "intuition about independence" coincides with our "definition of independence".

Note that this $\frac{1}{4}$ is *not* the same $\frac{1}{4}$ in the definition of product space $\frac{1}{4}$ = "probability of heads in a single flip" × "probability of heads in a single flip" $= \frac{1}{2} \cdot \frac{1}{2}$. Rather, it is $\Pr[\mathcal{E}] \cdot \Pr[\mathcal{E}'] = \frac{1}{2} \cdot \frac{1}{2}$ and thus $\mathcal{E}$ and $\mathcal{E}'$ are formally independent.

Never confuse: the probability Pr["HEADS in a single flip"] is a probability calculated in the space $\Omega_1 = \{H,T\}$, whereas the probability Pr["first coin comes HEADS"] is calculated over the space $\Omega' = \{(H,H),(H,T),(T,H),(T,T)\}$. The first $\frac{1}{2}$ is the probability of the event $\{H\}$ in Ω_1, whereas the second $\frac{1}{2}$ is the probability of the event $\{(H,T),(H,H)\}$ in Ω'.

Let us take things further. We can get a better understanding when working with an Ω, which has more than two elements. Say that $\Omega_1 = \{\text{face } 1, \ldots, \text{face } 6\}$ where all elementary probabilities are equal and say the same for $\Omega_2 = \{\text{face } 1, \ldots, \text{face } 6\}$. Now, consider the product space $\Omega' = \Omega_1 \times \Omega_2$. The event $\mathcal{E} =$ "the first die's outcome is 1" is $\mathcal{E} = \{(1,1),(1,2),(1,3),(1,4),(1,5),(1,6)\}$. Then, $\Pr[\mathcal{E}] = \frac{1}{36} + \frac{1}{36} + \frac{1}{36} + \frac{1}{36} + \frac{1}{36} + \frac{1}{36} = \frac{1}{6}$. This sum of $\frac{1}{36}$'s is not as boring as it looks like. By definition $\Pr[(1,1)] = \frac{1}{6} \cdot \frac{1}{6}$ and thus $\Pr[\mathcal{E}] = \frac{1}{6} \cdot \frac{1}{6} + \frac{1}{6} \cdot \frac{1}{6} + \frac{1}{6} \cdot \frac{1}{6} + \frac{1}{6} \cdot \frac{1}{6} + \frac{1}{6} \cdot \frac{1}{6} + \frac{1}{6} \cdot \frac{1}{6} = \frac{1}{6}\left(\frac{1}{6} + \frac{1}{6} + \frac{1}{6} + \frac{1}{6} + \frac{1}{6} + \frac{1}{6}\right) = \frac{1}{6} \cdot 1$. This factorization where one term equals to 1 is not a coincidence.

Say, more generally, that $(\Omega_1 = \{e_1, \ldots, e_k\}, \Pr_{\Omega_1})$ and $(\Omega_2 = \{h_1, \ldots, h_\ell\}, \Pr_{\Omega_2})$, and let the product space $\Omega' = \Omega_1 \times \Omega_2$. An event which refers to only the first part of the joint experiment in the product space can be *always* written as $\underbrace{\text{"event in the single space } \Omega_1\text{"}}_{\mathcal{E}_{\text{in } \Omega_1}} \times \Omega_2$. But since Ω_2 is a probability space $\Pr_{\Omega_2}[\Omega_2] = 1$. Therefore, for any event in Ω_1, say for example $\mathcal{E}_{\text{in } \Omega_1} = \{e_1, e_2, e_3\}$ we have[6]

$$\begin{aligned}\Pr_{\Omega'}[\mathcal{E}_{\text{in } \Omega_1} \times \Omega_2] = \quad & \Pr_{\Omega'}[(e_1,h_1)] + \Pr_{\Omega'}[(e_1,h_2)] + \cdots + \Pr_{\Omega'}[(e_1,h_\ell)] \\ & + \Pr_{\Omega'}[(e_2,h_1)] + \Pr_{\Omega'}[(e_2,h_2)] + \cdots + \Pr_{\Omega'}[(e_2,h_\ell)] \\ & + \Pr_{\Omega'}[(e_3,h_1)] + \Pr_{\Omega'}[(e_3,h_2)] + \cdots + \Pr_{\Omega'}[(e_3,h_\ell)]\end{aligned}$$

Now, we proceed similarly to the above and factor out appropriately.

[6] Recall that $\mathcal{E}_{\text{in } \Omega_1} \times \Omega_2$ is just a set. The subsets of the space $\Omega' = \Omega_1 \times \Omega_2$ are the events whose probabilities we are measuring.

$$
\begin{aligned}
\Pr_{\Omega'}[\mathcal{E}_{\text{in}\,\Omega_1} \times \Omega_2] = \quad & \Pr_{\Omega_1}[e_1]\big(\underbrace{\Pr_{\Omega_2}[h_1] + \cdots + \Pr_{\Omega_2}[h_\ell]}_{\text{this is the entire }\Omega_2}\big) \\
& + \Pr_{\Omega_1}[e_2]\big(\Pr_{\Omega_2}[h_1] + \cdots + \Pr_{\Omega_2}[h_\ell]\big) \\
& + \Pr_{\Omega_1}[e_3]\big(\Pr_{\Omega_2}[h_1] + \cdots + \Pr_{\Omega_2}[h_\ell]\big) \\
= \quad & \Pr_{\Omega_1}[e_1] \cdot 1 + \Pr_{\Omega_1}[e_2] \cdot 1 + \Pr_{\Omega_1}[e_3] \cdot 1 \\
= \quad & \Pr_{\Omega_1}[\{e_1, e_2, e_3\}] = \Pr_{\Omega_1}[\mathcal{E}_{\text{in}\,\Omega_1}]
\end{aligned}
$$

That is,

$$\Pr_{\Omega'}[\mathcal{E}_{\text{in}\,\Omega_1} \times \Omega_2] = \Pr_{\Omega_1}[\mathcal{E}_{\text{in}\,\Omega_1}]$$

Some attention is needed here. The probability we started to calculate $\Pr_{\Omega'}[\mathcal{E}_{\text{in}\,\Omega_1} \times \Omega_2]$ is over the product space Ω', whereas the probability we ended up with in this calculation $\Pr_{\Omega_1}[\mathcal{E}_{\text{in}\,\Omega_1}]$ is the probability computed over Ω_1.

None of these remarks is surprising. When we define a product space we multiply each element of the first Ω_1 space with all the elements in Ω_2 and furthermore we multiply the corresponding probabilities. Therefore, for every event that *refers only* to the first space in the final product space its second part gets multiplied with all possible outcomes of the second space (in the product). But, "all possible outcomes" themselves sum up to 1 and thus in a precise sense the second space does not affect the final calculation.

All told, a product space by definition corresponds to a space that has statistical independence between the constituent spaces – i.e. we can think of product spaces having *"built-in" independence*. For an event $\mathcal{E} = \mathcal{E}_{\text{in}\,\Omega_1} \times \Omega_2$ and a second event $\mathcal{E}' = \Omega_1 \times \mathcal{E}_{\text{in}\,\Omega_2}$, a calculation similar to the one we did above yields $\Pr_{\Omega'}[\mathcal{E} \cap \mathcal{E}'] =$

$\Pr_{\Omega_1}[\mathcal{E}_{\text{in }\Omega_1}] \cdot \Pr_{\Omega_2}[\mathcal{E}_{\text{in }\Omega_2}]$. You should make this calculation in its generality (try first for spaces that have 3-4 elements each) and formally derive $\Pr_{\Omega'}[\mathcal{E} \cap \mathcal{E}'] = \Pr_{\Omega_1}[\mathcal{E}_{\text{in }\Omega_1}] \cdot \Pr_{\Omega_2}[\mathcal{E}_{\text{in }\Omega_2}]$, which shows that $\mathcal{E}$, $\mathcal{E}'$ are formally independent.[7]

Therefore, the two definitions, the definition of product space and the definition of statistical independence, are very well related.

Do not go any further before you understand all of the above.

Let us now come back to the general notion of independence.

Example 6. *Often times a probability space will only be defined implicitly. That is, instead of a detailed measure-theoretic description, we may only have some properties of the space. This is* not *an informal treatment. In fact, in many common practical situations this will be the case. The information provided will be* sufficient to carry out exact, formal calculations. *Consider an experiment where we choose an individual who studies at Rutgers University right now. This choice is made using a given sampling method according to which the probability that a random student is "left-brained" is* 0.6 *and "right-brained" is* 0.4*. Say also that the probability that the student studies "sciences" is* 0.25*, and say also that the probability of being both left-brained and studying sciences is* 0.15*. Then, we can see that if we sample one student* $\Pr[\textit{student studies sciences AND student is left-brained}] = 0.15 = 0.6 \cdot 0.25$*. That is, the two events "student studies sciences" and "student is left-brained" are statistically independent.*

It just "happened" that the probability measurements worked in a way that happened to satisfy the definition of statistical independence (in which case we informally say that there is no statistical correlation between the events "student studies sciences" and "student is left-brained").

A few remarks are in order.

[7] This statement doesn't make sense because $\mathcal{E}$ and $\mathcal{E}'$ do not appear in the RHS of $\Pr_{\Omega'}[\mathcal{E} \cap \mathcal{E}'] = \Pr_{\Omega_1}[\mathcal{E}_{\text{in }\Omega_1}] \cdot \Pr_{\Omega_2}[\mathcal{E}_{\text{in }\Omega_2}]$. But, it's easy to see that $\Pr_{\Omega_1}[\mathcal{E}_{\text{in }\Omega_1}] = \Pr_{\Omega'}[\mathcal{E}]$ and $\Pr_{\Omega_2}[\mathcal{E}_{\text{in }\Omega_2}] = \Pr_{\Omega'}[\mathcal{E}']$.

First, note that Pr[student studies sciences] is perfectly formal. There is *some* probability space, which is associated with the sampling method (maybe given to us as a black box to use – this black box fully determines the space). We may not know the exact description of the sampling method, but this does not mean it does not formally exist. In the previous example we could do formal calculations without knowing its description. In particular, it does not prevent us from writing e.g. $\Pr[\mathcal{E}]$ where $\mathcal{E}$ = "student studies sciences", because $\mathcal{E}$, formally speaking, *is* a subset of something that is implicitly defined and thus itself is implicit. Still, everything is well-defined.

Second, here statistical independence was not induced by the way we defined any product space (there is no product space in Example 6). Rather, it is "hidden" in the nature of the experiment. This, "hidden" is an intuitive notion and not a mathematical notion (mathematically the events are simply called "independent").

A much more interesting example of "hidden" statistical independence is given latter on in Section 1.8 on p. 21.

1.5 Conditional spaces

Let us start with a picture (cf. Figure 1.1) that gives us a new perspective to what statistical independence means.

The idea of events that affect or not the possibility of realization of other events brings us to *conditional probability spaces*. We wish to quantify the statement "given that event A happens what is the probability of event B happenning?". For example, "conditioned on (given) the fact that the outcome is an even face of a fair die, what is the probability that the outcome is 'face 2 or face 1' ?". This concept is not as simple as it originally sounds[8]. Somehow, we care only to measure the probability A within the context of B, which means that

[8] The first philosophical treatise of the subject was about 250 years ago by reverend Bayes; published in the Philosophical Transactions of the Royal Society of London and is available online http://rstl.royalsocietypublishing.org/content/53/370.

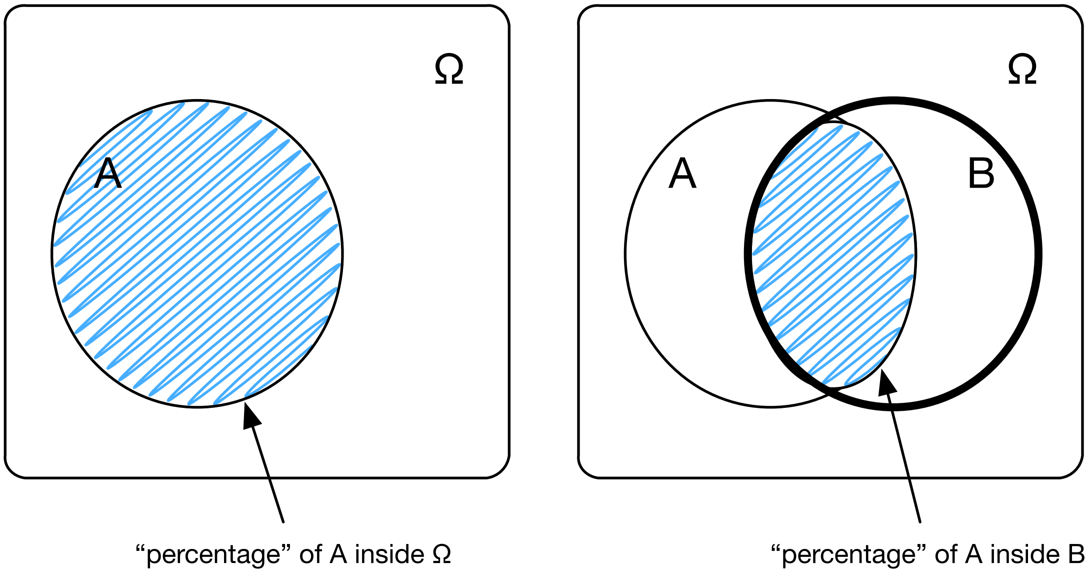


Figure 1.1: The *percentage* (fraction) of A inside Ω (left figure) is equal to the *percentage* of A inside B (right figure). That is, we can now think of B as a new probability space and this corresponds to the real-world intuition that we are measuring the probability of A happening after we know that B already occurred. But because the *percentages* in the left figure (the percentage of A inside Ω) are the same as the one in the right one (the percentage of A inside B) we do not experience any difference regarding the realization of A in the experiment even if someone tells us in advance that B has happened. Intuitively, this seems to be another way to say that B is independent of A.

in a sense event B now itself becomes a probability space. Probability spaces have measure 1, therefore a simple way to address this is to re-weight things by normalizing by $\Pr[B]$. The definition[9] of "probability of A given B" denoted as $\Pr[A|B]$ is

$$\Pr[A|B] = \frac{\Pr[A \cap B]}{\Pr[B]}$$

This is exactly what we intuitively expect (the part of A inside B).

Then, Pr[*the outcome of rolling a fair die is 'face 2 or face 1'*, given that *the outcome is an even face*] $= \frac{\frac{1}{6}}{\frac{1}{2}} = \frac{1}{3}$.

[9] For an event B with non-zero support $\Pr[B] > 0$.

The notation $\Pr[A|B]$ is *not* a probability measure when both A, B vary. But, if we fix B then $\Pr[\cdot|B]$ measures things that sum up to 1.

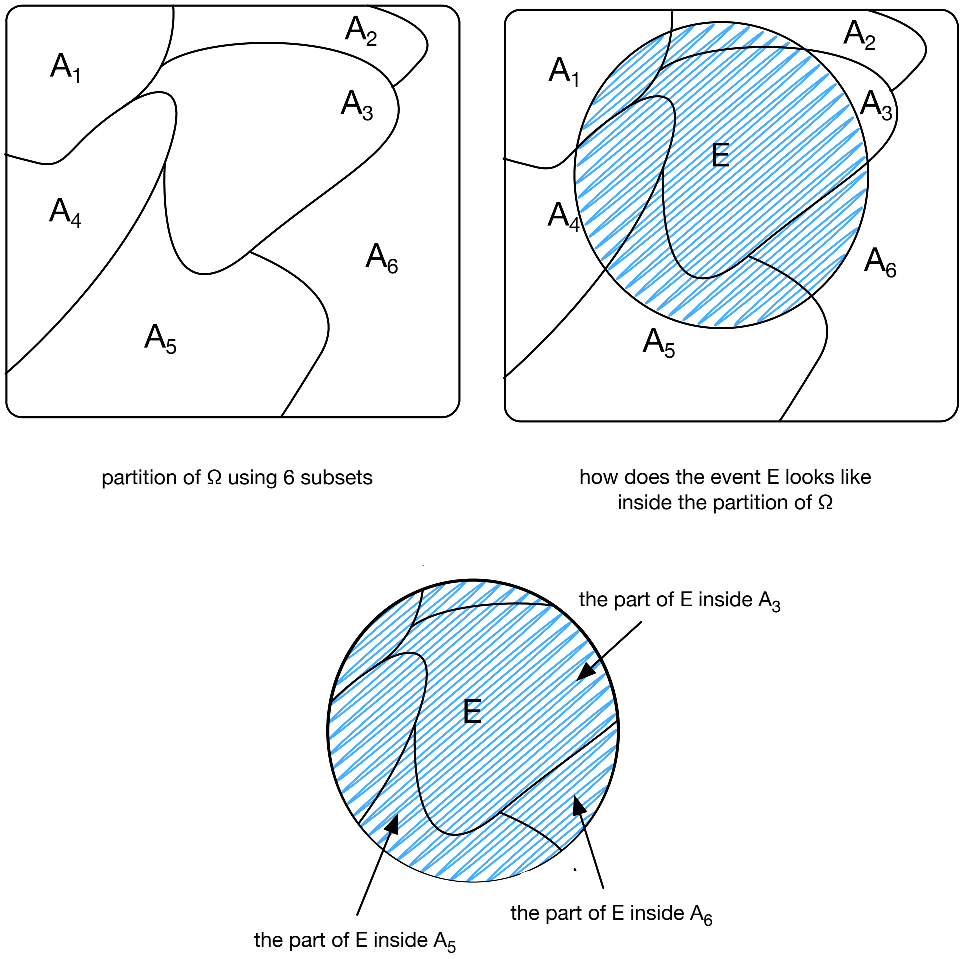
In fact, if we fix B to be a constant and we run over different events A then we have the following $\Pr[B|A] = \frac{\Pr[A\cap B]}{\Pr[A]} \implies \Pr[A \cap B] = \Pr[B]\Pr[B|A]$. But then, $\Pr[A|B] = \frac{\Pr[A\cap B]}{\Pr[B]} = \frac{1}{\Pr[B]}\Pr[A]\Pr[B|A]$. This immediate consequence of the definition is called *Bayes Theorem*. Since, in our specific application "$\Pr[B] =$ constant" we have that $\Pr[A|B] \propto \Pr[A]\Pr[B|A]$. Let us just mention that the probabilities $\Pr[A|B]$ and $\Pr[B|A]$ sometimes gain physical meaning and then we talk about the "a priori" and "a posteriori" probabilities.

Independence and conditional probability By our definitions of statistical independence and conditional probability, if A, B are independent and if $\Pr[B] > 0$, then $\Pr[A|B] = \frac{\Pr[A\cap B]}{\Pr[B]} = \frac{\Pr[A]\Pr[B]}{\Pr[B]} = \Pr[A]$. This $\Pr[A|B] = \Pr[A]$ formalizes better the concept that the outcome of B "does not affect" the probability of A happening. Again, we stress out that statistical independence is somewhat cumbersome. In some sense, it expresses that the *"**proportion** of A stays the same inside the original space and inside B"*. The notion of statistical independence is the most important notion over all probability theory. It gives probability theory meaning and context in places where the so-called general *Measure Theory* never cares to look at[10].

Conditional probability and an important consequence The formula $\Pr[A|B] = \frac{\Pr[A\cap B]}{\Pr[B]}$ is sometimes called "definition of conditional probability". This is just a definition and nothing more than that. Now, we state and prove Theorem 7. This is a mathematical statement (i.e. a property derived by manipulating the definitions).

[10] Measure Theory is a branch of modern mathematics to which probability theory can be understood as a special case. What we discuss here may become problematic when instead of a finite Ω we have an infinite one. Several types of infinity then become of interest. Furthermore, even over "simple" Ω's, e.g. $\Omega = [-1, 1]$, not every subset of Ω can be associated with probability measure. You read this and now you can promptly forget it.

We say that $A_1, \ldots, A_k \subseteq \Omega$ is a *partition of* Ω if for every $i \neq j \in \{1, \ldots, k\}$ we have that $A_i \cap A_j = \emptyset$ and $A_1 \cup A_2 \cup \cdots \cup A_k = \Omega$.



partition of Ω using 6 subsets

how does the event E looks like inside the partition of Ω

the partition of Ω **induces** a partition on E

Figure 1.2: A partition $A_1, A_2, \ldots, A_6$ of Ω *induces* the partition $\mathcal{E} \cap A_1, \mathcal{E} \cap A_2, \ldots, \mathcal{E} \cap A_6$ of $\mathcal{E}$.

Here is an easy exercise (just use the definition of $\Pr[\mathcal{E}]$). Let disjoint $\mathcal{E}, \mathcal{E}'$, i.e. $\mathcal{E} \cap \mathcal{E}' = \emptyset$. Then, $\Pr[\mathcal{E} \cup \mathcal{E}'] = \Pr[\mathcal{E}] + \Pr[\mathit{event}']$. Furthermore, show that in general (i.e. not necessarily for disjoint $\mathcal{E}'', \mathcal{E}'''$) it holds that $\Pr[\mathcal{E}'' \cup \mathcal{E}'''] = \Pr[\mathcal{E}''] + \Pr[\mathcal{E}'''] - \Pr[\mathcal{E}'' \cap \mathcal{E}''']$ (in particular, $\Pr[\mathcal{E}'' \cup \mathcal{E}'''] \leq \Pr[\mathcal{E}''] + \Pr[\mathcal{E}''']$).

Theorem 7. *Let a probability space Ω, a partition of the space $A_1, \ldots, A_k$, and an event $\mathcal{E} \subseteq \Omega$. Then,*

$$\Pr[\mathcal{E}] = \Pr[\mathcal{E}|A_1] \Pr[A_1] + \Pr[\mathcal{E}|A_2] \Pr[A_2] + \ldots + \Pr[\mathcal{E}|A_k] \Pr[A_k]$$

Proof. Note, that we can use the partition of Ω to partition $\mathcal{E}$. That is, $\mathcal{E} = \left(\mathcal{E} \cap A_1\right) \bigcup \cdots \bigcup \left(\mathcal{E} \cap A_k\right)$ and any two $\left(\mathcal{E} \cap A_i\right) \cap \left(\mathcal{E} \cap A_j\right) = \emptyset$ (draw a picture with three sets A_1, A_2, A_3 to visually verify this).

Since for two disjoint events A and B, $\Pr[A \cup B] = \Pr[A] + \Pr[B]$ and the same rule generalizes to unions of more than two sets, we have $\Pr[\mathcal{E}] = \Pr[\left(\mathcal{E} \cap A_1\right) \bigcup \cdots \bigcup \left(\mathcal{E} \cap A_k\right)] = \Pr[\mathcal{E} \cap A_1] + \ldots + \Pr[\mathcal{E} \cap A_k]$. Now, apply the condition probability definition: $\Pr[\mathcal{E}] = \Pr[\mathcal{E}|A_1] \Pr[A_1] + \ldots + \Pr[\mathcal{E}|A_k] \Pr[A_k]$. □

Later on, we will see uses of this theorem. Theorem 7 is used when we can easily compute $\mathcal{E}$ conditioned on the fact that say A_1 and A_2 has occurred, and we also know the probability measure of A_1, A_2.

1.6 Random variables

The spaces we encountered so far contain elements without any numerical meaning. For example, the space of a fair die roll $\Omega = \{\text{face } 1, \text{face } 2, \ldots, \text{face } 6\}$ does not consist of numbers. Of course, we could have written it as $\Omega = \{1, 2, \ldots, 6\}$, but it would have been the same. *The reason is that so far we did not use the outcomes as numbers; e.g. we did not add them up.*

For us, "use as numbers" means to add them up, multiply them, and compute averages. We kept writing "face 1" instead of 1 to emphasize that there was no other intended calculation with the outcome.

A *random variable* X is a function $X : \Omega \to \mathbb{R}$. We use the term "variable" to talk about an object, which is a function not because we want to cause confusion but for historical reasons.

For example, $X(\text{face } 1) = 1$, $X(\text{face } 2) = 2, \ldots, X(\text{face } 6) = 6$.

Not all random variables have such trivial connection to probability spaces. We typically care about one experiment, i.e. one probability space, over which we define many random variables.

We denote by $X(\Omega)$ the set of all possible values of X (aka the image of X). The *expected value* (or expectation) of X is defined as

$$E[X] = \sum_{\alpha \in X(\Omega)} \Pr[X = \alpha] \cdot \alpha$$

That is, $E[X]$ is the average value of X weighted with probability.

Remark on terminology 8. *In addition to the historical reason, we call X a "variable" because when it appears inside the "$\Pr[\ldots]$" notation it looks like a variable. For example, $\Pr[X(\omega) = 5]$. An $\omega \in \Omega$ is sampled and we consider the event associated with $X(\omega) = 5$. This looks like as if we sample at random a value directly from $X(\Omega)$. To make things more intuitive we abuse notation and write X instead of $X(\omega)$. Then, X really looks like a variable that assumes a random value, and we can instead write $\Pr[X = 5]$.*

In our fair die example $E[X] = 1 \cdot \frac{1}{6} + \ldots + 6 \cdot \frac{1}{6} = 3.5$.

Remark 9. *Our first example is an anti-example[11]. In this case the expectation is meaningless. There is no interesting physical meaning in the value 3.5; in the sense that we do not really "expect" that a fair die outcome is 3.5. We will see there is a reason for this.*

Another "averaging" quantity is that of *variance of X* defined as

$$\operatorname{Var}[X] = E\left[(X - E[X])^2\right]$$

We stress out that $E[X]$ is just a number, e.g. $E[X] = 42$. Whenever we see an "E" in front of a random variable then this "E" acts like an integral (or summation if you like) turning X into a number.

[11] This set of notes is not a replacement for any good textbook in probability and statistics. It lacks examples.

The expression "$(X - E[X])^2$" is a new random variable. If $E[X] = 42$, then we have a new function: $Y(\omega) = (X(\omega) - 42)^2$. New random variables are built by composing simpler ones.

The variable $Y = (X - E[X])^2$ measures the distance of a X from its average (expected value). Roughly speaking, $E[(X - E[X])^2]$ is the average of the distances of X from its average.

Here is why Remark 9 happens. In case of the fair die this number is very large, i.e. $\text{Var}[X] \approx 2.91$. "Very large" compared to its possible assumed values in the interval $[1, 6]$.

Variance is a very important parameter that describes the behavior of a random variable. If the variance (i.e. the expected squared distance from the expectation) is high then the value of the expectation tells nothing too interesting. Read over Remark 9.

Interesting examples will be developed in the sequel (Sections 1.11 and 1.12).

One property of expectation is that by definition is a *linear operator*.

Lemma 10. *Let X, Y be random variables over the same space Ω and $c \in \mathbb{R}$. Then,*

$$E[cX] = cE[X] \quad \textit{and} \quad E[X + Y] = E[X] + E[Y]$$

(or equivalently $E[cX + Y] = cE[X] + E[Y]$).

The proof is immediate by definition of E.

The same is not true for variance. Recall that formally when we say "Let X, Y" we mean "for all X, Y". Therefore, to prove[12] that the statement is *not* true for variance we should prove that the following *is* true:

$$\begin{aligned} &\text{NOT}\Big(\forall X, Y \text{ we have } \text{Var}[X + Y] = \text{Var}[X] + \text{Var}[Y]\Big) \\ =&\exists X, Y \text{ such that we have } \text{Var}[X + Y] \neq \text{Var}[X] + \text{Var}[Y] \end{aligned}$$

[12] Recall that *"NOT (for all + logical statement)"= "there exists + NOT(logical statement)"*. For example, "the negation of every day in New Jersey is sunny" is equivalent to "there exists a day in New Jersey, which is not sunny".

An example (formally) proves existence. You should make sure that you can give an example showing that Lemma 10 does not hold for variance.

Independent random variables Suppose that X, Y are random variables defined over the same probability space Ω. We will say that X, Y *are independent* if the following corresponding events are independent[13] that is

$$\forall x, y \in \Omega, \qquad \Pr[X = x \text{ AND } Y = y] = \Pr[X = x] \Pr[Y = y]$$

Therefore, in order to say that two *variables* are independent this should hold for every possible value the random variables assume.

If X, Y are independent then we can show that $E[XY] = E[X]E[Y]$. You should verify that this equality holds before going any further. Note that this does not hold for arbitrary X, Y (show this!). Starting from here, it does not take long to see that if X, Y are independent then $\text{Var}[X + Y] = \text{Var}[X] + \text{Var}[Y]$.

These simple facts are left to the reader to verify.

Random variables we care about The random variable that uniformly ranges over $\{1, 2, 3, 4, 5, 6\}$ (i.e. the fair die) is uninteresting. There are many interesting random variables in Statistics, which we are *not* going to discuss. For the courses where this set of notes is used, we have some very specific random variables of interest.

We say that X is an *indicator random variable (RV)*[14] if it takes values $0, 1$. Say that $\Pr[X = 1] = p$ and $\Pr[X = 0] = 1 - p$. Then, observe that $E[X] = p$.

The reason we care about indicator variables is because they will indicate success (=1) and failure (=0) of various events of interest.

[13] Before we defined independent events. Now, we use random variables to designate events.

[14] In the literature these are also called Bernoulli trials.

Also, by summing up indicator variables we can count the number of successes. For example, suppose that we have the indicator RVs X_1, X_2, X_3, X_4, X_5, say all of them parameterized with probability $p = 0.1$. Then, the "number of successes" is a *new random variable* $X = X_1 + X_2 + X_3 + X_4 + X_5$. The expectation of X is easy to compute by the linearity of expectation: $E[X] = E[X_1 + X_2 + X_3 + X_4 + X_5] = E[X_1] + E[X_2] + E[X_3] + E[X_4] + E[X_5] = 0.1 \cdot 5 = 0.5$. If instead we had n indicator variables each distributed with probability p then $E[X_1 + \ldots + X_n] = n \cdot p$.

The variance of an indicator variable X_1 with parameter p is $\text{Var}[X_1] = E\big[(X_1 - E[X_1])^2\big] = E[X_1^2] - E[X_1]^2$. Observe that $X_1^2 = X_1$ because X_1 takes only values 0 and 1. That is, $\text{Var}[X_1] = E[X_1] - E[X_1]^2 = p - p^2 = p(1-p)$.

Is it true that $\text{Var}[X_1 + \ldots + X_n] = np(1-p)$?

No, not in general[15], unless the X_i's are pairwise (i.e. every two of them) independent. This is a really very important point in our narrative.

> *It is not sufficient to know that the X_i's follow a certain probability distribution when we look each of them* in isolation.

For example, it may be the case that $X_1 = X_2$. Then, $E[X_1 + X_2] = 2p$, because expectation is linear regardless of any correlations between the random variables. But is it true that $\text{Var}[X_1 + X_2] = \text{Var}[X_1] + \text{Var}[X_2]$?

We conclude this section leaving two tasks to the reader.

First, you should verify that if the X_i's are pairwise independent then $Var[X_1 + \ldots + X_n] = np(1-p)$.

[15] For example, if $X_1 = X_2$ and $\Pr[X_1 = 1] = 1/2$ then $E[X_1 + X_2] = 2 \cdot \frac{1}{2} = 1$, but $\text{Var}[X_1 + X_2] = \text{Var}[2X_1] = E[4X_1^2] - E[2X_1]^2 = 4\text{Var}[X_1] \neq 2\text{Var}[X_1]$.

Second, try to understand if there exist variables, which are pair-wise independent but they are not three-wise independent[16].

1.7 How do we express things and why do we write them as such

Probability theory was properly formalized (axiomatized) by Kolmogorov[17] in the 1930s. Before 1930s people were also reasoning about probability. For example, Bayes' article was written 150 years before Kolmogorov's work. When the world was young, probability was a mess, often times wrong, and not usable. The pre-Kolmogorov era inherited us the notation $\Pr[\ldots]$. In fact, it inherited us more than the notation – a way of expressing ourselves about probabilities.

Think about it. We can define $\Omega = \{\text{face } 1, \text{face } 2, \ldots, \text{face } 6\}$, then $\Pr[\text{face } i] = \frac{1}{6}$ for every $i = 1, 2, \ldots, 6$. Then, say that $X(\text{face } i) = i$. Finally, define the event $\mathcal{E} = \big\{\omega \mid X(\omega) \text{ is even}\big\} = \{2, 4, 6\}$. That is, the event is defined by the predicate "$X(\omega)$ is even". At the end, we calculate $\Pr[\mathcal{E}] = \frac{1}{2}$.

Now, instead of all these we could have simply written, $\Pr[X \text{ is even}] = \frac{1}{2}$, which *means exactly the same thing* as the small paragraph above. Even those obsessed with mathematical formalism would have found "$\Pr[X \text{ is even}] = \frac{1}{2}$" much cleaner than the detailed formal description. We loose nothing in formality if the translation of "$\Pr[X \text{ is even}] = \frac{1}{2}$" can be done in our heads.

Often times we may begin by writing, for example: "Consider the random variables X, Y". This implies that there is an underlying probability space associated with these random variables. When obvious we will not explicitly mention the space (but it is always there).

[16] A collection of random variables $X_1, \ldots, X_n$ is three-wise independent if for *every* distinct three variables $X_i, X_j, X_k \in \{X_1, \ldots, X_n\}$ and for every $x, y, z \in \Omega$ we have $\Pr[X = x \text{ AND } Y = y \text{ AND } Z = z] = \Pr[X = x]\Pr[Y = y]\Pr[Z = z]$. Note that pair-wise, three-wise, four-wise, and so on, notion of independence are restrictions of the notion of independence of all variables (which coincides with n-wise independence).

[17] See http://www.kolmogorov.com/Foundations.html.

Another point of confusion is when one says "random variable" instead of simply saying a "sample". For example, consider a space that consists of binary strings $\{00, 01, 10, 11\}$ each with the same probability. Then, someone may write $\Pr[X = 00]$, calling X a "random variable" (instead of calling it "sample"). X is not real-valued[18] and we cannot compute expectations or variances for such an X. However, we will allow abuse of terminology and call X a "random variable".

Finally, we will use the terms "distribution" and "random variable" interchangeably. In fact, one could have introduced terms such as "probability mass/density", "probability distribution", and so on. This type of terminology is unnecessary for our purposes. We will also not explain why a function is different from a distribution. None of these are hard to explain, but they are not necessary for us.

1.8 Examples of "hidden" statistical dependence and independence

Let us now discuss some very interesting examples.[19].

Consider three indicator random variables X_1, X_2, X_3, and their sum, which is calculated over the integers, $X = X_1 + X_2 + X_3$.

Now, let us define two other random variables. Consider the representation of the sum of the X_i's in binary notation. Their sum can be $0, 1, 2, 3$, which in binary is $00, 01, 10, 11$. We associate the first (most significant) digit of X with the random variable b_1 and the second digit with b_0. That is, X is written in binary as $b_1 b_0$; i.e. the new random variables b_1 and b_0 take $\{0, 1\}$ values and put together they form the binary numbers $00, 01, 10, 11$.

Do you think that the *digits* of the sum of independent random variables are independent?

[18] Advanced comment: we can define X's over measurable spaces (not necessarily $\mathbb{R}$), but this X in the example is not measurable in any interesting way.

[19] **REMINDER:** This material is *copyrighted* (10/2015) and in particular the treatment in this example. Any use is prohibited, unless this set of notes is *explicitly cited* or with the *written permission* of the author.

Are the digits of the sum statistically correlated with each other? If b_0 and b_1 are independent[20] then *for all* $\alpha, \beta \in \{0,1\}$ holds

$$\Pr[b_0 = \alpha, b_1 = \beta] = \Pr[b_0 = \alpha]\Pr[b_1 = \beta]$$

(or that $\Pr[b_0 = \alpha | b_1 = \beta] = \Pr[b_0 = \alpha]$).

We begin by determining the probability that X equals $00, 01, 10, 11$. $X = 00$ (i.e. $b_1 = 0$ and $b_0 = 0$) only when $X_1 = X_2 = X_3 = 0$, i.e. $\Pr[X = 00] = \frac{1}{8}$; $X = 01$ if exactly one of the X_i's is 1, i.e. $\Pr[X = 01] = \frac{3}{8}$; similarly, $\Pr[X = 10] = \frac{3}{8}$ and $\Pr[X = 11] = \frac{1}{8}$.

Now, let us come back to checking independence of b_1 and b_0.

First check, $b_0 = 0$ and $b_1 = 0$. The summations that correspond to $b_1 = 0$ are $\{00, 01\}$ and to $b_0 = 0$ are $\{00, 10\}$, and thus $\Pr[b_0 = 0, b_1 = 0] = \Pr[X = 00] = \frac{1}{8} \neq \frac{1}{4} = \frac{1}{2} \cdot \frac{1}{2} = \Pr[b_0 = 0]\Pr[b_1 = 0]$. Since there exist α and β such that $\Pr[b_0 = \alpha, b_1 = \beta] \neq \Pr[b_0 = \alpha]\Pr[b_1 = \beta]$ the variables b_0, b_1 are statistically dependent. Therefore, although the digits b_0 and b_1 are the sum of statistically independent random variables, these digits statistically depend on each other. This is the first non-trivial fact about statistical intuition. It deepens our understanding how a random sum statistically looks like[21].

All told, as random variables the digits b_0 and b_1 depend on each other, because we can find values for b_0 and b_1 where the definition of independence does not hold. On the other hand, there are certain pairs of values for which the two digits do not depend on each other.

Mastering the above two examples significantly boosts one's understanding of statistical dependence and independence.

Let us take things just one step further. Digit b_0 depends on b_1 and this is witnessed by a difference between $1/4$ and $1/8$. What if the number of variables in the summation increases? Do the following

[20] In $\Pr[X = 1, Y = 2]$ comma means "AND". That is, $\Pr[X = 1 \text{ AND } Y = 2]$.

[21] For example, if it were the case that the digits of a random sum were independent then it would have been the case that we could have put together a simple statistical model to sample a random sum directly! (i.e. without first sampling random X_i's and then adding them up!)

exercise. Consider four indicator variables X_1, X_2, X_3, X_4. The possible sums written in binary are $000, 001, 010, 011, 100$. Let us now associate the most significant bit with b_2, the middle with b_1, and the least significant one with b_0. Then, are b_0 and b_1 statistically dependent? If yes, does this "dependence" look less important to you than the one before?

Are the digits b_1, b_0 of the sum $X = X_1 + X_2 + X_3$ statistically dependent with the variables X_i that form the sum? This question is extremelly interesting for someone who wants to understand what statistical independence means ("statistical independence" is not the same at all as some casual notion of independence).

Intuitively, we would expect that the variables X_1, X_2, X_3 are related to b_0 and b_1 (after all these X_i's determine b_0, b_1). The truth is not as simple as this intuition suggests.

Does b_0 statistically depends on X_1? We calculate $\Pr[b_0 = 0, X_1 = 0] = \frac{1}{4} = \Pr[b_0 = 0] \Pr[X_1 = 0]$. Same for $\Pr[b_0 = 0, X_1 = 1]$ and $\Pr[b_0 = 1, X_1 = 0]$ and $\Pr[b_0 = 1, X_1 = 1]$. Therefore, the least significant digit *is independent* of the value of X_1 (or of any other variable). Can you see why intuitively this is the case?

The same observation does *not* hold if instead of b_0 we consider b_1. It also does not hold if instead of only one variable X_1 we consider more, e.g. X_1, X_2, X_3 (i.e. when we consider the probability conditioned on $X_1 = \beta_1, X_2 = \beta_2, X_3 = \beta_3$).

Study all these examples very carefully.

Remark 11. *The examples in this section (Section 1.8) are probably the most indicative examples of statistical independence in the probability literature! Why does the least signficant digit of the number representing the sum does not statistically depend on what we are summing? Why every other digit does? Why when the first and the second digits differ then they are statistically independent and when they are equal they depend on each* *other? The formal explanations (through calculations) are all given above, but developping intuition about all these will probably take time.*

1.9 Common distributions and useful tools

The most basic distribution is the *Bernoulli trial*, it assumes values $\{0, 1\}$ with parameter p, where p is the probability of 1.

The distribution that quantifies the probability of k successes (i.e. k-many 1s) "until the first failure" using i.i.d. Bernouli trials is called *geometric distribution*.

We have a special interest in the behavior of sums of i.i.d. Bernoulli trials. This measures the number of 1s in $X = X_1 + X_2 + \cdots + X_n$, and is called the *binomial* distribution.

Task for the reader Given n and p the probability of $X_i = 1$ make a plot (e.g. use R or Mathematica) of the magnitude of $f(k) = \Pr[X = k]$ and explain where this distribution assumes its highest value. Which of the continuous distributions you learned in your first class that involved statistics has a similar shape?

For the geometric and the binomial distribution we are interested in understanding their "tails" (tail = what happens away from $E[X]$).

Here are some very useful expressions and inequalities.

- For an event $\mathcal{E} \subseteq \Omega$ and its *complement* (with respect to Ω) , i.e. $\bar{\mathcal{E}} = \Omega - \mathcal{E}$, we have $\Pr[\bar{\mathcal{E}}] = 1 - \Pr[\mathcal{E}]$.
- (union bound) For *any* collection (i.e. arbitrarily correlated) of events $\mathcal{E}_1, \ldots, \mathcal{E}_n$ we have $\Pr[\mathcal{E}_1 \cup \cdots \cup \mathcal{E}_n] \leq \Pr[\mathcal{E}_1] + \cdots + \Pr[\mathcal{E}_n]$
- $\left(\frac{n}{e}\right)^k \leq \binom{n}{k} \leq n^k$, where $e \approx 2.718$
- $\lim_{n\to\infty} \left(1 - \frac{1}{n}\right)^n = \frac{1}{e}$
- $\frac{1}{4} \leq \left(1 - \frac{1}{n}\right)^n \leq \frac{1}{e}$

1.10 Important inequalities

The most basic inequality is Markov's.

Theorem 12 (Markov's inequality)**.** *Let X be a non-negative random variable and $c > 0$ and arbitrary real number. Then,*

$$Pr[X \geq c] \leq \frac{E[X]}{c}$$

This inequality relates *probability* of a random variable attaining high values with its *expectation*.

This probability is as measuring what happens when we "do the experiment once", whereas the expectation is an average[22].

Markov's inequality is so general that it cannot be super useful on its own (there are only a few restricted cases where it is used on its own). For example, let us replace c with another constant $c = kE[X]$. Then, $Pr[X \geq kE[X]] \leq \frac{1}{k}$. This quantifies how unlikely is for a single execution of the experiment to yield a value for the variable that is k times away its expectation.

Here is what restricts its applicability. Let $X = X_1 + X_2 + \ldots + X_{10}$, where $E[X_i] = 0.5$ for X_i's, where each X_i is an independent coin flip of an unbiased coin (say 1=HEADS and 0=TAILS). Then, X counts the number of HEADS. Note that $E[X] = 5$. Then, $\Pr[X \geq k \cdot 5] \leq \frac{1}{c}$. Think of c as indicating a probability of error — i.e. how far away from the expectation we go. To bound this probability *using Markov* by less than 50% we should set $k > 2$. This means that that the event is $X > 10$ which can never happen. It is amusing that Markov is telling us that this event can happen with probability at most e.g. 49%. But we already know that this event can happen with probability at most 0% because we only have 10 variables. We do not need any inequality to tell us this.

[22] In fact, "probability" is also an averaging quantity of some short, but if this remark confuses you, then you read it and promptly forget it.

Markov is definitely not useless. It is useful in certain cases. Moreover, it is very important in deriving new, stronger inequalities, but in more restricted settings [23].

If we have information about the variance of a variable, and this variance is small, then much more can be achieved.

Theorem 13 (Chebyshev's inequality)**.** *For every a random variable X and $c > 0$ holds that*

$$\Pr[|X - E[X]| \geq c] \leq \frac{Var[X]}{c^2}$$

This inequality relates: (i) the probability of the value of X in one realization of the experiment, (ii) its expectation, and (iii) its variance.

We can prove Chebyshev's by directly substituting a new random variable $Y = (X - E[X])^2$ for X in Markov's inequality (do this).

To prove Markov's is also not hard (this proof can be skipped at a first reading).

Proof of Theorem 12 for discrete random variables. Define $f(x) = 0$ for all $x < c$, and $f(x) = 1$ for all $x \geq c$. Then, although we think of X taking random values it always holds that $c \cdot f(X) \leq X$. It is easy to see that for RVs Y, Z if $Y \leq Z$ then $E[Y] \leq E[Z]$. Therefore, $c \cdot f(X) \leq X \implies E[cf(X)] \leq E[X] \implies cE[f(X)] \leq E[X]$.

$$\begin{aligned} E[f(X)] &= \sum_x \Pr[f(X) = x]x = \Pr[f(X) = 1]1 + \Pr[f(X) = 0]0 \\ &= \Pr[f(X) = 1] = \Pr[X \geq c] \end{aligned}$$

Therefore, $cE[f(X)] \leq E[X] \implies \Pr[X \geq c] \leq \frac{E[X]}{c}$. □

[23] A restricted setting is an interesting one. Generic/abstract and unrestricted mathematical settings typically describe generic/kind-of-obvious facts.

1.11 The concentration of measure phenomenon

Suppose that we perform 1000 independent, unbiased coin flips. If X is the random variable whose value is the total number of HEADS, then $E[X] = 500$. In practice, we do not care only about the average but mostly about the value of X *with high probability*.

Remark on terminology 14. *"High probability" is loosely defined and is determined by context. In some cases it means any constant above* $\frac{1}{2}$*, e.g.* $\frac{2}{3}$*. The term "constant" is also undefined unless there is some quantity growing to infinity. For example, consider a probabilistic experiment*[24]*, parameterized by n, where n is the number of coin flips. More often, "high probability" means probability* $1 - \frac{1}{n}$ *or* $1 - \frac{1}{n^2}$ *or* $1 - \frac{1}{10^n}$*; e.g. for $n = 10$ we have* $1 - \frac{1}{n} = 0.9$ *whereas* $1 - \frac{1}{10^n} = 0.9999999999$*. Depending on the context we may want the "high probability" to converge polynomially fast to* 1*, or in other contexts "high" means exponential fast convergence to* 1*. Also, we may write* almost surely (a.s.) *instead of "with high probability".*

We continue with the goal of understanding the value of X a.s. in the experiment where we i.i.d. flip n unbiased coins. Let $X_i \in \{0, 1\}$ be the random variable, which is 1 if and only if the i-th coin flip is "HEADS". Then, we have $X = X_1 + \cdots + X_n$ and thus $E[X] = E[X_1] + \cdots + E[X_n] = \frac{1}{2} + \cdots + \frac{1}{2} = \frac{n}{2}$.

We are ready to derive our first *probability measure concentration result*, which is on its own quite impressive. By *measure concentration* we mean that most of the probability is around its expectation. "Around" means in a small interval centered at expectation.

For the calculation with Chebyshev we will need two facts. First, the variance of each X_i is $\text{Var}[X_i] = E[X_i^2] - E[X_i]^2$, and since $X_i \in \{0, 1\}$, we have $\text{Var}[X_i] = E[X_i] - E[X_i]^2 = \frac{1}{2} - \left(\frac{1}{2}\right)^2 = \frac{1}{2} - \frac{1}{4} =$

[24] We already saw that every intuitively described experiment corresponds to a formal probability space

$\frac{1}{4}$. Finally, since the X_i's are independent we have that $\text{Var}[X] = \text{Var}[X_1] + \cdots + \text{Var}[X_n] = \frac{n}{4}$.

Now, let us put everything together.

Theorem 15 (Chebyshev sampling)**.** *Let $\varepsilon, p > 0$ be constants. Consider n i.i.d. Bernoulli trials $X_1, \ldots, X_n$, where $E[X_i] = p$. Let $X = \sum_{i=1}^n X_i$, then,*

$$\Pr\left[X > (1+\varepsilon)E[X]\right] < O\left(\frac{1}{n}\right)$$

Proof. Note that $\Pr\left[X > (1+\varepsilon)E[X]\right] < \Pr\left[X > (1+\varepsilon)E[X] \text{ or } X < (1-\varepsilon)E[X]\right] = \Pr\left[\left|X - E[X]\right| > \varepsilon E[X]\right]$. Therefore, by Chebyshev we have $\Pr\left[\left|X - E[X]\right| > \varepsilon E[X]\right] \leq \frac{\text{Var}[X]}{(\varepsilon E[X])^2} = \frac{n\text{Var}[X_1]}{(\varepsilon n E[X_1])^2} = \frac{\text{Var}[X_1]}{n\varepsilon^2 E[X_1]^2} = \frac{1}{n} \cdot \frac{p-p^2}{\varepsilon^2 p^2} = \frac{1-p}{\varepsilon^2} \cdot \frac{1}{n} = O\left(\frac{1}{n}\right)$, since ε and p are constants. □

Thus, just computing the variance we can show that the probability of going e.g. 0.1% above the average decreases polynomially with the number of variables (in practice, each variable X_i corresponds to a repetition of an experiment, or a coin flip, or . . .).

Similarly, to Theorem 15 we obtain that $\Pr\left[X < (1-\varepsilon)E[X]\right] < O\left(\frac{1}{n}\right)$. Therefore, after "one full trial" for X (which consists of n small trials, one for each X_i), the probability that X falls *outside* $[(1-\varepsilon)E[X], (1+\varepsilon)E[X]]$ is at most $O(1/n)$ and thus with probability $1 - O(\frac{1}{n})$, X is inside $[(1-\varepsilon)E[X], (1+\varepsilon)E[X]]$ ("concentrated around $E[X]$").

The probability measure, which in total is 1*, is sharply concentrated around $E[X]$.*

The calculation in the proof says in fact more (in this document "proofs" are just calculations). Even if the variables are pairwise independent (i.e. not fully independent) we still have the same conclusion. The reason is that pairwise independence implies $E[X_iX_j] = E[X_i]E[X_j]$, which in turn suffices for showing $\text{Var}[X] = \text{Var}[X_1] +$

$\cdots + \mathrm{Var}[X_n]$. Recall that the latter in particular means that the *covariance* is $\mathrm{Cov}[X_i, X_j] = E[X_iX_j] - E[X_i]E[X_j] = 0$. In other words, variables that are uncorrelated, as measured by covariance[25], do exhibit measure concentration phenomena. We will see in the next section that full independence (i.e. stronger than pairwise independence) suffices to obtain exponential convergence to 1 (not only $1 - \frac{1}{n}$).

How close to the true concentration of n i.i.d. variables is this bound? Concentration around the expectation with probability $1 - O(\frac{1}{n})$ is very high, but it may be the case that we can do even better when we have independent random variables – recall that the bound holds even if the variables are pairwise independent.

Here is a computer experiment (in Mathematica) which goes as follows: (i) sample independent Bernoulli trials X_i, for $i = 1, \ldots, 10^5$ with probability parameter $\frac{1}{2}$; (ii) at the end sum them up; (iii) repeat fresh starting from (i) for 1000 times. That is, sample $X = X_1 + \cdots + X_{10^5}$ for 1000 times and then plot a histogram (Figure 1.11).

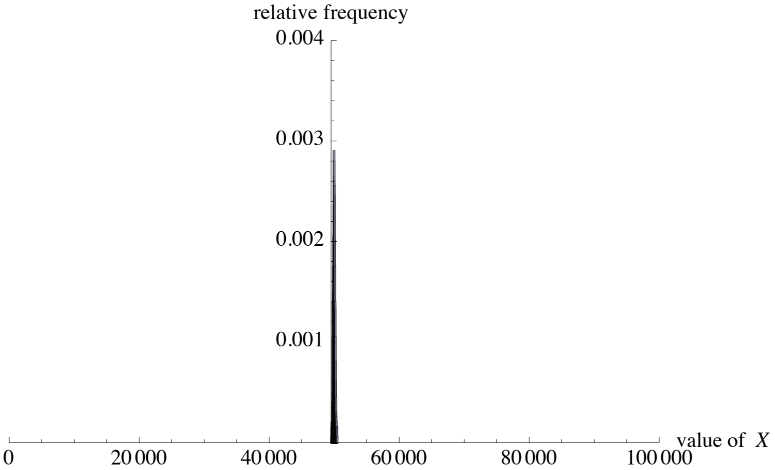


Figure 1.3: Histogram for the value of $X = X_1 + \cdots + X_{10^5}$

[25] Note that zero covariance does not preclude statistical correlations of other forms.

We can see that all the mass of the histogram we plotted is sharply concentrated around the expectation point. Now, if we magnify the region around the expectation we get a clearer picture of the same experiment (Figure 1.11).

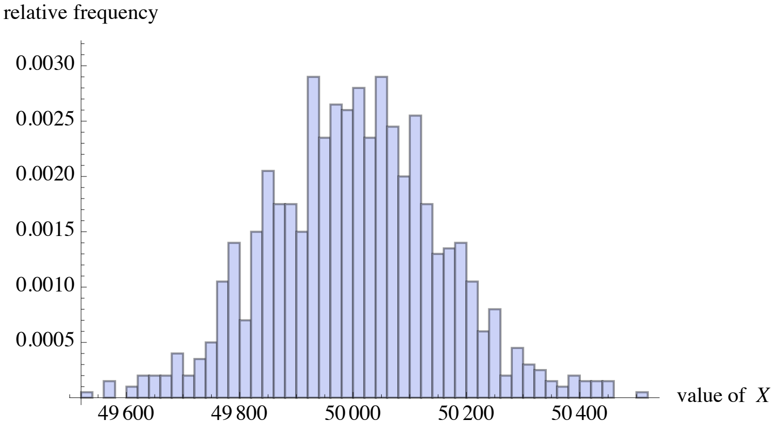


Figure 1.4: Same histogram as in Figure 1.11 but now magnified around $X = 50000$. Observe that no single one among these 1000 repetitions of the expriement resulted anything smaller than 49500 or bigger than 50500.

We observe that in this computer experiment sharp concentration did happen around the expectation. In the proof of Theorem 15 we have that the precise constant (in front of $1/n$) we calculated in the bound is $\frac{1-p}{\varepsilon^2}$. Here, $p = 1/2$ and let us set $\varepsilon = 0.01$. Then, by applying Theorem 15 we have that with probability which is *at most* $\frac{1}{20000} = 0.00005$ we can have the value of X is bigger than 50500 or smaller than 49500.

All these sound very good, since probability 0.00005 of being outside the concentration interval appears to be very small. But this is with probability *at most* 0.00005. Maybe our theorem is not very strong. Maybe a better calculation could have resulted in a better (better=smaller) upper bound. For instace, maybe the truth is that the real upper bound on this probability is even smaller, e.g. 0.0000000000000000005. Of course, at most

0.0000000000000000005 also means at most 0.00005, i.e. we do not challenge whether we have proved Theorem 15 correctly. We challenge whether this bound can be improved.

What if it were true that 0.00005 is the correct upper bound. This means that the bound is "tight", i.e. the probability can be exactly equal to 0.00005. Then, the probability that $X \in [49500, 50500]$ is $(1 - \frac{1}{20000})$ and the probability that *all* of 100000 independent executions are all inside $[49500, 50500]$ would have been $(1 - \frac{1}{20000})^{100000} = \left((1 - \frac{1}{20000})^{20000}\right)^5 \approx \frac{1}{e^5} \approx 0.0067$. That is, the probability that all 100000 independent executions of the experiment are inside $[49500, 50500]$ is about 6.7%. But in our computer simulation/experiment it happened to be the case that all of the executions were inside $[49500, 50500]$. Now, one of the two things have happened. Either we are unlucky and we just hit the event that happens with 6.7% or it is just the case that the 0.00005 is not a "tight" upper bound (can be improved to something smaller).

1.12 Strong measure concentration from independence

We saw that repeating an experiment with two outcomes (0 and 1) can result in concentration $1 - O(1/n)$ around the expected value. Recall that for this it was not necessary to have "full independence". Rather, pairwise independence between the executions was sufficient. Now, we show that there is an amazingly strong concentration around the expectation when we make "full use" of independence among the X_i's.

Let $X_1, \ldots, X_n$ be *independent and identically distributed (i.i.d)* Bernoulli trials with parameter p (i.e. $\{0, 1\}$ distributed random variables that come 1 with probability p). Let also $X = X_1 + \cdots + X_n$ be their sum. We also have that $E[X] = np$. We wish to upper bound the probability $\Pr[X > \Delta]$, for a Δ that we will choose conveniently later on. We remark that if we have any monotonically increasing function F then $\Pr[X > \Delta] = \Pr[F(X) > F(\Delta)]$, because the event "$X > \Delta$" is just a set that satisfies "..." inside "$\Pr[\dots]$" for the corresponding values of X, which are exactly the same as the values in e.g. "$X + 1 > \Delta + 1$" or "$2^X > 2^\Delta$" or more generally "$F(X) > F(\Delta)$". Therefore, for any $\lambda > 0$ we have

$$\Pr[X > \Delta] = \Pr[e^{\lambda X} > e^{\lambda\Delta}] \leq \frac{E[e^{\lambda X}]}{e^{\lambda\Delta}} \tag{1.1}$$

Now, the problem of bounding this probability reduces to the problem of bounding the average $E[e^{\lambda X}]$, where $X = X_1 + \cdots + X_n$. Now, the independence among the X_i's is used to assert that

$$E[e^{\lambda X}] = E[e^{\lambda(X_1+\cdots+X_n)}] = E[e^{\lambda X_1} \dots e^{\lambda X_n}] = E[e^{\lambda X_1}] \dots E[e^{\lambda X_n}] \tag{1.2}$$

, where the last equality is because of independence (this is the only place where we use independence – will be used nowhere else). By definition of expectation: $E[e^{\lambda X_1}] = pe^{\lambda \cdot 1} + (1-p)e^{\lambda \cdot 0} = pe^\lambda + q$, where we set $q = 1 - p$. Therefore, by (1.2) we have that $E[e^{\lambda X}] = (pe^\lambda + q)^n$.

We intentionally left up until now Δ not set to a specific value because this is the first time that it matters what it is. Let us set Δ similarly to $(1+\varepsilon)E[X] = np + \varepsilon pn$, i.e. $\Delta = (p+t)n$, which is a slightly more convenient form for the calculation that follows. Then, by (1.1) we have

$$Pr[X > (p+t)n] \leq \frac{(pe^\lambda + q)^n}{e^{\lambda(p+t)n}} = \left(\frac{pe^\lambda + q}{e^{\lambda(p+t)}}\right)^n$$

The reason that we introduced a λ is the same reason that $\lambda > 0$ is introduced in *Laplace Transform* (the serious reader should check the literature about Laplace Transform and understand why the choice of introducing a free parameter λ in the exponent is not "magic").

Since, the expression holds for all $\lambda > 0$ we apply the monotonicity study (see Calculus 101) to find the λ that minimizes $f(\lambda) = \left(\frac{pe^{\lambda}+q}{e^{\lambda(p+t)}}\right)^n$. By finding and substituting this λ back to (1.1) we have that for $t > 0$

$$\Pr[X > (p+t)n] \leq e^{-n\left((p+t)\ln\frac{p+t}{p}+(q-t)\ln\frac{q-t}{q}\right)}$$

This probability bound is called *Chernoff bound* or *Chernoff-Hoeffding Bound*. This form is the strongest (tightest) probability bound we will derive. However, it is somewhat messy – not very easy to use. By a simple (but not immediate) manipulation this expression easily yields the following theorem[26].

Theorem 16. *Let $X_1, \dots, X_n$ be i.i.d. Bernoulli trials with probability parameter p. Then,*

$$\Pr[X > (1+\varepsilon)E[X]] \leq e^{-\frac{\varepsilon^2}{3}E[X]} = e^{-\frac{\varepsilon^2}{3}pn}$$

and

$$\Pr[X < (1-\varepsilon)E[X]] \leq e^{-\frac{\varepsilon^2}{3}E[X]} = e^{-\frac{\varepsilon^2}{3}pn}$$

Therefore, for a constant probability p and constant ε if we do the experiment once (i.e. flip all n variables), then the probability that the outcome is just a little bit away from $E[X]$ is exponentially small, i.e. $\frac{1}{e^{\Omega(n)}}$. That is, with probability $1 - \frac{1}{e^{\Omega(n)}}$ the value of X will be inside $[(1-\varepsilon)E[X], (1+\varepsilon)E[X]]$. Compare this with the $1 - \frac{1}{\Omega(n)}$ rate we derived before using Chebyshev's inequality.

1.13 Statistical experiments over time: stochastic processes

Throughout this text we keep repeating that every informally (but reasonably) defined experiment immediately translates to a probability space Ω. What happens if the experiment changes over time?

[26] This is just a derivation by: manipulating symbols, use a standard Taylor expansion, making substitutions. It is simple to get and its proof does not provide any probabilistic insight.

What is time? Time can be a continuous quantity, e.g. time $t \in [0, \infty)$. For every application of interest to Elements of Probability and Statistics time progresses in *discrete time steps*, $t \in \{0, 1, 2, 3, \dots\}$. We occasionally *introduce time* in the analysis of an experiment. In these cases there is no physical notion of time associated with our introduced time steps. For example, when we consider n independent $X_1, \dots, X_n$ there is no notion of time here. But in order to be able to use the tools (developed in the next sections) we may artificially assume that there is a time order for the X_i's. A detailed example will be given later on.

How to formalize time? One option is to consider different probability spaces, e.g. $\Omega_1, \Omega_2, \dots$. Another option would be to consider product spaces with possibly infinite coordinates. However, for (mathematically) technical reasons it helps to have *one* space Ω over which we define random variables $X_1, X_2, \dots$, with X_i corresponding to the i-th time-step. Such a sequence of X_i's is called a *stochastic process*. Then, the theory is developed by studying the relations between X_i's. The more interesting and useful findings are when the X_i's are strongly related – the more the restrictions the more meaningful the study.

Discrete memoryless processes An example of a severely restricted stochastic process is one where the next step depends only on the previous step. Formally, for every $i > 1$ and $\alpha, \beta_1, \dots, \beta_{i-1} \in X(\Omega)$, $\Pr[X_i = \alpha | X_1 = \beta_1, \dots, X_{i-1} = \beta_{i-1}] = \Pr[X_i = \alpha | X_{i-1} = \beta_{i-1}]$. This restriction is also great for visualizing such a memoryless process. The fact that the i-th step depends only on the previous one allows us to draw the stochastic process on papers: use one paper for each time step.

A further restriction is when the discrete memoryless process is *time-homogeneous*, i.e. when the behavior of the process is the same for every time step. Formally, $\Pr[X_i = \alpha | X_{i-1} = \beta_{i-1}] = \Pr[X_{i-1} = \alpha | X_{i-2} = \beta_{i-1}]$, i.e. the distributions of the X_i's do not depend on i. They only depend on the value of the previous step (whichever this is). Now, a single graph defines the process. Maybe we will need a paper of infinite size, but still just one paper.

Remark on terminology 17. *Time-homogeneous, discrete memoryless processes are usually called* stationary Markov chains.

Such processes are common in supply chains, actuarial sciences, process engineering, computer engineering, and computer science.

1.14 Martingales and Azuma's inequality

A martingale is a concept different than a Markov process[27]. Markov processes "are processes without memory". Martingales are processes that they "maintain the expected value".

The typical example of a martingale is a fair gambling game. To understand this we need the notion of *conditional expectation*. Let X be a random variable and $\mathcal{E}$ be an event.

$$E[X|\mathcal{E}] = \sum_{\alpha} \alpha \Pr[X = \alpha | \mathcal{E}]$$

In this notation, $E[X|Y]$ is a random variable because it depends on Y (Y is not one event $\mathcal{E}$ – for different values β of Y we consider the event $\mathcal{E} = "Y = \beta"$).

A stochastic process $X_1, X_2, \ldots$ is a martingale if for all $i \geq 2$ holds:

$$E[X_i | X_1, \ldots, X_{i-1}] = X_{i-1}$$

We have a special interest in martingales that do not change too rapidly. Specifically, we say that a martingale $X_1, X_2, \ldots$ satisfies the

[27] There are examples of Markov processes that are not martingales, and of martingales that are not Markov processes.

bounded difference condition if for constants $c_i \geq 0$ and every $i \geq 2$ we have that

$$|X_i - X_{i-1}| \leq c_i$$

Theorem 18 (Azuma's inequality)**.** *Let $X_1, X_2, \ldots$ be a martingale satisfying the bounded difference condition with parameters c_i. Fix $n > 0$ and let $c = \sum_{i=1}^{n} c_i^2$. Then,*

$$\Pr[X_n > X_0 + t] \leq e^{-\frac{t^2}{2c}}$$

and also

$$\Pr[X_n < X_0 - t] \leq e^{-\frac{t^2}{2c}}$$

So, how to use the above in order to show measure concentration?

The serious reader should give serious thought on martingales. Here we presented exactly what we will need for the rest of the class. However, their importance is disproportional to the length of their current presentation. After mastering all topics mentioned in this set of notes you should study what is a filtration of a probability space, what is a Doob's filter, and other related topics.

1.15 Suggested readings

Here is what I consider as the best sources to study the subject.

Introduction to Probability, 2nd Edition
by Dimitris P. Bertsekas and John N. Tsitsiklis

An Introduction to Probability Theory and Its Applications, Vol.1, 3rd ed.
by William Feller

A more advanced text mostly on "continuous" spaces:

Probability, 2nd ed.
by Albert N. Shiryaev

A glimpse on the philosophical interpretation of probability:

Interpretations of Probability (Stanford Encyclopedia of Philosophy)
`https://plato.stanford.edu/entries/probability-interpret/`
by Alan Hajek