A crash course in probability

Periklis A. Papakonstantinou

LECTURE NOTES IN Elements of Probability and Statistics

Periklis A. Papakonstantinou
Harlem, New York, May 2024

Acknowledgments

I am grateful to Maria Pouliopoulou for carefully reading a previous version of the manuscript and for offering constructive criticism. I am also grateful to Tanay Talukdar for repeating many of the calculations and arguments and for discovering subtle mistakes. I would like to thank Jia Xu for the detailed expository remarks and technical suggestions. Finally, let me thank Hafiz Asif, Galib Khan, and Andrea Nemes, my teaching assistants, and students, for their remarks.

Contents

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 material from proper sources, such as the textbooks suggested at the end.

1.1 What is a sample space?

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

i. a set Ω, together with

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

where f over Ω sums to 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 sample spaces may cause confusion. For example, when in a calculation two distinct sample 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 simply a subset of Ω, e.g. $\mathcal{E} \subseteq \Omega$.
For $\mathcal{E} = \{e_1, e_2, \ldots, e_k\}$, define $\Pr[\mathcal{E}] = \Pr[e_1] + \cdots + \Pr[e_k]$.
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 its chance:*

"The probability that the outcome of rolling 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 $\Pr[$fair die outcome is 2 or 4 or 6$]$? Is this a correct probability expression? No, unless there is a rigorously defined sample 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 sample space.*

In real life, we may say "formal statistical model" instead of "sample 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 $\Omega = \{$face 1, face 2, $\ldots$, face 6$\}$ and all faces[2] *are equiprobable, i.e. $\Pr[$face 1$] = \frac{1}{6}, \ldots, \Pr[$face 6$] = \frac{1}{6}$. This is our model of the world.*

[1] The gain of having a formal model is that we can forget about the real world (the 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.

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 "$\Pr[\text{fair die outcome is 2 or 4 or 6}] = \frac{1}{2}$" 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.

It is important to remember that *a sample 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 cannot even be asked. 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 sample space, can we construct other, more interesting spaces?

Below, we answer this question by specifying a sample space where the question can be asked and thus the statistical calculation can be carried out unambiguously.

1.2 Product spaces

Let $(\Omega, \Pr_\Omega)$ be a sample 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.*

Remark on terminology 3. *We decided to subscript* Pr *with each of the corresponding sample 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* sample spaces is itself a uniform space. Recall that "uniform" is a sample 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 natural relationship between product spaces and the notion of "chance" and "probability" in real-life.

[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.

[4] For example, why not add 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 sample 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 elementary event. There is a far more important reason why we decided to define $\Pr[(H, T)]$ as $\Pr[H] \Pr[T]$.

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 – it 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) in our product space. We note that (H, T) is a single elementary outcome in this space; i.e., H or T are not elementary outcomes, only (H, T) is. 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 embodies a *special case* of a phenomenon we call *statistical independence*. There are many ways in which statistical independence arises and "product spaces" is one such way.

We are not restricted to defining the product space over the same Ω. For two sample 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 all elementary outcomes $x \in \Omega_1$ and $y \in \Omega_2$.

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

1.3 From intuition to definition

We will now explain what is the relationship between statistical independence and the values we have chosen to assign to the probabilities in product spaces.

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

[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).

"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, even if I know the outcome of the other coin flip.

The quantitative problem we have to solve now is to give a *formal* definition of independence. Whichever definition we give, this should formalize precisely the above intuitive idea we have about independence.

Statistical independence

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

The above is just a definition. That is, we have no choice but to accept that from this point on the term "independence" means "$\Pr[\mathcal{E} \cap \mathcal{E}'] = \Pr[\mathcal{E}] \Pr[\mathcal{E}']$". Nevertheless, at first glance, it is unclear why a product of probabilities like this one formalizes the concept of statistical independence.

Have we succeeded in transferring our intuition into quantitative reasoning?

Statistical independence and product spaces

Now, we will show the relationship between the two notions we defined/introduced above: product sample space and statistical independence.

Intuition: Consider an experiment that has two phases. We will use our notion of the product space to model these two phases in an experiment that we like to think that it consists of two sub-experiments. And we will argue that the way we defined the product space *and* the way we defined statistical independence are such that "the two phases do not affect each other". To demonstrate this idea we will define two events $\mathcal{E}$ and $\mathcal{E}'$ each referencing exclusively to a different

phase of the experiment and show that $\mathcal{E}$ and $\mathcal{E}'$ are independent. Here is an example that puts the above in proper context.

Verification of independence: Say that Ω^2 is derived from Ω as in Example 4. Let $\mathcal{E}$ =“the first coin flip is heads” and $\mathcal{E}'$ =“the second coin flip is heads”. Let us now verify that $\mathcal{E}$ and $\mathcal{E}'$ are independent. If we spell out as subsets the verbal description of the events we have $\mathcal{E} = \{(H,H),(H,T)\}$, and $\mathcal{E}' = \{(H,H),(T,H)\}$. Note that $\Pr[\mathcal{E}] = \frac{1}{4} + \frac{1}{4} = \frac{1}{2}$ and same for $\Pr[\mathcal{E}'] = \frac{1}{2}$. Therefore, $\Pr[\mathcal{E}]\Pr[\mathcal{E}'] = \frac{1}{4}$. Furthermore, the event $\mathcal{E} \cap \mathcal{E}' = \{(H,H)\}$, and thus $\Pr[\mathcal{E} \cap \mathcal{E}'] = \frac{1}{4}$. 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*.

We remark that the two $\frac{1}{4}$ we calculated above are one-over-fours derived in two different ways. They happen to be equal in value.

Here is what we have done so far. We gave two *definitions*: one for product space and one for independence. Then, we modeled two intuitive events, one referencing only the first coin flip and the second only 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”.

Never confuse: the probability $\Pr[\text{“HEADS in a single flip”}]$ is a probability calculated in the space $\Omega = \{H,T\}$, whereas the probability $\Pr[\text{“first coin comes HEADS”}]$ is calculated over the space $\Omega^2 = \{(H,H),(H,T),(T,H),(T,T)\}$. The first $\frac{1}{2}$ is the probability of the event $\{H\}$ in Ω, whereas the second $\frac{1}{2}$ is the probability of the event $\{(H,T),(H,H)\}$ in Ω^2.

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 that 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 sample 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.

Statistical independence outside product spaces

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

Example 5. *Often times a sample 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 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 a major university. 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 $\Pr[\text{student studies sciences}]$ is perfectly formal. There is an underlying sample space, which is associated with the sampling method. We are not given its exact description, but we are given everything we need to know (about this space) in order to carry out our formal calculations.

Second, here statistical independence was not induced by any product space. There is no product space in Example 5.

A much more interesting example of "implicit" statistical independence is given latter on in Section 1.8 on p. 26.

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?

"Disjointness" is *commonly mistaken* for "independence".

Consider an experiment and two disjoint events $\mathcal{E}, \mathcal{E}'$. Say, for example, $\mathcal{E}$="a die roll is even" and $\mathcal{E}'$="a die roll is odd". These events are disjoint. Knowing, that $\mathcal{E}$ happens we at the same time know for sure that $\mathcal{E}'$ cannot happen. Therefore, "disjointness" is a very strong form of dependence; i.e., in this sense disjointness is the opposite of independence. Formally, for two disjoint events $\Pr[\mathcal{E} \cap \mathcal{E}'] = \Pr[\emptyset] = 0 \neq \Pr[\mathcal{E}] \Pr[\mathcal{E}']$ (unless one of the events has zero probability).

Statistically disjoint events

Here are more remarks about disjoint events. For $\mathcal{E}, \mathcal{E}'$ disjoint events we have $\Pr[\mathcal{E} \cup \mathcal{E}'] = \Pr[\mathcal{E}] + \Pr[\mathcal{E}']$.

Here is an example showing why this is true. Let $A = \{1,2,3\}$ and $B = \{4,5,6\}$. Then $\Pr[A \cup B] = \Pr[\{1,2,3,4,5,6\}] = \underbrace{\Pr[1] + \Pr[2] + \Pr[3]}_{\text{equals } \Pr[A]} + \underbrace{\Pr[4] + \Pr[5] + \Pr[6]}_{\text{equals } \Pr[B]} = \Pr[A] + \Pr[B]$. The

general case (for general disjoint $\mathcal{E}$ and $\mathcal{E}'$) you can similarly verify that for disjoint $\mathcal{E}$ and $\mathcal{E}'$ we have $\Pr[\mathcal{E} \cup \mathcal{E}'] = \Pr[\mathcal{E}] + \Pr[\mathcal{E}']$.

We stress 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 sample space.

- In contrast, for independent $\mathcal{E}, \mathcal{E}'$ we have $\Pr[\mathcal{E} \cap \mathcal{E}'] = \Pr[\mathcal{E}] \Pr[\mathcal{E}']$, which is a definition (not some provable consequence).

Remark 6. *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?

1.5 Conditional spaces

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

The idea of events that affect or not the possibility of realization of other events brings us to *conditional sample 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

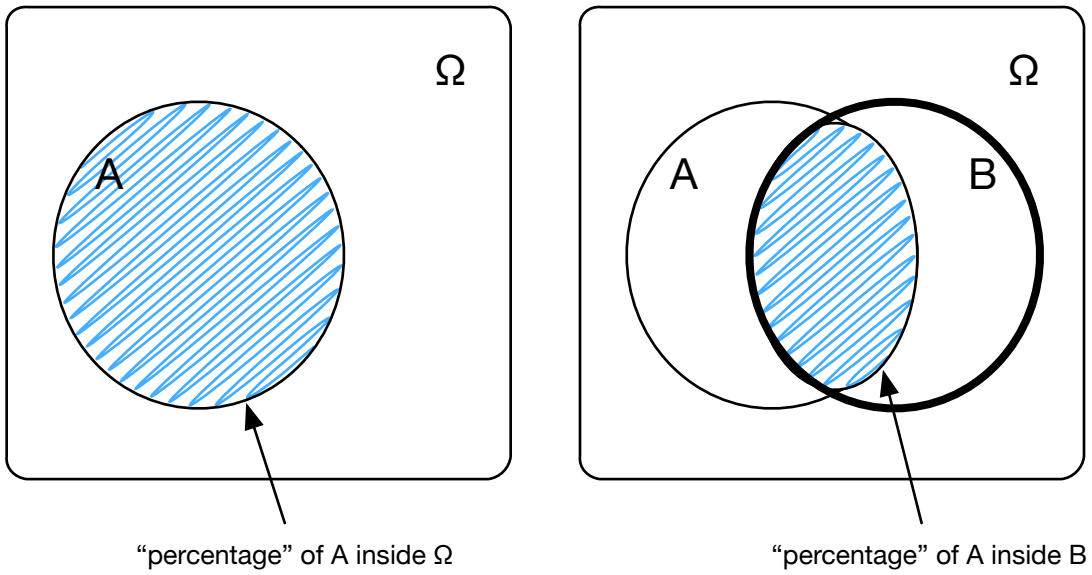

Figure 1.1: The *percentage* (fraction) of A inside Ω (left figure) is equal to the *percentage* of A inside B (right figure). That is, we can now think of B as a new sample space and this corresponds to the real-world intuition that we are measuring the probability of A happening after we know that 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.

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 in a sense event B now itself becomes a sample space. Sample 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).

[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.

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

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

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

[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.

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).

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 $A_1 \cup A_2 \cup \cdots \cup A_k = \Omega$.

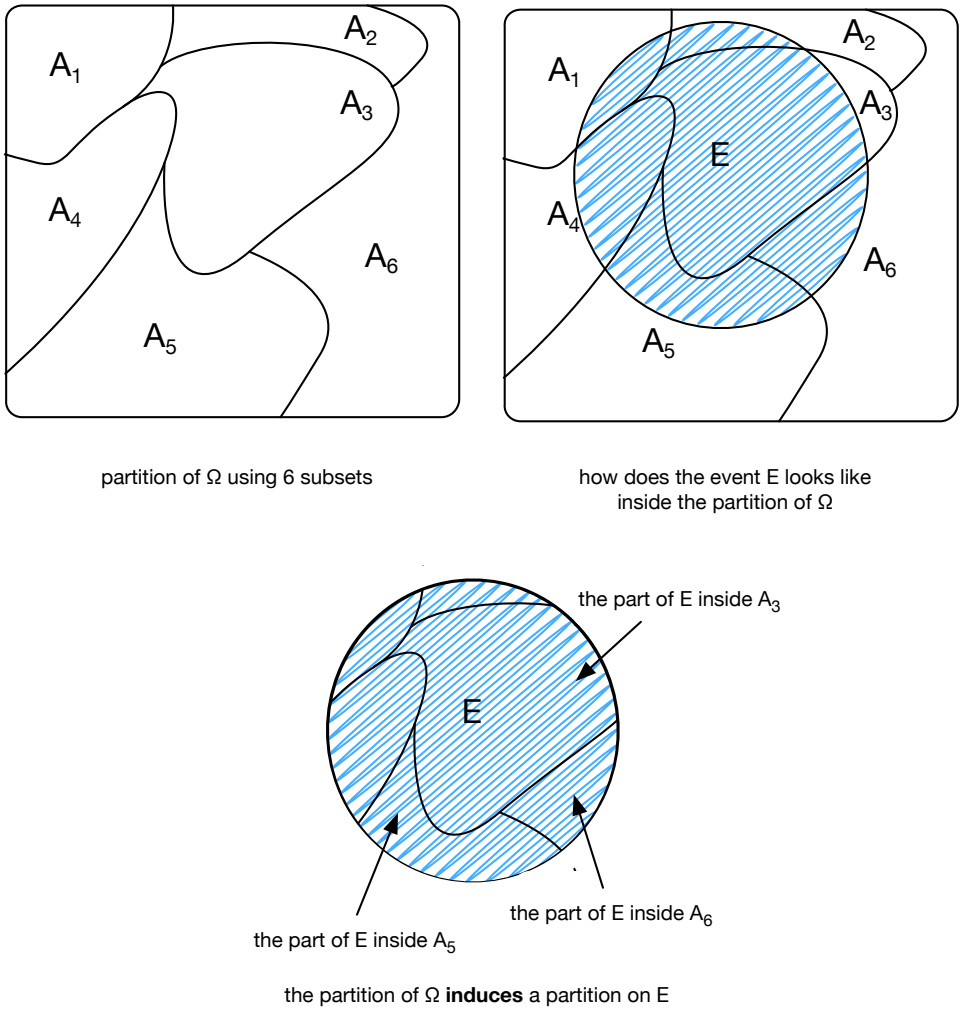

Figure 1.2: A partition $A_1, A_2, \ldots, A_6$ of Ω *induces* the partition $\mathcal{E} \cap A_1, \mathcal{E} \cap A_2, \ldots, \mathcal{E} \cap A_6$ of $\mathcal{E}$.

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, $\Pr[\mathcal{E} \cup \mathcal{E}'] = \Pr[\mathcal{E}] + \Pr[\mathcal{E}']$. Further-

more, 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 sample 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 the partition of Ω induces a partition on $\mathcal{E}$; i.e., $\mathcal{E}$ can be thought of as gluing together the pieces of $\mathcal{E}$ as per the partition of Ω. Formally, $\mathcal{E} = (\mathcal{E} \cap A_1) \bigcup \cdots \bigcup (\mathcal{E} \cap A_k)$ and any two $(\mathcal{E} \cap A_i) \cap (\mathcal{E} \cap A_j) = \emptyset$ (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 $\Pr[\mathcal{E}] = \Pr[(\mathcal{E} \cap A_1) \bigcup \cdots \bigcup (\mathcal{E} \cap A_k)] = \Pr[\mathcal{E} \cap A_1] + \ldots + \Pr[\mathcal{E} \cap A_k]$. Now, apply the condition probability definition: $\Pr[\mathcal{E}] = \Pr[\mathcal{E}|A_1]\Pr[A_1] + \ldots + \Pr[\mathcal{E}|A_k]\Pr[A_k]$. □

Theorem 7 is used when we can easily compute $\mathcal{E}$ conditioned on the fact that say A_1 and A_2 have 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 em-

phasize 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 for historical reasons.

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

Not all random variables have such a trivial connection to sample spaces. We typically care about one experiment, i.e., one sample 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 "$\Pr[\dots]$" 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 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 $\Pr[X = 5]$.*

Remark on terminology 9. *Let us consider the example of the fair die roll and the random variable X defined above. Then, the expression "$X \geq 4$", which is the same as "$X(\omega) = 4$", is satisfied by $\omega =$ face 4, $\omega =$ face 5, and $\omega =$ face 6. That is, $\Pr[X \geq 4] = \Pr[\{\text{face } 4, \text{face } 5, \text{face } 6\}] = \frac{1}{2}$. As a side remark, useful when proving Chebychev's inequality later on, is that "$|X(\omega)| \geq 4$" is satisfied by the same elementary outcomes of Ω as the expression "$X(\omega)^2 \geq 16$".*

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

Remark 10. *Our first example is an anti-example. 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 that there is a reason for this.*

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

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

We stress 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.

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 10 happens. In the case of the fair die, this number is very large, i.e. $\mathrm{Var}[X] = \Pr[(X - 3.5)^2 = 6.25]6.25 + \Pr[(X - 3.5)^2 = 2.25]2.25 + \Pr[(X - 3.5)^2 = 0.25]0.25$. This is because the possible values of the function/random variable $Y = (X - E[X])$ are $\{6.25, 2.25, 0.25\}$. But, $\Pr[(X - 3.5)^2 = 6.25] = \Pr[|X - 3.5| = 2.5] = \Pr[X = 1 \text{ OR } X = 6] = \frac{2}{6} = \frac{1}{3}$. Similarly, we get $\Pr[(X - 3.5)^2 = 2.25] = \frac{1}{3}$ and $\Pr[(X - 3.5)^2 = 0.25] = \frac{1}{3}$. Thus, $\mathrm{Var}[X] \approx 2.91$, which is "very large" compared to its possible assumed values in the interval $[1, 6]$.

Variance is an 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 10.

At the end of this section, we discuss an example that is indicative of the role of variance. Some more, and very interesting examples,

explaining the role of variance in sampling and estimation will be developed in the sequel in Sections 1.11 and 1.12. Before we discuss these examples we would like to discuss some important properties of the expectation and variance.

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

Lemma 11. *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 the 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[11] that the statement is *not* true for variance, we should prove that the following *is* true:

$$\text{NOT}\Big(\text{for all } X, Y \text{ we have } \text{Var}[X+Y] = \text{Var}[X] + \text{Var}[Y]\Big)$$
$$= \text{there is } X, Y \text{ such that we have } \text{Var}[X+Y] \neq \text{Var}[X] + \text{Var}[Y]$$

An example (formally) proves existence. You should make sure that you can give an example showing that Lemma 11 does not hold for variance.

An indicative example of the role of expectation and variance

Consider two betting games where in each we press a button and we either get a reward or we pay.

Betting game I: With probability $1/10$ we uniformly sample an integer from $\{1, 2 \ldots, 10\}$. If the number is from $1, 2, 3, 4$, then we pay \$2. If the number is from $5, 6, 7, 8, 9, 10$ then we gain \$4.

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

Betting game II: In this game, we press the button independently (i.e. product space) twice. In each time with probability $1/10$ we get a number from $\{1, \ldots, 10\}$ and if it is at most 4 then we pay \$1, whereas if it is 5 or bigger we gain \$2.

We have three questions to answer.

- Would you play in any of these betting games?
- Does it matter which game you play?

To answer the first question we determine the expectation. If the expectation is positive then this means that on average we make a profit and it makes sense to play. The expectation of which variable?

For the betting game I, the variable $X = -2$ with probability $4/10$ and, $X = 4$ with probability $6/10$. That is, $E[X] = -2 \cdot \frac{4}{10} + 4 \cdot \frac{6}{10} = \frac{16}{10} = 1.6$.

For betting game II, $Y_1 = -1$ with probability $4/10$ and $Y_1 = 2$ with probability $6/10$, this is the gain/loss of the first trial. Similarly, for the second trial $Y_2 = -1$ with probability $4/10$ and $Y_2 = 2$ with probability $6/10$. The total gain/loss is $Y = Y_1 + Y_2$. The expected value of Y is $E[Y] = E[Y_1 + Y_2] = E[Y_1] + E[Y_2] = (-1 \cdot \frac{4}{10} + 2 \cdot \frac{6}{10}) + (-1 \cdot \frac{4}{10} + 2 \cdot \frac{6}{10}) = 0.8 + 0.8 = 1.6$

Therefore, in both betting games in expectation, we have a profit of 1.6 dollars. Thus, it does make sense to play in any of the betting games. In practice, to "see" this gain we must play many times and if we average over the number of times we played then on the average we will see that each time we played we gained the amount of \$1.6.

Since the expected values are the same, does this mean that both games are equally good in terms of profit?

Here is where the role of variance is important.

If we do the calculation we find that $Var[X] = 8.64$, whereas $Var[Y] = Var[Y_1] + Var[Y_2] = 2 \times 2.8 = 5.6$. That is, in both cases,

we have in expectation the same profit, but in Betting Game II the variance is much smaller. In other words and informally speaking, if we play the game once we are better off to play in Betting Game II because this profit will be realized since statistically, we are much closer to the expected profit.

Independent random variables

Suppose that X, Y are random variables defined over the same sample space Ω. We will say that X, Y *are independent* if the following corresponding events are independent[12] that is

$$\text{for all } x \in X(\Omega), y \in X(\Omega), \qquad \Pr[X = x \text{ AND } Y = y] = \Pr[X = x] \Pr[Y = y]$$

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

Example Consider the sample space of independently rolling a fair die twice $\Omega = \{(1,1), (1,2), \ldots, (6,6)\}$ where the probability of each elementary outcome (i, j) is $\Pr[(i, j)] = 1/36$. Now, let X be the random variable which is 1 if the first die roll is even and 0 otherwise. Also, let Y be the random variable which is 1 if the second die roll is equal to 3 and 0 otherwise. Then, the event "$X = 1$" corresponds to the set $\{(2,1), (2,2), \ldots, (2,6), (4,1), (4,2), \ldots (4,6), (6,1), \ldots, (6,6)\}$. We can similarly, calculate the sets corresponding to the events $X = 0$, $Y = 1$, and $Y = 0$. As we saw before, and after doing detailed calculations, we can show that $\Pr[X = i, Y = j] = \Pr[X = i] Pr[Y = j]$ for all *i-j* combinations where $i = 0, 1$ and $j = 0, 1$.

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

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

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)*[13] 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 that they will indicate success (=1) and failure (=0) of various events of interest.

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 $\text{Var}[X_1] = E\left[(X_1 - E[X_1])^2\right] = E[X_1^2] - E[X_1]^2$. 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[14], unless the X_i's are pairwise (i.e. every two

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

[14] 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 $\text{Var}[X_1 + X_2] = \text{Var}[2X_1] = E[4X_1^2] - E[2X_1]^2 = 4\text{Var}[X_1] \neq 2\text{Var}[X_1]$.

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 at 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 by leaving two tasks to the reader.

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

Second, try to understand if there exist variables, which are pairwise independent but they are not three-wise independent[15].

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

Probability theory was properly formalized (axiomatized) by Kolmogorov[16] in the 1930s. Before the 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, oftentimes 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 $\mathcal{E} = \{\omega \mid X(\omega) \text{ is even}\} = \{2, 4, 6\}$. That is, the event is defined by the predicate "$X(\omega)$ is even". At the end, we calculate $\Pr[\mathcal{E}] = \frac{1}{6} + \frac{1}{6} + \frac{1}{6} = \frac{1}{2}$.

[15] 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, are restrictions of the notion of independence of all variables (which coincides with n-wise independence).

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

Now, instead of all these, we could have simply written, $\Pr[X \text{ is even}] = \frac{1}{2}$, which *means exactly the same thing* and can replace the whole 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 lose 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 sample space associated with these random variables. When obvious we will not explicitly mention the space (but it is always there).

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[17] and we cannot compute expectations or variances for such an X. However, we will occasionally abuse 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.[18].

[17] 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.

[18] **Reminder:** This material is *copyrighted* 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.

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?

Are the digits of the sum statistically correlated with each other? If b_0 and b_1 are independent[19] 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 the 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

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

first non-trivial fact about statistical intuition. It deepens our understanding of how a random sum statistically looks like[20].

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 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 very interesting for someone who wants to understand what statistical independence means.

Does b_0 statistically depend 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

[20] 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!)

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.

Statistical dependence does not mean that things are "dependent" in some natural/everyday commonsense notion. Commonsense suggests that the digits of a summation depend on the variables we are summing. *Statistically*, dependence means that "knowing the value" of one variable affects what we "predict" about a digit of the summation (of more variables). It turns out that we learn nothing about the LSB when we know the value of one of the summands. The same is not true about more significant digits. For example, if we are summing three independent binary variables, the probability that the MSB (second digit) is 1 with different probability if we know that one of them is 0 or whether we know that the same variable is 1.

1.9 Common distributions and useful tools

The most basic distribution is the *Bernoulli trial*, which 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 $c > 0$ an 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[21].

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

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

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 it is for a single execution of the experiment to yield a value for the variable that is k times away from 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.

Markov is definitely not useless. It is helpful in certain cases. Moreover, it is very important in deriving new, stronger inequalities, but in more restricted settings [22].

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.

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

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 recalling that $\Pr[(X - E[X])^2 \geq c^2] = \Pr[|X - E[X]| \geq c]$ for $c > 0$).

To prove Markov's is also not hard (this proof can be skipped at 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, $cE[f(X)] \leq E[X] \implies \Pr[X \geq c] \leq \frac{E[X]}{c}$. □

Question: Modify Betting Game I and Betting Game II (cf. p. 21) in the following way: consider the payments to happen with probability $1/10$ and the profits with $9/10$ and maintain the same loss and profit values as in the original games. Recalculate the expectations and variances and then use Chebychev's inequality to derive which of the two betting games is more likely (and how much more likely) with a single instantiation of the game to have profit (i.e. the corresponding random variable to be bigger than 0).

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*[23]*, 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 on expectation.

For the calculation with Chebyshev we will need two facts. First, the variance of each X_i is $\mathrm{Var}[X_i] = E[X_i^2] - E[X_i]^2$, and since $X_i \in \{0, 1\}$, we have $\mathrm{Var}[X_i] = E[X_i] - E[X_i]^2 = \frac{1}{2} - \left(\frac{1}{2}\right)^2 = \frac{1}{2} - \frac{1}{4} = \frac{1}{4}$. Finally, since the X_i's are independent we have that $\mathrm{Var}[X] = \mathrm{Var}[X_1] + \cdots + \mathrm{Var}[X_n] = \frac{n}{4}$.

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* $X = \sum_{i=1}^{n} X_i$*,*

[23] We already saw that every intuitively described experiment corresponds to a formal sample space

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 $\Pr\left[\left|X - E[X]\right| > \varepsilon E[X]\right] \leq \frac{\mathrm{Var}[X]}{(\varepsilon E[X])^2} = \frac{n\mathrm{Var}[X_1]}{(\varepsilon n E[X_1])^2} = \frac{\mathrm{Var}[X_1]}{n\varepsilon^2 E[X_1]^2} = \frac{1}{n} \cdot \frac{p-p^2}{\varepsilon^2 p^2} = \frac{1-p}{\varepsilon^2} \cdot \frac{1}{n} = O\left(\frac{1}{n}\right)$, since ε and p are constants. □

Thus, just by 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, 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 $\mathrm{Var}[X] = \mathrm{Var}[X_1] + \cdots + \mathrm{Var}[X_n]$. 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[24], do exhibit measure concentration phenomena. We will see in the next section

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

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) that 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 1.3).

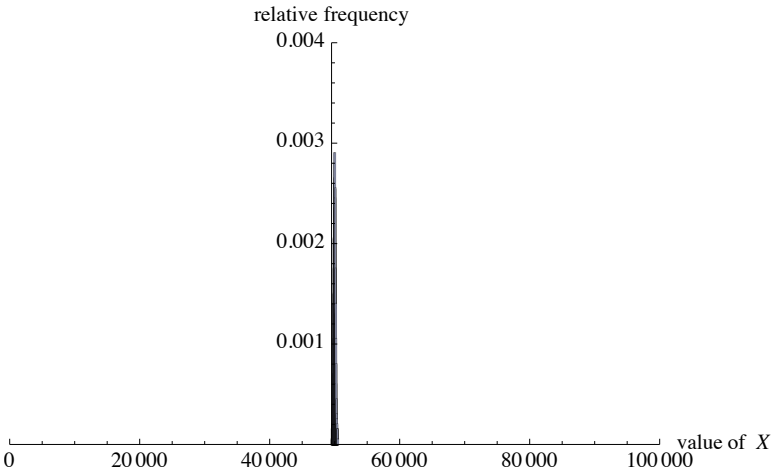

Figure 1.3: Histogram for the value of $X = X_1 + \cdots + X_{10^5}$

We can see that 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 1.4).

We observe that in this computer experiment sharp concentration did happen around the expectation. In the proof of Theorem 15 we

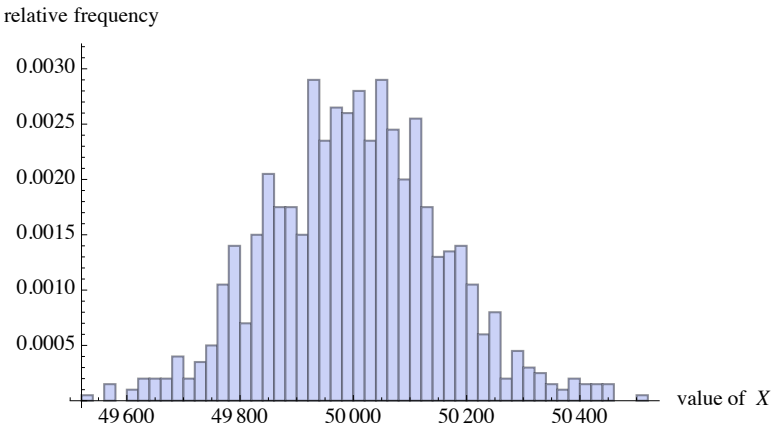

Figure 1.4: Same histogram as in Figure 1.3 but now magnified around $X = 50000$. Observe that no single one among these 1000 repetitions of the experiment resulted anything smaller than 49500 or bigger than 50500.

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 the probability 0.00005 of being outside the concentration interval appears to be very small. But, our theorem calculates that this happens with probability *at most* 0.00005. Can it be that our theorem is not very strong? Maybe a better analysis could have resulted in a better (better=smaller) upper bound. For instance, 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. the questions we ask here do not challenge whether we have proved Theorem 15 correctly. We challenge whether this bound can be improved.

Let us make the following thought experiment. Let us suppose that 0.00005 is the true upper bound; i.e. the probability is exactly equal to 0.00005. Then, the probability that $X \in [49500, 50500]$ is

$\left(1-\frac{1}{20000}\right)$ and the probability that *all* of 100000 independent executions are all inside $[49500, 50500]$ would have been $\left(1-\frac{1}{20000}\right)^{100000} = \left(\left(1-\frac{1}{20000}\right)^{20000}\right)^5 \approx \frac{1}{e^5} \approx 0.0067$. That is, the probability that all 100000 independent executions of the experiment are inside $[49500, 50500]$ is about 0.67%. Recall that in our computer simulation/experiment it happened to be the case that all of the executions were inside $[49500, 50500]$. Now, one of two things has happened. Either we were unlucky and we just hit the event that happens with 0.67% or it is just the case that the 0.00005 is not a "tight" upper bound ("not tight" = "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 "$\Pr[\ldots]$" 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

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

, 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 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[25].

Theorem 16. *Let $X_1, \ldots, X_n$ 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 informal (but reasonably) defined experiment immediately translates to a sample space Ω. What happens if the experiment changes over time?

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*, $t \in \{0, 1, 2, 3, \ldots\}$. We occasionally *introduce time* in the analysis of an experiment. In these

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

cases, there is no physical notion of time associated with our introduced time steps. For example, when we consider n independent $X_1, \ldots, X_n$ 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 sample spaces, e.g. $\Omega_1, \Omega_2, \ldots$. 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 $X_1, X_2, \ldots$, 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 $\alpha, \beta_1, \ldots, \beta_{i-1} \in X(\Omega)$, $\Pr[X_i = \alpha | X_1 = \beta_1, \ldots, X_{i-1} = \beta_{i-1}] = \Pr[X_i = \alpha | X_{i-1} = \beta_{i-1}]$. 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 piece of 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[26]. Markov processes "are processes without memory". Martingales are processes that "maintain the expected value".

A 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 *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*

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

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 to 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 sample space, what is a Doob's filter, and other related topics.

1.15 Suggested readings

Here is what I consider 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