Elementary probability theory

What is probability? And why does it work?

Everybody has some intuitive feeling about randomness. Yet to many people it is not far from magic. And it is easy to get tricked or make mistakes (see appendix at the end of the chapter).

To say that probability works means that if the probability of some event is, say, $1/3$, and we reproduce many times the experiment producing the event, then it will happen roughly one third of the times. But as we know, this doesn't always work nicely. Sometimes the event will happen more often, or less often, than it should. These are exceptions. And by definition of exceptions they are rare.

Any way we want to pinpoint a good definition of probability, it seems to escape like quicksilver. Our explanations end up involving... probability. But let's give it a try.

In nature there are experiments which, when we replicate them, keep producing exactly the same results. They are predictable or deterministic. They are those we mentioned earlier, appearing for instance in classical mechanics when we can apply Newton's laws.

And there are experiments whose outcomes vary. Those are said to be random. The randomness comes either from some fundamental randomness of nature, like certain phenomena in quantum mechanics, or from our incomplete knowledge, as said, of the initial conditions or other things. Yet the results display some experimental stability in the proportions of occurrences, when the experiment is repeated many times, which we will come back to.

This fundamental distinction established, to work with probability theory we need some primitive ideas about randomness, and then we need to construct a framework.

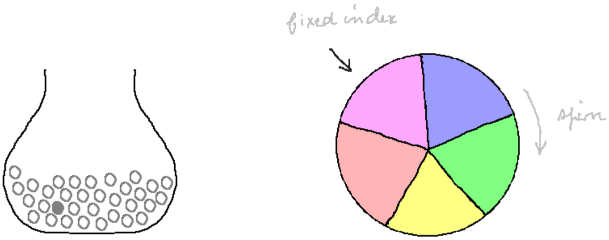

Figure 1: Rare events, and equiprobable events.

There are two primitive ideas, which we present with two illustrative examples, figure 1:

1. *Rare events*: In an urn, suppose there is a *very large* number of white marbles and one black marble. If we plunge the hand in the urn without looking, and pick a marble at random, then picking the black marble would be rare, and we can *"safely"* count on picking a white marble[1].

2. *Equiprobable events*: If the possible outcomes of an experiment display some *symmetry*, then each event is considered as likely as any other. Thus when we spin the wheel shown on the right side of figure 1, and wait until it has stopped and the fixed index points to a color, the five colors are said to be equiprobable. Equiprobable distributions will play an important role, at first, when we begin to talk about the entropy of systems.

We mentioned the stability of the proportions of occurrences, in a random experiment reproduced many times. We intuitively – and correctly – feel that it is linked to the *second idea* listed above. But we shall see in a moment that it is also actually an instance of the *first idea* about rare events.

[1] One of the many counterintuitive aspects of probability, which we will have to get used to, is that *any marble*, if we can distinguish each of them, is rare. Yet we do pick one. Seen that way, whatever outcome we obtain is rare. The paradox clears when we distinguish states and events, which are explained in the sequel of the text.

Let's now turn to the framework to work with a *random experiment* and probabilities. The experiment itself is usually denoted $\mathcal{E}$. Each time we reproduce $\mathcal{E}$, at the end of the experiment the world is in some state ω, which may be different from replication to replication.

We are interested in some aspects of ω: we may want to measure some quantity which has a numerical value, or note a color, or whatever. Sometimes we are interested in only one feature of ω and sometimes in several, like picking someone at random, in a well-specified procedure, and recording both his height and weight, or level of income and academic degree, or, of a more physical flavor, measuring the position and velocity of a particle.

The main source of mistakes when working with probabilities is the ill-definition of the *experiment* $\mathcal{E}$. For instance, when we simply say "let's pick a person at random", it is ill-defined. The experiment is not sufficiently specified. Do we mean, in the lecture hall, or out of a subway entrance in New York, and in that case in which borough, at what time of day, or in the United States, or on Earth?

Another example – of impossibility this time – is when we say "let's pick a point at random uniformly over the entire line". It is actually impossible. There is no such thing as a density of probability with the same value from $-\infty$ to $+\infty$[2]. So we have to be careful.

Once the experiment $\mathcal{E}$ has been specified, we consider that each time it is performed, the world comes out in some state ω. The set of the ω's is denoted Ω and is called the space of states, or set of states. Probabilists call it the universe of possible states attached to performing $\mathcal{E}$.

We are interested in measuring some features, numerical data or non numerical data, about ω. Suppose we are interested in a

[2] But a sequence of densities can become more and more flat and therefore with almost the same value – necessarily close to zero – everywhere. Similarly, there is no such thing as a function whose value is zero everywhere except at one point and whose integral is one. But *distributions*, which, in one theoretical approach, are defined as limits of functions, can be like that.

feature denoted X. It depends on ω. In other words it is a function of ω. It is called a *random variable* and the measurement made on a given ω is denoted

$$X(\omega) \tag{1}$$

It is the result, after having performed the experiment $\mathcal{E}$ once, of the measurement of X on the state ω that $\mathcal{E}$ produced.

$X(\omega)$ takes its values in the set of possible outcomes of X. Let's denote it A. Without going into category theory, let's mention that mathematicians write that X is a function from Ω to A as follows

$$X : \Omega \to A \tag{2}$$

The set A can be finite, infinite and countable, or infinite and continuous.

For instance if the experiment $\mathcal{E}$ is the throw of a die, X is the number shown on top after a throw, and the set A is

$$A = \{1,\ 2,\ 3,\ 4,\ 5,\ 6\} \tag{3}$$

In this case the abstract set Ω can just be taken to be the set of results of throwing the die, i.e. $\Omega = A$ itself. Then the ω's are simply the six possible results, and we don't even have to think about X. It is simply the identity.

Although this simplification is often appropriate, it is better – particularly when there are several random variables that we want to measure on the outcome of $\mathcal{E}$ – to clearly distinguish the set of states Ω that is once and for all attached to $\mathcal{E}$ and the various sets of outcomes of random variables.

Let's stress again that A doesn't have to be numerical. Our die, for instance, could have faces painted with different colors rather than bearing numbers.

Once $\mathcal{E}$, Ω, X and A have been defined, the last fundamental concept to introduce in the framework is a *probability* P.

In advanced probabilities[3], mathematicians technically talk about a measure of probability P on Ω that is σ-additive, etc. And they introduce it before any random variable. But in this review of elementary probability we don't need to go into this. Furthermore in an elementary approach, it may obfuscate the link with probabilities as we intuitively know them.

P is defined such that any subset of Ω – called an *event* – has a probability. In the case of the die, it is particularly simple. Each ω is itself an interesting event. There are six of them. And if the die is well balanced they are equiprobable. In other words we assign to them the same probability. Thus we write

$$P\{X = 5\} = \frac{1}{6} \tag{4}$$

meaning that the probability of getting a 5, when throwing the die, is $1/6$. Events are any subsets of Ω, not only the ω's themselves. We can also write

$$P\{X \leq 2\} = \frac{1}{3} \tag{5}$$

The principle that *symmetry*, or equivalence somehow, between the possible states ω implies equiprobability – in other words, whatever makes them different doesn't affect their propensity to

[3] In the theory of probability, there are two main parts:

1. Elementary probability, developed by Fermat (c. 1605-1665) and Pascal (1623-1662) in the XVIIth century to clarify questions about randomness that had puzzled people for centuries, in particular card players but not only. The elementary theory was further developed by people like Chebyshev (1821-1894), Markov (1856-1922), etc. That is the one we briefly sketch here.

2. The mathematical theory of probability, developed by Lévy (1886-1871), Khinchin (1894-1959), Kolmogorov (1903-1987) and others in the XXth century to give it firmer foundations than elementary probability. It uses measure theory, Lebesgue integral, σ-algebra, etc. This theory is not necessary for our purpose and won't concern us in the course.

occur – is often invoked to figure out the distribution of probability P attached to an experiment.

Another possible way to figure out P is through a large number of replications of the experiment. We will explain it in a moment.

Finally there may be various probabilistic calculations which we can also make to relate the distribution of a random variable X to those of other random variables we already know. This pertains to the calculus of probability. And we shall work out many examples in this course.

Our framework is now complete. It consists of a random experiment $\mathcal{E}$, a big set Ω of possible states of the world after the performance of $\mathcal{E}$, and a probability P.

$$\text{framework} = [\ \mathcal{E},\ \Omega,\ P\ ] \tag{6}$$

And we are interested in measuring various random variables X, Y, Z, etc. after having performed $\mathcal{E}$[4].

Let's now simplify a bit the setting and the notations. For the time being, the set of possible states will be

$$\Omega = \{\ \omega_1,\ \omega_2,\ \omega_2,\ \dots\ ,\ \omega_n\ \} \tag{7}$$

that is a finite set of outcomes of $\mathcal{E}$. The states are indexed by i running from 1 to n. For example, when flipping a coin once,

[4] In maths manuals, the reader will usually see the framework described as $[\ \Omega,\ \mathcal{A},\ P\ ]$, the experiment $\mathcal{E}$ not being mentioned – which in our teaching experience is regrettable. And the extra $\mathcal{A}$, not to be confused with the target set of any random variable, is the collection of subsets of Ω, but not quite all of them, only the "measurable" ones. Again, we don't need to be concerned here with those subtleties.

Interested readers are referred to the book by Alexandrov, Kolmogorov, Lavrentiev, *Mathematics*, Dover, 1999, which gives an excellent presentation of most of the mathematics necessary to do physics, and more. Chapter XV in the book explains what is measurability, and chapter XI, written by Andreï Kolmogorov presents elementary probabilities.

A standard reference to study probabilities is the two volume book by William Feller (1906-1970), *An Introduction to Probability Theory and Its Applications*, John Wiley & sons there are two states: ω_1 = heads, and ω_2 = tails.

Later on, we will extend this to an infinite countable, and then even an infinite continuous set Ω^5.

Staying with a finite set Ω, the probabilities of the single states ω_i's will simply be denoted

$$P(i) \tag{8}$$

Sometimes they will even be simply denoted P_i (see lecture 2).

They must satisfy

$$\begin{gathered} P(i) \geq 0 \\ \sum_i^n P(i) = 1 \end{gathered} \tag{9}$$

Indeed, probabilities are positive numbers. And the total probability, when we add up the probabilities of all possibilities, should be equal to one. When performing $\mathcal{E}$ we certainly should get some result.

Probabilities have all sorts of interesting, beautiful and sometimes surprising properties. The most useful one for us in this course is the *law of large numbers*.

Here is what it says. Suppose that we either make many replicas of the same system, or do the same experiment $\mathcal{E}$ over and over a large number N of times, and we count how many times we get the i-th outcome ω_i. That is some count that we denote N_i. Then the law says that

$$\lim_{N \to \infty} \frac{N_i}{N} = P(i) \tag{10}$$

[5] In this last case probabilities will be replaced by *densities of probability*. Instead of considering $P\{X = x\}$, which would usually be equal to 0, we will consider $P\{\ X \in [\ x,\ x + dx\ ]\ \} = p(x)dx$. And, following the custom in physics, we will often still denote it $P(x)dx$, keeping in mind that each random variable has its own density. In other words, when it is necessary to be explicit, we will write $P_X(x)dx$.

This is a statement about probabilities, which can be stated more precisely and rigorously within the framework $[\,\mathcal{E},\ \Omega,\ P\,]$. But let's approach it at an intuitive level. It says that when we replicate $\mathcal{E}$ a larger and larger number of times, and measure the *experimental frequency* of occurrence of the i-th outcome, this experimental frequency (which is itself a random variable) deviates less and less i.e. gets closer and closer to the actual probability $P(i)$.

For instance, if we toss a coin a thousand times, the frequency of heads will be close to $1/2$. If we throw it 10 000 times, the frequency will be even closer to $1/2$. In each case, it is only a probabilistic statement. There can be – and in fact most of the times there will be – a discrepancy. That discrepancy is itself a random variable. But it will have a distribution more and more concentrated, relatively to its range, around 0.

We said earlier that the convergence of experimental frequencies toward their corresponding theoretical probabilities is actually an instance of the rare event idea, see figure 1 and its comments.

The law of large numbers is neither magic, nor some kind of eerie principle of nature. It stems from the fact that in an urn with one black ball and very many white balls, if we pick one at random, we can assume safely that we will pick a white one. It is not always the case, but it will be extremely rare to pick a black ball. And for all practical purpose it can be neglected.

Let's see why, in the case of tossing coins, the law is a *simple result in numbering*. Consider the experiment $\mathcal{F}$ which consists in tossing the coin 1000 times, i.e. repeating $\mathcal{E}$ 1000 times. The space $\Omega_{\mathcal{F}}$ attached to $\mathcal{F}$ has 2^{1000} elements – that is a huge number. Each are equiprobable. When we perform $\mathcal{F}$ once, i.e. when we repeat $\mathcal{E}$ a thousand times, we pick one element in $\Omega_{\mathcal{F}}$.

It turns out – and it is not hard to show, although we won't do it – that most elements in $\Omega_{\mathcal{F}}$ contain about as many heads as tails. View them as the white balls in the urn if you like. And the black balls would be very few. So when we pick one, we pick a white one.

To try to shed even more light on the phenomenon, rather than do some combinatorics, consider the 16 possible results, displayed below, of throwing the coin four times. The reader can check that there is only one result with zero head. Four results with 1 head. Six results with 2 heads. Four results with 3 heads, and one result with 4 heads.

$$\begin{array}{r}
T,\ T,\ T,\ T \\
T,\ T,\ T,\ H \\
T,\ T,\ H,\ T \\
T,\ T,\ H,\ H \\
T,\ H,\ T,\ T \\
T,\ H,\ T,\ H \\
T,\ H,\ H,\ T \\
T,\ H,\ H,\ H \\
H,\ T,\ T,\ T \\
H,\ T,\ T,\ H \\
H,\ T,\ H,\ T \\
H,\ T,\ H,\ H \\
H,\ H,\ T,\ T \\
H,\ H,\ T,\ H \\
H,\ H,\ H,\ T \\
H,\ H,\ H,\ H
\end{array}$$

So there is a kind of concentration around an equal number of heads and tails. The counts of the number of heads actually correspond to the so-called *Pascal triangle*. They are also the coefficients in the development of the polynomial $(a+b)^N$. The concentration about half and half is more marked, of course, when N is larger than 4, and it grows more and more marked as N increases.

That is what the law of large numbers is about. In probability theory, it is stated more rigorously than we have done here. It is proved via an intermediate result called *Bienaymé-Chebyshev inequality*[6]. It is not particularly hard, and is rather elegant. But

[6] Named after Irénée-Jules Bienaymé (1796-1878), French mathematician, and Pafnuty Chebyshev (1821-1894), Russian mathematician.

it is outside what we want to do in this review[7].

The law of large numbers, expressed by equation (10), says that the ratio N_i/N converges "probabilistically" toward $P(i)$ when N gets very large. In other words,

when we have repeated an experiment a large number of times, we can use the experimental frequency N_i/N of occurrence of the i-th outcome as an estimate of $P(i)$.

We use this result all the time.

Now let's go back to a random variable we want to measure, which is not the outcome of $\mathcal{E}$ itself. To get closer to physics concepts and notations, let's call the random variable F. So let's suppose that there is a quantity, denoted $F(i)$, that is associated with the i-th state ω_i. Recall expression (7) defining the set of states.

F can be some meaningful physical quantity. We can also make it up. For example if our system is heads and tails, and nothing but heads and tails, we could assign

$$\begin{aligned} F(\text{heads}) &= +1 \\ F(\text{tails}) &= -1 \end{aligned} \tag{11}$$

If our system has many more states, we may want to assign a larger number of possible values taken by F – not necessarily the same number as the number of elements in Ω though. F is simply some function of the states. We already mentioned this in expression (2), let's write it again

$$F : \Omega \rightarrow A \tag{12}$$

The random variable F acts on the set of states Ω and takes its value in the set A. In the case of the coin, $\Omega = \{H,\ T\}$, and the set A in which F takes its values is $\{+1,\ -1\}$. Thus, we

[7] Another beautiful and useful result is the *Central Limit Theorem*, which shows in essence that the Pascal triangle looks more and more, when properly rescaled, like a bell-shaped curve called a Gaussian. And it is true in a much more general setting than just flipping a coin many times. That is in fact the essential reason why Gaussian distributions pop up all the time in physics and engineering.

have made up a numerical random variable – or *measurement* – attached to flipping a coin.

As said, $F(i)$ can also be some meaningful physical quantity. It could be the energy of the i-th state ω_i. Given the state in which is some system, it has an energy. Its measurement would perhaps be called in that case

$$E(i) \tag{13}$$

Or we could be interested in the momentum of the i-th state. We would have to choose a good notation not interfering with probabilities. Or it could be something else. It could be whatever we happen to like to measure on our system.

Then an important quantity is the *average* of $F(i)$. After this probability refresher, we will mostly use the quantum mechanical notation for it, even though we are not doing quantum mechanics. It is a nice notation. Physicists tend to use it all over the place. Mathematicians hate it. We just put a pair of brackets around F to mean its average. It is defined as follows

$$< F > = \sum_{i}^{n} F(i)P(i) \tag{14}$$

In words, it is the average of the values $F(i)$ weighted by their respective probabilities $P(i)$.

Notice that the average of $F(i)$ does not have to be any of the possible values that F can take. For example, in the case of the coin, where $F(H) = +1$, and $F(T) = -1$, and we flip it a million times, and the probability is $^1/_2$ for heads and $^1/_2$ for tails, the average of F will be 0. It is not one of the possible outcomes of F in the experiment. Yet it is its average. There is no rule why the average of a measure should be one of its possible experimental values.

Thanks to the law of large numbers, we can write equation (14) another way.

$$< F > \approx \sum_i^n F(i) \frac{N_i}{N} \tag{15}$$

This approximate equality becomes a true equality in the limit when there is a large number of measurement.

That is it for our mathematical preliminary. We need to know what a *random experiment* is, what a *state* ω is, what a *probability* is, what a *random variable* or random measurement is, and what is an *average*, because we will use them over and over.

Before leaving our probability refresher, let's mention that a more complete presentation of elementary probabilities, would include the following topics, for which we refer the reader to any good manual:

1. Variance of a r.v. X. It is the expected value of the squared deviation of X from its average, and is often denoted σ^2

   $$\sigma^2(X) = E\{ \ [X - E(X)]^2 \ \}$$

   Both the concept of average, also called mean or expectation, and the concept of variance of a numerical r.v. were introduced by Pafnuty Chebyshev in the XIXth century.

2. Standard deviation of a r.v. X. It is the square root of the variance, therefore it is denoted σ or σ_X.

3. Independent and dependent random variables.

4. Conditional probabilities

5. Bayes theorem[8]

[8] Bayes theorem is first of all the formula relating the probabilities of events A and B and their conditional probabilities

$$P\{A|B\} = \frac{P\{B|A\}P\{A\}}{P\{B\}}$$

It is also an *interpretation* of this result in terms of "prior" and "posterior" probabilities. Note that Bayes theorem, named after Thomas Bayes (1701-1761), and its interpretation would more properly be called Laplace theorem, because it is Laplace (1749-1827) who really clarified Bayes ideas and result.

6. Covariance and correlation
7. Gaussian distributions of probabilities

Customary mistakes in elementary probabilities

When talking about probabilities, even educated people may say wrong things like: since having two accidents is very rare, and since I already had one, now the chances that I have another one are much smaller.

On a more sophisticated level, the paradox of Monty Hall, for instance, can befuddle even the best minds. Here is how it goes: there are two people, one is the guesser who must provide a best guess, the other is the operator of the game. There are three closed doors, A, B and C, facing the guesser. Behind one and only one is a prize. Step one: the guesser must make a guess, for instance guess door A. Step two: the operator, who knows where the prize is, doesn't open A, but selects among B and C a door where the prize is not, and opens it. The guesser sees this new piece of information. Step three: the guesser is invited to guess again where the prize is.

Question: Should the guesser change his or her guess, and now choose the other non opened door, or it doesn't matter?

Answer: It *does* matter. Even though in some intuitive – but misguided – view of the game it seems that the new piece of information is irrelevant, the guesser should change guess, and now choose the other non opened door. The probability of winning will go from $^1/_3$ to $^2/_3$.

Before looking up the solution, try to solve it by yourself.[9] [10]

[9] Solution: One way to see that it is judicious to change guess is to note that, if the guesser follows this strategy, he or she will lose only when the prize was behind the initial guess.

[10] This brief appendix mentioned only two examples of customary mistakes when thinking about probabilities (one very simple, one just marginally less simple). For another introduction to elementary probability, more recent than Feller's two volume book already cited, and oriented toward building an intuition about probability, we refer the reader to the book by Henk Tijms, *Understanding probability*, Cambridge, 2012.