1

Some Philosophy of Probability, Statistics, and Forensic Science

Before we embark on our task of discussing the mathematics of probability and statistics in legal and forensic science, we need to think about how probabilistic and statistical statements could or perhaps should be interpreted and understood in this context. That is, we need to agree on a philosophical interpretation of probability and statistics suitable for our purposes. This is a very important and subtle issue. In any real life application of probability it is crucial to agree (as much as we can) about what we mean when we write down probabilistic assessments about hypotheses or events. Probabilistic statements are known for their difficult interpretation, and certainly in legal and forensic affairs, vagueness about the very meaning of such questions should be avoided. Our discussion will be rather brief and we do not claim completeness in any way. The philosophy of probability is a huge subject of which we will barely scratch the surface. Apart from a philosophical position towards probability theory and statistics, we will also spend a few words on our philosophical position about forensic science itself.

1.1 The Kolmogorov Axioms of Probability

Let us start with the basic mathematical setup of probability theory. Since our focus will be on situations in which the truth is one out of only finitely many possibilities, we can restrict ourselves to probability distributions on finite outcome spaces. If Ω is such a finite outcome space, then a probability distribution on Ω is a mapping P from all subsets of Ω to $[0, 1]$ with the properties that

$$P(\Omega) = 1 \tag{1.1}$$

and

$$P(A \cup B) = P(A) + P(B), \tag{1.2}$$

for all $A, B \subset \Omega$ such that $A \cap B = \emptyset$. The quantity $P(A)$ is supposed to represent the probability that the outcome of the experiment is an element of A.

Other properties quickly follow from this definition, and we mention some of them, without proof:

(1) $P(\emptyset) = 0$;
(2) If $A \subset B$ then $P(A) \leq P(B)$. More precisely, $P(B \backslash A) = P(B) - P(A) \geq 0$;
(3) $P(A \cup B) = P(A) + P(B) - P(A \cap B)$, for all $A, B \subset \Omega$.

These properties align with our intuitive understanding of probabilities. Indeed for (2), if A implies B then the probability of B must at least be as large as the probability of A. Property (3) can be understood when you realize that when you add up the probabilities of A and B, you have counted the contribution of the intersection twice. Hence you must subtract it from the sum to obtain the probability of the union of A and B.

It is not difficult to see that a probability measure P is determined by the probabilities of the singletons in Ω. Indeed, for all $A \subset \Omega$ we have

$$P(A) = \sum_{\omega \in A} P(\{\omega\}).$$

One of the most important ideas in probability theory is the notion of conditional probability. Given a probability measure P, the conditional probability of A given B, denoted $P(A \mid B)$ is defined (if $P(B) > 0$) by

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}. \tag{1.3}$$

This quantity is supposed to represent the new probability of A once we have learned (or supposed) that B occurs. It can be understood heuristically: when we condition on B, we restrict ourselves to only outcomes in B, hence the $P(A \cap B)$ in the numerator. But in order to make sure that the total probability is still one, we have to normalize and divide by $P(B)$. Hence, B now plays the role of Ω.

The very useful Bayes rule

$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)} \tag{1.4}$$

follows from applying (1.3) twice. Although mathematically elementary, the rule is very significant in forensic and legal affairs, where it is often used in the so-called *odds form*:

$$\frac{P(H_1 \mid E)}{P(H_2 \mid E)} = \frac{P(E \mid H_1)}{P(E \mid H_2)} \times \frac{P(H_1)}{P(H_2)}. \tag{1.5}$$

Here, H_1 and H_2 typically are unobserved events, often called *hypotheses*. An example of such an event is "suspect is guilty of the crime." The event E can typically be observed, and refers to certain evidence, for instance the event that the suspect has a certain DNA profile. The left-hand side is the ratio of the probabilities of the hypotheses H_1 and H_2 after we have learned about the evidence E. The fraction at the far right is the same ratio before knowing this evidence, and the remaining fraction

$$\frac{P(E \mid H_1)}{P(E \mid H_2)} \tag{1.6}$$

is called the *likelihood ratio*. We will come back to this rule and its use and interpretation extensively and in great detail in the chapters to follow, but note already at this point that we view hypotheses as events which can be assigned a probability.

A very important question to ask now, is why the axioms in (1.1) and (1.2) would be acceptable as axioms for the theory of probability. Why can probability be axiomatized this way? Without agreement on what "probability" actually means, this question will be an impossible one to answer. The axioms are abstract mathematical notions, but they are supposed to capture the bare essentials of what we would like to call "probability." The axioms would be useless if there would be no instantiations of them that correspond to some intuitive notion of probability, whatever that notion may be.

Consider the simplest example of a probability distribution, namely the situation where each outcome is assigned the same probability. Specifically, writing $|A|$ for the number of elements in the set A, one can set

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

Clearly this P satisfies (1.1) and (1.2), and as a mathematical object P is clear. But how useful is it? How can we know that this is an applicable or useful model of what we are considering it to model? The point we want to make has been very clearly explained by Poincaré [122]:

"The definition, it will be said, is very simple. The probability of an event is the ratio of the number of cases favorable to the event to the total number of possible cases. [. . .] So we are compelled to define the probable by the probable. How can we know that two possible cases are equally probable? Will it be by convention? If we insert at the beginning of every problem an explicit convention, well and good! We then have nothing to do but to apply the rules of arithmetic and algebra, and we complete our calculation, when our result cannot be called into question. But if we wish to make the slightest application of this result, we must prove that our convention is legitimate, and we shall find ourselves in the presence of the very difficulty we thought we had avoided."

For example, suppose that we want to model tosses of a coin we have in our pocket. If we define $P(H) = 1/2$ and $P(T) = 1/2$, where H stands for heads and T for tails, then we assign equal probability to each, and hence also say that no other outcomes (e.g., landing on its edge) are considered possible. But how can we know, Poincaré asks, whether this is the "correct" probability, and what would that even be? Our mathematical definition of a function that we call probability is an abstract entity, and its definition sheds no light on the question on if or how we can see the probabilities as properties of physical objects.

We have not defined what probability is, outside the abstract mathematical framework above, is what Poincaré's quote above suggests. Instead, we can ask for concrete and useful instantiations of concepts that satisfy the axioms of Kolmogorov, and which also somehow align with our intuitive idea about what probability is.

One of the more powerful attempts in that direction is to interpret the probability of an event as the relative frequency of the occurrence of the event in a sequence of repetitions of the same experiment. A first natural idea when applying probability is that, if we were to repeat the same experiment many times independently, we would expect to see the outcomes having certain relative frequencies which are a property of the experiment, so that we may define their probabilities as these relative frequencies. It is useful and interesting to see how the natural axioms (1.1) and (1.2) and the equally natural definition (1.3) follow from interpreting probability in such a frequentistic way, and this is what we now first explore. We will conclude, however, that such a frequentistic interpretation does not solve the circularity dilemma of Poincaré: it will re-appear in a slightly disguised form. As such, it will not tell us what probability is in any ontological sense, but this does not prevent the notion of frequencies from being very useful in many circumstances.

1.2 The Frequentistic Interpretation

In many applications of probability theory we have a process or experiment in mind that can be repeated. To take a canonical example, when we flip a coin, we are inclined to think that the probability for heads to come up is 1/2. And indeed, repeating the coin flip many times will often show that about half of the time heads does come up. This empirical fact can be seen as support for the statement that the probability of heads is 1/2. Moreover, we humans have the capacity to imagine infinitely many coin flips. When we do so, we imagine that if we were able to carry out infinitely many such repetitions of the coin flip, the *relative frequency* of heads coming up would converge to 1/2. That is, if we throw the coin k times, and k_h denotes the number of heads, we imagine that k_h/k tends to 1/2 as k gets larger.

This kind of frequentistic reasoning leads to the standard Kolmogorov axioms of probability stated above in a very straightforward way. Indeed, when we flip the coin k times, and the number of heads among these k flips is k_h, then the relative frequency of heads is equal to k_h/k. The relative frequency of tails can be written as k_t/k, where k_t is the number of tails among the k flips. Since obviously

$$\frac{k_h}{k} + \frac{k_t}{k} = 1, \tag{1.7}$$

and this identity is preserved under taking the limit of more and more experiments, we see that we arrive at the first axiom (1.1), since heads and tails are the only possible outcomes.

To see that the *addition* of probabilities in (1.2) is also reasonable in such a context, imagine that we roll a die k times. Denoting the number of 1's and 2's (say) among the k outcomes by k_1 and k_2 respectively, then the relative frequency of seeing a 1 or a 2 is equal to the sum $(k_1 + k_2)/k$ of the original relative frequencies. As above, this then remains true in the limit as the number of experiments tends to infinity. This argument extends to any two subsets A and B of the outcome space which are disjoint, and leads to the second Kolmogorov axiom (1.2).

Hence frequentistic reasoning adheres to the axioms of probability. In fact, it might even be the case that many people accept (1.1) and (1.2) as the axioms of probability, *because* they are satisfied by limiting relative frequencies.

Relative frequencies also lead to the notion of conditional probabilities defined above in an elegant way. Indeed, suppose that we repeat a certain experiment k times (where k is large), and that on each occasion we observe whether or not the events A and B occur. The number of occurrences of an event E is denoted by k_E. Conditioning on B means that we only look at those outcomes for which B occurs, and disregard all other outcomes. In this smaller collection of trails, the fraction of the outcomes for which A occurs is $k_{A\cap B}/k_B$ which is equal to

$$\frac{k_{A\cap B}/k}{k_B/k}, \tag{1.8}$$

and this is indeed about $P(A \cap B)/P(B)$. Hence this latter expression is a very natural candidate for the conditional probability of A given B, again from the frequentistic point of view.

The question now is whether or not this notion of probability solves the circularity dilemma of Poincaré that we discussed above. The answer is no, and we explain this now.

Obviously, not all infinite sequences of coin flips have the property that the relative frequency of heads converges to $1/2$. In fact, if any toss can be heads or tails, then any infinite sequence can in principle occur. We could, for instance, only throw tails. In fact it is also not difficult to construct infinite sequences of coin flips for which the relative frequencies of heads do not converge at all; not to 1/2 and not to any other number. This being undeniably true, the reader might object that such a sequence of only tails will not occur in real life. Indeed, anyone repeatedly flipping a coin will see that the relative frequencies of heads will – after some time – be close to 1/2. But the key question now is: on what grounds are we going to differentiate between sequences that lead to the "right" probability of 1/2, and those that do not? That is, we need to worry about those sequences for which the relative frequency of heads does not converge to 1/2. These anomalous sequences have to be somehow ruled out as not being relevant. The only possible way to do this seems to be to say that such exceptional sequences rarely occur and form a very small minority compared to the collection of possible sequences.

This argument would boil down to saying that the overwhelming majority of the infinite sequences of coin flips lead to a relative frequency of 1/2 for heads, and that the exceptional sequences form a tiny minority. But now we face a problem: what do we mean by "exceptional"? How do we define that? The only way to do that is to say that such exceptional sequences essentially do not occur, but quantifying this again presupposes some notion of chance already. The circularity of Poincaré appears again, albeit in a slightly different form.

So although interpreting probability frequentistically is very useful, it does not solve our problem to the extent that it does not tell us what probability *is*. Relative frequencies form an instantiation of the axioms of probability, that much is clear. They do not define probability in any way, but at the same time, they are extremely useful in situations where repetitions of similar events play a role. For instance, casinos make a profit by relying on the calculus of probability, interpreted as relative frequencies. Casinos are not concerned at all with philosophical discussions about the nature of probability. They are concerned with making money, and for that purpose, relative frequencies are very useful.

In fact, the application story of relative frequencies gets even better. It turns out that in the standard Kolmogorov axiomatization of probability, laws of large numbers can be proven which are consistent with the empirical facts about convergence of relative frequencies. For instance, the probability that after k flips with the coin, the fraction heads deviates by more than a given $\epsilon > 0$ from 1/2 tends to zero as k tends to infinity.

For casinos, it is important to have a very precise idea about the relative frequencies: if a coin would come up heads with frequency 0.501, then this could harm the casino. They can use results such as those we just mentioned, to decide that a coin seems safe for use in the casino based on its relative frequency in a long series being sufficiently close to 0.5. On the other hand, if a coin is used for the toss at the beginning of a soccer match, then 0.501 as the long run frequency is clearly less relevant. Thinking of probabilities in terms of relative frequencies is extremely useful, and leads to a mathematical theory which is consistent with many empirical facts.

Closely related to the concept of relative frequencies in repeatable experiments, is the notion of population proportions. If we have a population of, say, 100 individuals of which 40 carry a certain characteristic, then one may claim that upon randomly choosing a population member, the probability that the chosen member has the characteristic is 40/100. The idea behind this is that proportions, like frequencies, satisfy the Kolmogorov axioms, as is easily verified. The fact that the phrase "randomly choosing a population member" already assumes some notion of probability in which each member has the same probability to be chosen, implies that the idea of proportion cannot be used to define what probability is either. It is, however, another instantiation of the axioms of Kolmogorov. Whenever we speak of a population *frequency* instead of a population proportion, we think of the fraction of an infinite population that would have a certain characteristic.

All this makes it tempting to associate probability with (at least conceptual) repeatability, or even to require repeatability for the assignment of probabilities. We next explain why we disagree with that.

1.3 The Inadequacy of Relative Frequencies in the Legal and Forensic Context

For our purposes, things may not be so easy and straightforward, for at least two reasons. First of all, we would like to make probabilistic statements about, say, the guilt of a certain suspect, or about the hypothesis that a given man is the father of a given child. These kind of probabilistic statements, if they make sense at all, cannot be interpreted in the same vein as the coin flips because every single legal or forensic situation is genuinely unique, and cannot be repeated. Repetitions of the coin flip would not make sense from a frequentistic point of view if the circumstances of different coin flips would not be essentially the same. The implicit assumption in the context of coin flips, or for that matter in any application of a frequentistic interpretation of probabilities, is that we repeat the *same* experiment many times. This, then, obviously often does not make sense in the legal context that we are interested in. No two legal cases are the same and this means that a phrase like "The probability that John Smith is the criminal" does not make sense from a frequentistic point of view. We have to give this statement another meaning, that is, we need another instantiation of probability than the frequentistic one.

Moreover, when two individuals have the same information, it need not be the case that they make the same probabilistic assessment. Their background knowledge and expertise may lead them to incorporate this information differently leading to different probabilistic assessments. As a result, the probabilistic analysis of a situation will differ among different people, and each such interpretation should therefore be interpreted subjectively, not objectively. So if we want to use the axioms of Kolmogorov described above in (1.1) and (1.2) we need to argue that such a subjective view on probabilities can be based on these very same axioms. In short, we need a subjective instantiation of probability.

In the next section we will give such an instantiation, namely *epistemic* probability, which intends to assign probabilities based on one's *information* or *knowledge*, and which also allows one to assign probabilities to events that already have taken place. Many probabilistic statements in legal and forensic cases can, and we think should, be interpreted from such an epistemic point of view. A judge will base the verdict on his or her personal conviction of the truth based on the information he or she has, not on the truth itself. This illustrates the fact that most probabilistic statements in legal and forensic affairs are genuinely epistemic in nature.

So, many probabilistic statements in legal and forensic affairs do not seem to make sense when interpreted frequentistically, especially when they relate to individual cases. This is, of course, not to say that frequentistic probability should be abandoned from forensic and legal contexts. To give but one example, it might be useful or even necessary to think about the quality of a legal system in terms of the fraction of wrong convictions. Such questions should be approached frequentistically, despite the fact that individual cases typically should not. Another example is the fact that one tends to think about DNA profiles in a frequentistic manner. If the actual population frequency of a given profile would be known to us, then one may interpret the subsequent checking of people's DNA in the same vein as flipping a coin: one postulates a fixed p representing the probability for any person to have the profile (ignoring familial relationships). That is, one may interpret this number p frequentistically.

But even this rather straightforward application is not easily possible without complications. This is due to the fact that in practice p is typically not known precisely. If we have no certainty about p, then upon checking people's DNA one by one, one may actually learn something about the unknown p, and then we are back in the epistemic context where one's probability assessment depends on the information one has. This information may in part, but not entirely, consist of a sample of the population and the relative frequency therein. We discuss this issue briefly in Section 1.6 below, and more extensively in Chapters 3 and 7.

Hence, there is at first glance a certain tension between epistemic and frequentistic interpretations of probability and this leads to rather interesting questions. For instance, is it possible to combine two different interpretations of probabilities (epistemic and frequentistic) into a single formula, like Bayes' rule (1.4)?

Does such a formula still make sense when the probabilities involved in the formula should be interpreted in different ways?

To briefly elaborate on this, suppose that in (1.4), H_1 is the hypothesis that John Smith committed a certain crime, H_2 the hypothesis that an unknown person did it, and that E is the evidence of finding a certain DNA profile at the scene of the crime, which we assume comes from the culprit, and which is identical to the profile of John Smith. The probabilities $P(H_i)$, $P(H_i \mid E)$ represent epistemic probabilities for a hypothesis to be true, while $P(E \mid H_i)$ can represent frequentistic probabilities for persons to have the relevant DNA profile. So we have two different interpretations of P within the same formula. The only way to meaningfully interpret the formula then, it seems, is to embrace a frequentistic number into one's epistemic probability. This makes sense, since there is no reason why empirical facts and observed frequencies cannot be incorporated into one's epistemic analysis. Epistemic probabilities are based on all our knowledge, and this knowledge may lead us to conclude that a frequentist interpretation is (exactly or approximately) appropriate, such as in the example with DNA profiles. In other words, it is reasonable to assume that epistemic interpretations include, but are more general than, frequentistic ones and that we should interpret Bayes' rule epistemically in situations like this. A remark to more or less the same effect was made by Cooke [38], where we read on page 108:

> "Subjective probabilities can be, and often are, limiting frequencies. In particular, this happens when a subject's belief state leads him to regard the past as relevant for the future in a special way."

Let us next study epistemic probability in some more detail, and investigate how such an epistemic interpretation can adhere to the axioms of Kolmogorov (1.1) and (1.2) as well. Since we want to use classical probability calculus like Bayes' rule, such an axiomatic foundation of epistemic probability is very important for our purposes.

1.4 Epistemic Probability

As we explained in the previous section, we need a well-founded interpretation of epistemic probability suitable for our purposes. Epistemic probability allows different people to make their own probabilistic analysis of a given situation. The basic idea is that the probabilistic analysis of a situation of an agent, depends on the information (including both case specific information and general knowledge) he has, and not directly on the actual truth itself. This information can vary greatly among various people.

We saw in the previous section that the idea behind frequentistic interpretations of probability is the concept of relative frequencies. For epistemic probability interpretations, a different approach is needed. The epistemic probability of an event A of a certain agent is sometimes called the *degree of belief* this agent has in the event A. If the agent knows for sure that A does not occur, then his degree of belief in A is 0, while complete confidence in the occurrence of A implies that his belief in A is equal to 1.

Although the extremal point 0 and 1 are rather easy to interpret, the values in between are not. In fact, there is no consensus in the philosophical literature that belief comes in continuous degrees at all, but for our purposes we do assume that it does. How should we define and quantify one's degree of belief in a certain event A? One approach (but not the only one) is to ask people how much money they would be willing to pay for the bet on the occurrence of the event A which pays out 1 if A actually occurs. Since the amount of money an agent wants to bet crucially depends on her or his information (and indeed probably on many other things as well), different people will be willing to pay different amounts for the same bet. Let us therefore define the notion of degree of belief this way. Here and in the sequel we refer to the agent as "he." This is meant to include both sexes.

Definition 1.4.1 The degree of belief that an agent has in an event A is the maximal amount of money he is willing to pay for a bet which pays out 1 if A actually occurs.

Obviously the degree of belief will differ among different agents. For one thing they may have different information, and moreover, not all people are willing to take the same risks. This makes this approach a genuine subjective one. Note that it is important that we ask for the *maximal* amount an agent is willing to pay. Many people would probably be willing to pay a very small amount for a bet on an event they know (or think) occurs. The key point is to determine what they are maximally willing to pay.

If we adopt this idea, then one can also assign degrees of belief to factual things. For instance, if agent A throws a die in such a way that agent A sees the result but agent B does not, then agent B will still be able to express his degree of belief in the event that the outcome is 2. The fact that the die has already been thrown is completely irrelevant for this. Agent A, however, will have a very different degree of belief in the outcome 2, because he knows whether or not it has occurred.

If you are worried about assigning nontrivial probabilities to events that we know are either factually true or not, then you should be aware of the fact that this is precisely what needs to be done in legal and forensic circumstances. If we want to assign a probability to the event that a given individual is the donor of a certain DNA profile, then this is factually either true or not. But this is not important. What matters is what we can say about this question on the basis of the information that