Probability Theory

Jacob Denson

September 10, 2016

Table Of Contents

Chapter 1

Foundations

What is probability? What do we mean when we say that you are 80% more likely to develop lung disease if you are a smoker rather than an average person. To a mathematician, probability theory is just a subfield of measure theory, in which we study the properties of abstract functions on a measure theory. To a natural scientist, probability theory is viewed in a different vein. In this chapter, we will explore the two major interpretations of probability theory in real life, each of which use the same underlying mathematical theory to make judgements about the world. After exploring these interpretations, we will make axiomatic definitions of probability, and explore basic consequences.

1.1 Frequentist Probability

Classical probability theory was developed according to the intuitions of what is now known as the frequentist school of probability theory, and is the simplest type of probability to understand. Suppose you are repeatedly performing some experiment. Even under rigorously controlled conditions, the experiment will not always result in the same outcome – we will instead have a range of outcomes which we may observe from a single result in an experiment. Nonetheless, some outcomes will occur more frequently than others. Let us perform an experiment as often as desired, obtaining an infinite sequence of outcomes

$$\omega_1, \omega_2, \omega_3, \dots$$

Let D be a proposition decidable from the outcome of the experiment (e.g. D may represent whether a flipped coin lands heads up or heads down). Mathematically, this is a subset of the set of all outcomes in an experiment – the elements for which the proposition is true. We may then define the relative frequency of this proposition being true to be

$$P_n(D) := \frac{\#\{k \leqslant n : \omega_k \in D\}}{n}$$

It is the claim (and key assumption) of the frequentist school that, if our experiments are suitably controlled, then regardless of the specific sequence of outcomes measured, our relative frequencies will always converge to a well defined, invariant number, which we define to be the probability of a certain event:

$$\mathbf{P}(D) := \lim_{n \to \infty} P_n(D)$$

Let's explore some consequences of this doctrine. First, we note $0 \leqslant P_n(D) \leqslant 1$ for any event D, so that $0 \leqslant \mathbf{P}(D) \leqslant 1$. If we let Ω denote the set of all possible outcomes to the experiment (a proposition that is true for all outcomes), then

$$P_n(D) = \frac{\#\{k \leqslant n : \omega_k \in \Omega\}}{n} = \frac{\#\{1, 2, \dots, n\}}{n} = 1$$

Thus we must define $\mathbf{P}(\Omega) = 1$. If $A_1, A_2, \dots$ is a sequence of disjoint events, representing prepositions of which only at most one can occur at any one time, then

$$P_n\left(\bigcup_i A_i\right) = \frac{\#\{k \leqslant n : \omega_k \in \bigcup A_i\}}{n} = \frac{\sum_i \#\{k \leqslant n : \omega_k \in A_i\}}{n} = \sum_i P_n(A_i)$$

Hence, in the limit, $\mathbf{P}(\bigcup_i A_i) = \sum_i \mathbf{P}(A_i)$.

1.2 Bayesian Probability

The frequentist school is sufficient to use probability theory to model experiments of science, but our own use of probability is much more general. For instance, a common utterance on the news is that "there is an

	Payoff
D occurs	$\mathbf{P}(D) - 1$
D does not occur	$\mathbf{P}(D)$

Figure 1.1: Payoff for betting against an event D

	Payoff
D occurs	$(1 - \mathbf{P}(D))$
D does not occur	$-\mathbf{P}(D)$

Figure 1.2: Payoff for betting for an event D

80% chance of downpour this evening". It is difficult to interpret this as a frequentist. Even if we see each night's temperament as an experimental trial, it is hard to convince yourself that these experiments are controlled enough to converge to a probabilistic result. The Bayesian school of probability defines a probability to be a person's individual belief in some proposition being true.

Now if a person has free rein over choosing these beliefs, we will never be able to make emperical decisions on probability theory. Thus we require that beliefs are chosen in a logical manner, which is known as a consistant choice of beliefs. Consistancy can be formulated in various ways, but my favourite is the Dutch book method, developed by the Italian probabilist Bruno de Finetti; if you assign an event D a probability $\mathbf{P}(D)$, then you are agreeing to make bets of the following character. Suppose I bet a dollar. Either I can bet against D occuring, and I stand to gain $\mathbf{P}(D)$ if I am correct, or lose $1 - \mathbf{P}(D)$ if I am wrong, or I can bet that D will occur, and lose $\mathbf{P}(D)$ if I am wrong, and gain $(1 - \mathbf{P}(D))$ if I am right. A person's probability function is inconsistant if it possible to make bets against them that will guarantee a profit: a Dutch book.

Here's an example of how the Dutch book method can be employed to obtain general rules of probability. We claim that for any event D, $0 \leqslant \mathbf{P}(D) \leqslant 1$. Suppose $\mathbf{P}(D) > 1$. Lets bet against D happening. If D occurs, I lose $1 - \mathbf{P}(D)$, an amount less than zero, so that I gain money. If D does not occur, I gain $\mathbf{P}(D) > 0$. In both circumstances, I come out on top. Similarily, if $\mathbf{P}(D) < 0$, then I can make a dutch book by betting that D will not occur. Therefore, if a probability function is consistant, it lies between 0 and 1 at any event.

It can be shown, via similar arguments, that the following pair of propositions hold:

1. $\mathbf{P}(\Omega) = 1$.

2. If $\{A_i\}$ is a countable collection of disjoint events, then $\mathbf{P}(\bigcup_i A_i) = \sum_i \mathbf{P}(A_i)$[1].

What we have shown is that consistant degrees of belief in the Bayesian system have similar properties of experimental frequencies to a frequentist. Regardless of which philosophy you agree with, you will eventually have to agree on the same fundamental principles of probability theory. Neither of these systems rest of mathematical foundations, so we need to make a rigorous model, from which we can avoid the philosophical controversies that arise. Just as the game of chess does not have to be about knights and castles, the game of probability theory does not have to be about frequencies nor degrees of belief.

1.3 Axioms of Probability

Mathematically rigorous probability theory is specified under the banner of measure theory, with a few more hypothesis. This enables us to avoid some paradoxes when performing probability theory over an infinite sample space. To the uninitiated, I apologize for the abstraction – you will have to take some propositions for granted, and trust that the terseness is necessary.

Definition. A **Probability Space** is a measure space $(\Omega, \mathcal{F}, \mathbf{P})$ such that $\mathbf{P}(\Omega) = 1$. Stylistically, Ω is known as the **Sample Space** and $\mathbf{P}$ is known as the **probability distribution** or **measure**.

Finite and countable examples are easy to construct, whereas basic examples involving uncountably many points require deep results in measure theory. Therefore, we assume the measure theory is given, and develop the probability theory as an afterthought.

Example (Uniform Probability Measure)**.** *Let Ω be a finite, non-empty set, and, for $A \subset \Omega$ define $\mathbf{P}(A) = |A||\Omega|^{-1}$. Then $(\Omega, \mathcal{P}(\Omega), \mathbf{P})$ is a probability*

[1] Definetti would have only accepted this statement for finite collections of events. Here, we allow one to make a countable number of bets at once, rather than only finitely many at any point - Allowing limit operations is very useful!

space, known as the uniform probability space. Analysis of this distributions just involves combinatorial methods.

Example (Defining Probabilities with Integrals)**.** *Let f be a positive measurable function from some measure space $(\Omega, \mathcal{F}, \mu)$ to $\mathbf{R}$, with $\int f \, d\mu = 1$. For any measurable $A \in \mathcal{F}$, define $\mathbf{P}(A) = \int_A f \, d\mu$. Then $(\Omega, \mathcal{F}, \mathbf{P})$ is a probability space.*

Example (Uniform Probability Measure on the Unit Interval)**.** *We wish to construct a uniform measure on $[0,1]$, analogous to the finite case. Since $[0,1]$ has infinitely many points, we cannot assign positive probability to singletons. Let μ be the Lebesgue measure on $[0,1]$, and $f = 1$ be defined on $[0,1]$. By the last example, we obtain a probability space, known as the uniform probability measure.*

Example (Normal Distribution)**.** *Define a function $f : \mathbf{R} \to \mathbf{R}$ by the formula*

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

The probability space generated by f is known as the **normal distribution** *with mean μ and standard deviation σ.*

Theorem 1.1. *Let $(\Omega, \mathcal{F}, \mathbf{P})$ be an arbitrary probability space:*

1. *For $A \in \mathcal{F}$, $\mathbf{P}(A^c) = 1 - \mathbf{P}(A)$.*

Note that this is not a mathematic definition of probability, but instead an instead an interpretation of the axiomatic foundations. Another, more complicated interpretation is the bayesian, which will come later. It is also important to see that just because the probability of some event is 0 does not imply that A is impossible, just that is almost certainly will not happen (Just because a sequence converges to 0 does not imply the sequence is 0 everywhere). Similarily, just because the probability of some event is 1 does not imply that the measurement will always occur in an experiment. The only certain event is Ω, and the only impossible event is $\varnothing$.

Using the basic axioms of probability theory, some easy properties of probabilities occur:

- For any event A, $\mathbf{P}(A^c) = 1 - \mathbf{P}(A)$:

Proof. A and A^c are disjoint, and the union of the two is X □

- $\mathbf{P}(A^c) = 1 - \mathbf{P}(A)$

  Proof. $\Omega = A \cup A^c$ □

- $\mathbf{P}(\varnothing) = 0$

  Proof. $\varnothing = \Omega^c$ □

- If A is a subset of B, $\mathbf{P}(A) \leqslant \mathbf{P}(B)$

  Proof. The property follows as $\mathbf{P}(B) = \mathbf{P}(A) + \mathbf{P}(B - A)$, and $\mathbf{P}(B - A) \geqslant 0$ □

- For any events A and B, $\mathbf{P}(A \cup B) = \mathbf{P}(A) + \mathbf{P}(B) - \mathbf{P}(A \cap B)$:

  Proof. $A \cup B$ is the union of three disjoint events $A \cap B^c$, $B \cap A^c$, and $A \cap B$. The following calculation results in the equation:

$$\begin{aligned}
\mathbf{P}(A \cup B) &= \mathbf{P}(A \cap B^c) + \mathbf{P}(A \cap B) + \mathbf{P}(A^c \cap B) \\
&= \mathbf{P}(A \cap B^c) + \mathbf{P}(A \cap B) + \mathbf{P}(A^c \cap B) + \mathbf{P}(A \cap B) - \mathbf{P}(A \cap B) \\
&= \mathbf{P}((A \cap B^c) \cup (A \cap B)) + \mathbf{P}((A^c \cap B) \cup (A \cap B)) - \mathbf{P}(A \cap B) \\
&= \mathbf{P}(A) + \mathbf{P}(B) - \mathbf{P}(A \cap B)
\end{aligned}$$

□

Less obvious is the following theorem, called 'the continuity of probabilities'. Let (A_i) be a sequence of events such that $A_i \subseteq A_j$ for all $j \geqslant i$. Then $\mathbf{P}(\cup_{i=1}^{\infty} A_i) = \lim_{n \to \infty} A_n$:

Proof. Give the sequence (A_i), define a new sequence (B_i) recursively by $B_1 = A_1$, and $B_n = A_n - \cup_{i=1}^{n-1} B_i$. These are disjoint and there is union is the same as the union of A_i, so $\mathbf{P}(\cup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} \mathbf{P}(B_i) = \lim_{n \to \infty} A_n$. □

It is easy to dualize the continuity of probabilities. If the set sequence is decreasing, the same property holds for intersections.

Bibliography

[1] Larry Wasserman, *All of Statistics*

[2] Walter Rudin, *Real and Complex Analysis*