INTRODUCTION TO PROBABILITY

Josef Lauri ©
Department of Mathematics
University of Malta

1 Introduction

We all have strong intuitive notions of what it means to say that an event is likely or unlikely to happen, or that one event is more likely to happen than another. We also frequently refer to the "average " behaviour of some random phenomenon, and we expect that, if the phenomenon is repeated often, the aggregate of the all outcomes will tend towards this average. These notions are very often sufficient for us to make every-day decisions. However, we sometimes want to give a more precise quantitative measure of our idea of the probability of an event. We then use these measures in calculations and the whole concept of probability takes on the status of a mathematically exact discipline. But if we want to use probability this way, we must first define it in the language of mathematics (sets, functions, etc) and we must define clear unambiguous rules (axioms) for carrying out calculations with probabilities.

It has taken mathematicians many years to work out which is the best set of axioms for the theory of probability. We shall describe them in the next section. Here we shall only give a very brief justification for these axioms. However, you should see the subsequent sections and the rest of the course as the real justification, because the ultimate test of any set of axioms is the theory which flows out from it: Do the axioms mirror our intuitive notions of probability? Do we gain new insights? Do we obtain useful results? Is the theory created mathematically interesting? Hopefully this course will give you affirmative answers to these questions.

So let us start with the simplest definition of probability, one with which you might already be familiar. Suppose we carry out an "experiment" which could have several possible outcomes. Suppose we are interested in a particular "event" A. Some of the outcomes of the experiment result in A others do not. Then, *if all possible outcomes are equally likely*, we often define the probability of A as

$$\frac{\text{Number of outcomes which result in } A}{\text{Number of all outcomes}}.$$

For example, suppose that the experiment consists in throwing a fair (or unbiased) die (that is, the numbers 1 to 6 are all equally likely to come up on top). Let A be the event "A prime number is obtained with the toss". There are six possible outcomes for this experiment and three of these, 2, 3, and 5, give the event A. The above definition therefore gives that the probability of A is $2/6 = 1/2$.

What are the components which make up this example? We have a set of outcomes or "elementary events", $\{1, 2, 3, 4, 5, 6\}$. Let us denote this set by Ω. These elementary events are not the only ones in which we are interested, as we saw above. We could be interested in events which are made up of a number of elements of Ω. In the above example, the event A can be considered to be the subset $\{2, 3, 5\}$ of Ω. Therefore, apart from the set Ω we need to work with "events" which can be represented as subsets of Ω. With each of these events or subsets we have to associate a number which we shall then call the probability of the event. If we adopt the above definition of probability, then given such

a subset A, the probability of A becomes $|A|/|\Omega|$. The three components are therefore: the set Ω, a collection of subsets of Ω, and an assignment of a number (the probability) to each subset in the collection.

In many simple cases the following setup is therefore sufficient: Let Ω be a set (which will be chosen to model appropriately the particular situation at hand), let $\mathcal{F}$ be the set of all subsets of Ω (that is, all possible combination of elementary events will be considered), and, for any set A in $\mathcal{F}$, let the probability of A be defined to be $|A|/|\Omega|$.

There are, however, difficulties with this. What if the elementary events are not equally likely (biased die, for example)? What if Ω is an infinite set? The axioms below overcome these and other difficulties by slight refinements of the above ideas. The crucial point to keep in mind is that these axioms will not tell us how to assign probabilities to events (as we did above with the formula $|A|/|\Omega|$). That will depend on the situation in question. The axioms will however lay down rules which these assigned probabilities *must* follow if we are to call them probabilities in a mathematical sense. The theory which then follows from these axioms would then apply to any situation we would study, provided we have followed these rules in assigning probabilities.

These rules are mostly a formalisation of what properties we would naturally expect probabilities to possess. We can illustrate this with two examples.

Suppose we are throwing a biased die. If you are told that the probabilities of obtaining 1, 2, 3, 4, 5, or 6 are, respectively, $\frac{1}{9}, \frac{1}{9}, \frac{1}{9}, \frac{1}{3}, \frac{1}{6}, \frac{1}{6}$, you probably would not object. However you might find it strange if these probabilities were $\frac{1}{9}, \frac{1}{9}, \frac{1}{3}, \frac{1}{3}, \frac{1}{6}, \frac{1}{6}$ or $\frac{1}{9}, \frac{1}{9}, \frac{1}{9}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}$. (Would you? Why?)

Suppose, again, we are throwing a fair die. Let A be the same event as above, that is, that a prime number comes up. Let B be the event that 1 or 6 comes up. The probability of A is $\frac{1}{2}$, and that of B is $1/3$. Let X be the event that either A or B occurs. Note that, as subsets of Ω, A and B are disjoint and $X = \{1, 2, 3, 5, 6\} = A \cup B$. The probability of X is $\frac{5}{6}$ and this equals the sum of the probabilities of A and B. But now let C be the event that an odd number comes up (therefore it has probability $\frac{1}{2}$), and let Y be the event that either A or C occurs. Now, A and C are not disjoint, $Y = \{1, 2, 3, 5\}$ which is again $A \cup B$, and the probability of Y is $\frac{2}{3}$ which is not equal to the sum of the probabilities of A and B.

With these motivations at the back of our minds we can proceed to the presentation of the basic axioms of probability.

2 Axioms and examples

Definition. A *probability space* is defined to be a triple $(\Omega, \mathcal{F}, \mathrm{P})$ such that:

1. Ω is a nonempty set. It is called the *sample space*. An element $\omega \in \Omega$ is called an *elementary event*.

2. $\mathcal{F}$ is a collection of subsets of Ω. It is called the *event space* and any $A \in \mathcal{F}$ is called an *event*. The collection $\mathcal{F}$ must satisfy the following conditions:

(a) $\mathcal{F}$ is non-empty.

(b) If A is in $\mathcal{F}$, then its complement $A^c = \Omega - A$ is also in $\mathcal{F}$.

(c) If $A_1, A_2, \cdots \in \mathcal{F}$, then $\cup_i A_i$ is also in $\mathcal{F}$.

3. P is a function from $\mathcal{F}$ to $[0,1]$. It is called a *probability measure* on $(\Omega, \mathcal{F})$. P must satisfy the following conditions:

(a) $\mathrm{P}(\Omega) = 1$.

(b) If $A_1, A_2, \ldots$ are mutually disjoint events in $\mathcal{F}$, then $\mathrm{P}(\cup_i A_i) = \sum_i \mathrm{P}(A_i)$.

□

In many of the simpler examples we shall encounter, the situation being studied can be modelled by a probability space in which Ω is a finite or a countable set, and $\mathcal{F}$ is the set of all subsets of Ω (such an $\mathcal{F}$ would certainly satisfy the above conditions required of it). In this case we only need to define $\mathrm{P}(\{\omega\})$ (which we shall write as $\mathrm{P}(\omega)$ for simplicity) for every $\omega \in \Omega$, because then, for any $A \in \mathcal{F}$, $\mathrm{P}(A)$ is given by

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

since $A = \cup_{\omega \in A} \{\omega\}$ and since P is additive over a union of a countable number of disjoint sets. Moreover, if $\Omega = \{\omega_1, \ldots, \omega_N\}$ is finite and the elementary events are equiprobable (that is, $\mathrm{P}(\omega_i) = \mathrm{P}(\omega_j)$ for all $1 \leq i, j \leq N$), then, each $\mathrm{P}(\omega)$ must equal $\frac{1}{N}$, and, for each $A \in \mathcal{F}$,

$$\mathrm{P}(A) = \frac{|A|}{N},$$

recovering the intuitive definition we had given in the first section. However, we shall see that we cannot always employ this simple definition of probability measure, even when we have equiprobable elementary events.

Examples

1. Two fair coins are tossed. There are four possible outcomes for a single toss of the two coins. We can let Ω be the set $\{HT, TH, HH, HT\}$ (the meaning of the symbols should be clear: H denotes heads and T denotes tails; the first letter refers to one coin and the second one to the other), and we let $\mathcal{F}$ be the set of all subsets of Ω. Since all elementary events are equally likely, we define $\mathrm{P}(\omega) = \frac{1}{4}$ for every $\omega \in \Omega$.

2. Two biased coins are tossed. For each coin, heads are twice as likely to come up as tails. We can let Ω and $\mathcal{F}$ be as in the previous example. But what assignment of probabilities to the elementary events would model the situation now? We let $\mathrm{P}(HT) = \mathrm{P}(TH) = \frac{2}{9}, \mathrm{P}(HH) = \frac{4}{9}$ and $\mathrm{P}(TT) = \frac{1}{9}$. Then, for example, the probability of the event "Different values are obtained when the coins are tossed" is given by

$$\mathrm{P}(\{HT, TH\}) = \mathrm{P}(HT) + \mathrm{P}(TH) = \frac{4}{9}.$$

You might well ask how the above probabilities were assigned. It seems to involve a "multiplication" of probabilities: for each coin, the probability of heads is $\frac{2}{3}$, probability of tails is $\frac{1}{3}$ (this would make heads twice as likely as tails), and the probability of a pair is the product of the corresponding probabilities. But we still have no rules which tell us if and how we can multiply probabilities, so this calculation at this stage remains an intuitive justification. What you are expected to understand this point is simply that the triple $(\Omega, \mathcal{F}, \mathrm{P})$ which we have defined does satisfy the axioms of a probability space (this is easy to check) and, equally importantly, that it represents satisfactorily the situation we are trying to model. Thus, let A be the event "The first coin comes up heads" and let B be the event "The first coin comes up tails". As subsets of Ω, $A = \{HH, HT\}$ and $B = \{TH, TT\}$. Therefore $\mathrm{P}(A) = \mathrm{P}(HH) + \mathrm{P}(HT) = \frac{2}{3}$ and $\mathrm{P}(B) = \mathrm{P}(TH) + \mathrm{P}(TT) = \frac{1}{3}$, verifying that, for the first coin, heads is twice as likely to come up as tails. The same verification can be carried out for the second coin.

3. Two fair dice are thrown. The sample space consists of all the possible pairs of numbers which can be obtained (thirty-six pairs in all), that is $\Omega = \{(1,1), (1,2), \ldots, (6,5), (6,6)\}$. $\mathcal{F}$ can be defined to be the set of all subsets of Ω, and $\mathrm{P}((i,j)) = \frac{1}{36}$ for every pair (i,j). Let A be the event "A total of five is obtained". Then A is the subset $\{(1,4), (2,3), (3,2), (4,1)\}$ and $\mathrm{P}(A) = \frac{1}{9}$. What number is most likely to be obtained as a total?

4. Sometimes, the same "experiment" can give rise to different probability spaces, depending on what events we are interested in. For example, suppose ten fair coins are tossed. What would be the probability spaces if (i) every outcome is of interest; (ii) the number of tails is of interest.

In the first case, Ω consists of all sequences such as $HHTHTTHTTH$ (compare Example 1). Therefore Ω contains 2^{10} elements, $\mathcal{F}$ is the set of all subsets of Ω and for any event $A \subseteq \Omega$, $\mathrm{P}(A) = |A|/2^{10}$. In the second case, there are eleven possible outcomes, that is, $\Omega = \{0, 1, 2, \ldots, 10\}$. Therefore for any event $A \in \Omega$ we now have $\mathrm{P}(A) = |A|/11$.

5. Greater care has to be taken when Ω is infinite. It is easy to ask meaningless questions. For example, suppose the experiment consists in picking a positive integer, all choices being equally likely. Then $\Omega = \mathbb{N}$. We might ask: What is the probability that the number 13 is chosen? Or, what is the probability that an even number is picked? One might think that the answer to this second question is $\frac{1}{2}$, but it is not difficult to discover that this "probability space" is meaningless, at least in a mathematical sense. For, since all integers have an equal probability of being chosen, let this probability be k. Then $\mathrm{P}(\Omega) = \sum_1^\infty k$. If $k > 0$ then this sum is ∞ and if $k = 0$ then it is equal to 0, and both cases are wrong since we require $\mathrm{P}(\Omega) = 1$. Example 18 in Chapter 1 of [GW] describes a way of making mathematical sense of this situation.

However, we shall encounter simple instances of infinite probability spaces. Consider this example. A number is chosen in the interval $[\alpha, \beta]$. We let $\Omega = [\alpha, \beta]$ and, for any $\omega \in \Omega$, we must have $\mathrm{P}(\omega) = 0$, otherwise $\mathrm{P}(\Omega)$ would be infinite. However, the questions we should be asking are of the following type: What is the probability that the number chosen is in the range $[a, b]$ for given

$\alpha \leq a < b \leq \beta$. We let $\mathcal{F}$ be the set of all subintervals and union of subintervals of $[0,1]$. If A is the event "The number chosen is in the interval $[a,b]$" then $\mathrm{P}(A)$ is defined to be $(b-a)/(\beta-\alpha)$. This way we obtain a probability space.

6. Here is an example of an infinite but countable probability space. Suppose a fair coin is thrown until the first occurrence of heads. What are the elementary events here. We could have heads on the first throw, and we denote this event by H, or on the second throw, TH, or on the nth throw, $T^{n-1}H$. It could also happen that we keep tossing the coin and no heads ever appears. Let us denote this event by T^∞. Therefore let $\Omega = \{T^\infty, H, TH, T^2H, \dots\}$ and let $\mathcal{F}$ be the set of all subsets of Ω. Now, what probabilities are we to assign to the elementary events? Anticipating what should become clearer after we consider independent events and products of probabilities, it would seem reasonable to let $\mathrm{P}(T^{n-1}H) = (\frac{1}{2})^n$. Also, it seems intuitively acceptable that the chance of obtaining no heads in an infinite sequence of tosses is zero (a rigorous proof of this is given in Example 25 of Chapter 1 in [GW]). Therefore we let $\mathrm{P}(T^\infty) = 0$. Is this a probability space? What should be checked is that $\mathrm{P}(\Omega) = 1$. But $\mathrm{P}(\Omega) = \sum_{\omega\in\Omega} \mathrm{P}(\omega) = \sum_{n=0}^{\infty} (\frac{1}{2})^n = 1$, as required. □

The axioms of a probability space give us only the starting point for calculating with probabilities. But just as is the case for other axiomatic systems you are studying (groups, vector spaces, etc) we can deduce a number of useful rules from the given axioms. Several of these are obtained in [GW] to which the student is referred. We shall here have a look at a few cases. More details are given in Chapter 1 of [GW].

For example, it follows from the axioms that $\mathrm{P}(\emptyset) = 0$, because $1 = \mathrm{P}(\Omega) = \mathrm{P}(\Omega \cup \emptyset) = \mathrm{P}(\Omega) + \mathrm{P}(\emptyset) = 1 + \mathrm{P}(\emptyset)$, since Ω and $\emptyset$ are disjoint. It is equally easy to show that $\mathrm{P}(A^c) = 1 - \mathrm{P}(A)$ (that is, if the probability of an event happening is p, say, then the probability of it not happening is $1-p$). For, $1 = \mathrm{P}(\Omega) = \mathrm{P}(A \cup A^c) = \mathrm{P}(A) + \mathrm{P}(A^c)$, since A and A^c are disjoint.

While the axioms tell us that P is additive over disjoint unions, we still have to work out what happens when calculating probabilities of arbitrary unions of events. Suppose A and B are events, not necessarily disjoint. Then $A = (A-B) \cup (A \cap B)$, therefore

$$\mathrm{P}(A) = \mathrm{P}((A-B) \cup (A \cap B)) = \mathrm{P}(A-B) + \mathrm{P}(A \cap B)$$

since $A - B$ and $A \cap B$ are disjoint. Similarly

$$\mathrm{P}(B) = \mathrm{P}(B-A) + \mathrm{P}(A \cap B).$$

Adding we obtain that

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

where we have used the fact that $A - B$, $A \cap B$ and $B - A$ are mutually disjoint (drawing a Venn diagram would make the above very clear). Now this

is beginning to look like the inclusion-exclusion principle, and in fact the above can be generalised to the following.

Lemma 1. The Inclusion-Exclusion Principle in Probability. *Let $A_1, A_2, \ldots, A_n$ be events, where $n \geq 2$. Then*

$$P\left(\bigcup_{i=1}^{n} A_i\right) = \sum_i P(A_i) - \sum_{i<j} P(A_i \cap A_j) + \sum_{i<j<k} P(A_i \cap A_j \cap A_k) - \cdots + (-1)^{n+1} P(A_1 \cap A_2 \cap \cdots \cap A_n).$$

Proof. We can prove this in various ways. Firstly, if we are dealing with a finite probability space in which, for each event A, $\mathrm{P}(A) = |A|/|\Omega|$, then we can simply invoke the inclusion-exclusion principle and divide by $|\Omega|$. But we want to prove this for a general probability space. We could prove it by induction on n, the case $n = 2$ having been dealt with above. Instead we shall prove it in a way very similar to how we have proved the inclusion-exclusion principle. First we need a definition.

Let $S = A_1 \cup A_2 \cup \cdots \cup A_n$. Let us partition S into what we shall call *atoms* in this way: two elements x, y of S will be in the same atom iff x and y belong exactly to the same sets from $A_1, A_2, \ldots, A_n$. For example, the atoms of the union of sets shown in the figure below are $X_1, X_2, \ldots, X_8$.

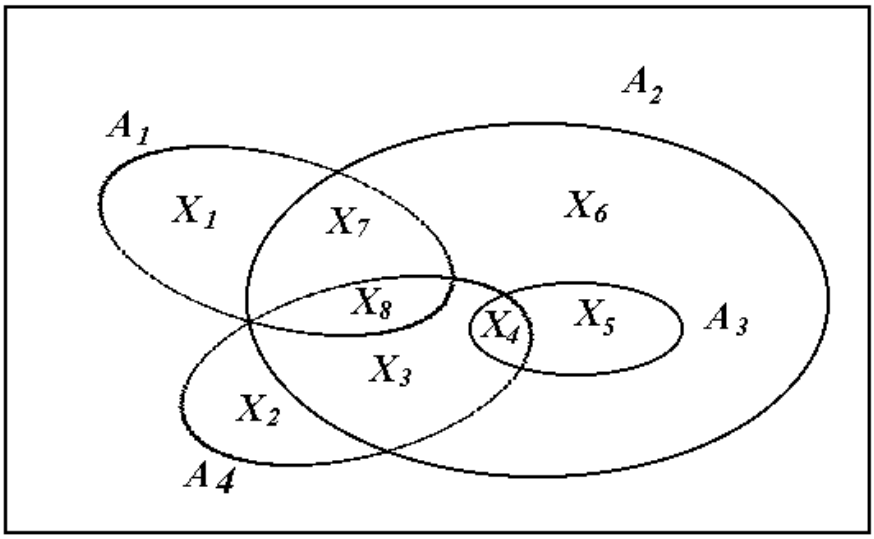

Figure 1: The X_i are the atoms formed by the intersection of the A_j

Now, let $X_1, X_2, \ldots, X_r$ be the atoms of S. Since the atoms form a partition of S, $\mathrm{P}(S) = \mathrm{P}(X_1) + \cdots + \mathrm{P}(X_r)$. Therefore, what we have to do is to count, for

every atom, its contribution to the sum of the probabilities on the right-hand side of the above equation. If, for every atom X, this contribution equals $\mathrm{P}(X)$, then the result is proved.

So, suppose that an atom X belongs to exactly t of the sets $A_1, \ldots, A_n$. The net contribution of X to the right-hand side is

$$\begin{aligned}\binom{t}{1}\mathrm{P}(X) \qquad & -\binom{t}{2}\mathrm{P}(X) + \cdots + (-1)^{t+1}\binom{t}{t}\mathrm{P}(X) \\ & = [1 - (1 - \binom{t}{1} + \binom{t}{2} - \cdots + (-1)^t\binom{t}{t})]\mathrm{P}(X) \\ & = \mathrm{P}(X)\end{aligned}$$

as required. □

Exercises

1. `Go through the first five sections of Chapter 1 of [GW].`

2. `Do exercises 9, 14, 15 and 16 of Chapter 1 in [GW].`

Problem.

1. Do Problem 1 of Chapter 1 in [GW]. For this you might find helpful the identity

$$(1+x)^n + (1-x)^n = 2\sum_{0}^{\lfloor \frac{n}{2} \rfloor} \binom{n}{2i} x^{2i}$$

which, of course, you should be able to prove.

2. A survey was held to determine viewership of radio stations. Radio STB (Simply The Best) did not do particularly well: The survey involved a sample of eight hundred respondents (we can assume that the sample was randomly chosen, that is, all members of the population had equal chances of being chosen), and of these only twelve said that they listened to STB. The directors of STB were quite irate, so, to check the survey's veracity they picked three hundred of their regular listeners (again randomly, of course) and asked them if they had been interviewed for this survey. None of them had, and the directors of STB concluded that, obviously, something must have been very wrong with the survey.

Assuming a population of 300,000, and supposing that the survey's findings were correct, that is, the proportion who listen to STB is 12 out of 800, find the probability that, picking 300 STB listeners at random, none of them were interviewed for the interview. Is this an unlikely event? Were the directors of STB justified in dismissing the original survey because of the result given by their survey?

3 Conditional probability and independent events

Suppose we are working with a probability space and we are given that an event B has happened. We therefore need to calculate probabilities subject to this extra information. Therefore, instead of asking for $\mathrm{P}(A)$, the probability of A, we now ask for $\mathrm{P}(A|B)$, the *probability of A given B*. Let us, for simplicity, first consider this in the context of a finite probability space with all elementary events equally likely. $\mathrm{P}(A)$ would therefore simply be $|A|/|\Omega|$. But now, since we are given that B has happened, we can restrict our "universe" of elementary events to those contained in B. Only those elements of A which are contained in B would now contribute to the probability of A. Therefore we would have that

$$\begin{aligned}\mathrm{P}(A|B) &= \frac{|A\cap B|}{|B|}\\ &= \frac{|A\cap B|}{|\Omega|}\frac{|\Omega|}{|B|}\\ &= \frac{\mathrm{P}(A\cap B)}{\mathrm{P}(B)}.\end{aligned}$$

It turns out that the rule $\mathrm{P}(A|B) = \mathrm{P}(A\cap B)/\mathrm{P}(B)$ which has been obtained in this simple situation provides just the right property which can be used as the *definition* of conditional probability in the general case.

Definition. Let $(\Omega, \mathcal{F}, \mathrm{P})$ be a probability space, and let $A, B \in \mathcal{F}$, with $\mathrm{P}(B) > 0$. Then the *(conditional) probability of A given B*, denoted by $\mathrm{P}(A|B)$ is defined by

$$\mathrm{P}(A|B) = \frac{\mathrm{P}(A\cap B)}{\mathrm{P}(B)}.$$

□

Examples

1. Suppose that an urn contains three red balls and seven white balls. One ball is removed at random, and then, without replacing it, a second ball is picked. Let B be the event "First ball is white" and let A be the event "Second ball is white". Then $\mathrm{P}(A) = \frac{7}{10}$ and, without even resorting to the definition, we would expect that $\mathrm{P}(A|B) = \frac{6}{9} = \frac{2}{3}$, since, for the second choice, there would be six white balls left from a total of nine. Let us see how this checks with the definition. First we define Ω. Let the red balls be denoted by $r_i, 1 \le i \le 3$ and let the white balls be denoted by $w_j, 1 \le j \le 7$. Then we can let Ω be the set of all pairs (x, y) where x and y can be any of the elements r_i, w_j, provided they are distinct; the first and second entries in each pair denote the results of the first and second choices, respectively. Therefore $|\Omega| = 10.9 = 90$. Then B contains all those pairs with a w as the first element; that is $|B| = 7.9 = 63$. Similarly, A contains all those pairs with a w as a second element, and therefore $A\cap B$ (which is the event that *both* A and B occur) contains all those pairs with w's in both positions; therefore $|A\cap B| = 7.6 = 42$. Hence, according to the definition, $\mathrm{P}(A|B) = \frac{42}{63} = \frac{2}{3}$.

2. As we can see from the previous example, it is sometimes easier to work out the conditional probability $\mathrm{P}(A|B)$ directly from the problem than to calculate $\mathrm{P}(A \cap B)$ and then use the definition. In fact, we often employ conditional probabilities to calculate the probability of intersections of two events (that is, the probability that both occur) rather than the other way round. For example, suppose we have the same urn as in Example 1, and two balls are picked at random. What is the probability that both are white if the balls are picked (i) without replacement and (ii) with replacement.

Let A be the event that the second ball is white and B the event that the first ball is white. We require $\mathrm{P}(A \cap B)$, which equals $\mathrm{P}(B).\mathrm{P}(A|B)$. In the first case this equals $\frac{7}{10}.\frac{6}{9}$ whereas in the second case it equals $\frac{7}{10}.\frac{7}{10}$, since $\mathrm{P}(A|B) = \mathrm{P}(A)$ because the first ball is replaced. □

Exercises

1. Redo the second example by elementary counting. That is, in the first case, there are $\binom{10}{2}$ possible elementary events, and $\binom{7}{2}$ of these give the event ''both balls white'', while in the second case there are 10^2 possible outcomes and 7^2 give the event ''both balls white''.

2. Suppose that both $\mathrm{P}(A)$ and $\mathrm{P}(B)$ are not zero. Prove that

$$\mathrm{P}(A|B)\mathrm{P}(B) = \mathrm{P}(B|A)\mathrm{P}(A).$$

This simple formula will be very useful for ''inverting'' conditional probabilities. Consider the following simple example. The probability that it is sunny on any given morning is 0.4. A particular student is on time for the morning lecture with probability 0.8, but if it is sunny this probability goes down to 0.6. If the student turns up on time for the morning lecture, what is the probability that it is sunny?

Example 2 above introduces us to the idea of independent events. When balls are removed with replacement, the second event is independent of the result obtained for the first event, that is $\mathrm{P}(A|B) = \mathrm{P}(A)$. In this case, $\mathrm{P}(A \cap B) = \mathrm{P}(B)\mathrm{P}(A|B) = \mathrm{P}(B)\mathrm{P}(A)$, that is, the probability of two independent events happening is the product of the probabilities of the two events. This leads us to make the following *definition* of independent events.

Definition. Two events A, B are said to be *independent* if

$$\mathrm{P}(A \cap B) = \mathrm{P}(A)\mathrm{P}(B).$$

□

Looking back at Examples 2 and 6 in Section 2 we see that the probabilities were assigned to the elementary events in a way which made successive tosses of the coin independent events according to the above definition.

Example

3. A biased coin is tossed repeatedly. The probability of heads is p and the probability of tails is $q = 1 - p$. Define the following events:

$$\begin{aligned} A_n &= \text{First heads occurs on the } n\text{th throw} \\ B &= \text{First heads appears after the third throw.} \end{aligned}$$

Then, $\mathrm{P}(A_n) = q^{n-1}p$ and

$$\begin{aligned}\mathrm{P}(B) &= \mathrm{P}(A_4 \cup A_5 \cup \dots) \\ &= \mathrm{P}(A_4) + \mathrm{P}(A_5) + \dots \\ &= \sum_{n=4}^{\infty} q^{n-1}p \\ &= \frac{q^3 p}{1-q} = q^3.\end{aligned}$$

Also,

$$\begin{aligned}\mathrm{P}(&\text{ At least one head in the next two tosses}|B) \\ &= \mathrm{P}(A_1 \cup \dots \cup A_5 | A_4 \cup A_5 \cup \dots) \\ &= \frac{\mathrm{P}(A_4 \cup A_5)}{\mathrm{P}(A_4 \cup A_5 \cup \dots)} \\ &= \frac{q^3 p + q^4 p}{q^3} = p + qp.\end{aligned}$$

Note that this is the same as the probability of at least one head in the first two tosses, that is, the previous tails do not effect future outcomes—if we are waiting for the first heads and we have obtained a run of tails, it is as if we are starting the process from scratch. This is what one would expect from a sequence of independent trials. □

Exercise

3. `Go through Sections 1.6 and 1.7 of [GW].`

Problems

1. A FAIR DIE IS THROWN REPEATEDLY. IT HAS TWO FACES COLOURED BLUE, TWO RED AND TWO GREEN (SO IT REALLY IS AN UNBIASED DIE). SHOW THAT THE PROBABILITY THAT NOT ALL COLOURS OCCUR IN THE FIRST $k\,(>0)$ THROWS IS

$$3\left(\frac{2}{3}\right)^k - 3\left(\frac{1}{3}\right)^k.$$

[HINT: CONSIDER THE UNION OF THREE EVENTS, AND USE INCLUSION-EXCLUSION.]

2. TWO MARKSPERSONS, JACK AND JILL, SHOOT AT A TARGET ALTERNATELY, JACK STARTING FIRST. AT EACH TURN, THE PROBABILITY THAT JACK HITS THE TARGET IS $p_0\,(\leq 0.5)$ WHEREAS THE PROBABILITY THAT JILL HITS IS p_1. SHOW THAT

$$\mathrm{P}(\text{JACK WINS}) = \frac{p_0}{1 - q_0 q_1}$$

AND

$$\mathrm{P}(\text{JILL WINS}) = \frac{q_0 p_1}{1 - q_0 q_1}$$

WHERE $q_i = 1 - p_i$.

4 The Partition Theorem

The proof of the following result is easy and can be found in Section 1.8 of [GW].

The Partition Theorem. *Let $\{B_1, B_2, \dots\}$ be a partition of Ω such that $P(B_i) > 0$ for each i. Then, for every $A \in \mathcal{F}$,*

$$P(A) = \sum_i P(A|B_i)\,P(B_i).$$

□

Example

1. There are two urns, Urn I contains three red balls and four white balls, and Urn II contains two red balls and six white balls. A fair die is thrown. If the result is a six, a ball is picked at random from Urn I, but otherwise it is picked from Urn II. What is the probability that the ball chosen is white?

Let W be the event "The ball chosen is white", and let S be the event "A six comes up". Then $\{S, S^c\}$ forms a partition of the sample space, so that

$$\begin{aligned} \mathrm{P}(W) &= \mathrm{P}(W|S)\mathrm{P}(S) + \mathrm{P}(W|S^c)\mathrm{P}(S^c) \\ &= \frac{3}{7} \times \frac{1}{6} + \frac{2}{8} \times \frac{5}{6} = \frac{147}{168}. \end{aligned}$$

□

The Partition Theorem is often combined with the trick of "inverting" conditional probabilities which we have seen in Exercise 2. Suppose, in the previous example, we are given that a white ball was picked. What is the probability that Urn I was chosen (that is, a six came up)? We now require $\mathrm{P}(S|W)$ which, as we have seen in Exercise 2, equals

$$\mathrm{P}(W|S).\frac{\mathrm{P}(S)}{\mathrm{P}(W)} = \frac{3}{7}.\frac{1/6}{47/68} = 0.10.$$

This method gives what is usually called Bayes' Theorem.

Bayes' Theorem. *Let $\{B_1, B_2, \dots\}$ be a partition of Ω such that $P(B_i) > 0$ for each i. Then, for every $A \in \mathcal{F}$,*

$$P(B_j|A) = \frac{P(A|B_j)P(B_j)}{\sum_i P(A|B_i)P(B_i)}.$$

Proof. This simply involves inverting conditional probabilities and applying the Partition Theorem to the denominator.

$$\begin{aligned} \mathrm{P}(B_j|A) &= \frac{\mathrm{P}(A|B_j)\mathrm{P}(B_j)}{\mathrm{P}(A)} \\ &= \frac{\mathrm{P}(A|B_j)\mathrm{P}(B_j)}{\sum_i \mathrm{P}(A|B_i)\mathrm{P}(B_i)}. \end{aligned}$$

□

Example

2. A test for a certain type of cancer has a 5% chance of giving a positive result even when the patient is not ill. Also, it always gives a positive result in the case of an ill patient. Suppose that 0.1% of the population have the disease. If a test applied on a randomly chosen member of the population gives a positive result, what is the probability that the person really has the disease?

Let us first do this using Bayes' Theorem. Let B_1 be the event that the patient is ill and let $B_2 = B_1^c$ be the event that the patient is not ill. Let A be the event that the test is positive. Then we require

$$\begin{aligned} \mathrm{P}(B_1|A) &= \frac{\mathrm{P}(A|B_1)\mathrm{P}(B_1)}{\mathrm{P}(A|B_1)\mathrm{P}(B_1) + \mathrm{P}(A|B_2)\mathrm{P}(B_2)} \\ &= \frac{1 \times \frac{1}{1000}}{1 \times \frac{1}{1000} + \frac{5}{100} \times \frac{999}{1000}} \\ &\simeq \frac{1}{51}. \end{aligned}$$

Let us now do this by a method which involves simple counting and little knowledge of probability; we shall simply interpret the probability of a set as a measure of the proportion of elements in that set (that is, the simple interpretation of $\mathrm{P}(A)$ as $|A|/|\Omega|$ which, as we know, holds in finite probability spaces with equiprobable elementary outcomes). Thus, suppose the size of the population is N, and suppose all of them are tested. Then $x = N/1000$ members of the population are ill, and the test marks all of them positive. There are $999N/1000$ healthy members of the population, and the test marks 5% of these positive; that is, there are $y = 0.05 \times 999N/1000$ healthy members of the population tested positive. The proportion of those marked positive who are really sick is therefore given by $x/(x+y$ and, substituting values for x and y gives exactly the same calculations as with Bayes' Theorem.

Note that the answer is not 95%, as one might at first think. This would be the answer to a different question, namely, what is $\mathrm{P}(A^c|B_2) = 1 - \mathrm{P}(A|B_2)$, that is, what is the probability that the test is negative given that the patient is not ill. □

Exercises

1. `Go through Sections 1.8 and 1.10 of [GW].`

2. `Do Exercise 26 in Chapter 1 of [GW].`

Problems

1. Do Problems 5, 6, 7 and 14 in Chapter 1 of [GW].

2. The rules for the game of craps are as follows. The shooter throws a pair of dice. He wins if he obtains 7 or 11 on the first throw and he loses if he obtains 2, 3, 12 on the first throw. Otherwise, he keeps on playing until he either obtains again his first throw (which would be one of 4, 5, 6, 8, 9 or 10), in which case he wins, or he gets a seven, in which case he loses. Let $p_i, 2 \leq i \leq 12$

DENOTE THE PROBABILITY THAT A TOTAL OF i IS OBTAINED WHEN THE TWO DICE ARE THROWN. SHOW THAT THE PROBABILITY THAT THE SHOOTER WINS IS

$$p_7 + p_{11} + \frac{p_4^2}{p_4 + p_7} + \frac{p_5^2}{p_5 + p_7} + \frac{p_6^2}{p_6 + p_7} + \frac{p_8^2}{p_8 + p_7} + \frac{p_9^2}{p_9 + p_7} + \frac{p_{10}^2}{p_{10} + p_7}.$$

SHOW THAT IF THE DICE ARE FAIR THEN THIS PROBABILITY IS APPROXIMATELY EQUAL TO 0.49.

5 The use of recurrence relations: The Gambler's Ruin

The Partition Theorem often enables us to find probabilities by setting up and solving a linear recurrence relation. The most famous problem of this type is the Gambler's Ruin Problem. Here is a variant of this problem.

Two gamblers A and B play a game against each other. They repeatedly flip a coin which comes up heads with probability p and tails with probability $q = 1 - p$. Gambler A starts with Lmn and gambler B starts with Lm$(N - n)$. If tails comes up, A wins Lm1 from B, whereas if heads comes up, B gives Lm1 to A. The game proceeds until one of the two gamblers ends up with Lm0 (is therefore "ruined").

We are interested in the probability that A wins. Since this depends on A's starting capital n, we let this probability be a_n. Clearly $a_0 = 0$ and $a_N = 1$. Let H be the event that the first flip of the coin gives heads. Let us "condition on the result of the first toss": by the Partition Theorem,

$$\mathrm{P}(A \text{ wins }) = \mathrm{P}(A \text{ wins } | H)\mathrm{P}(H) + \mathrm{P}(A \text{ wins } | H^c)\mathrm{P}(H^c).$$

But if H occurs, then A's capital increases from Lmn to Lm$(n + 1)$, therefore game starts again but with A's capital standing at $(n+1)$. Hence, $\mathrm{P}(A \text{ wins } | H) = a_{n+1}$. Similarly, $\mathrm{P}(A \text{ wins } | H^c) = a_{n-1}$. Therefore, substituting into the above equation gives

$$a_n = pa_{n+1} + qa_{n-1} \qquad \text{for } 1 \leq n \leq N - 1.$$

Using standard techniques, you can solve this second order recurrence relation, subject to the conditions $a_0 = 0$ and $a_N = 1$. The solution is

$$a_n = \left\{ \begin{array}{cc} \frac{(q/p)^n - 1}{(q/p)^N - 1} & p \neq q \\ n/N & p = q = \frac{1}{2} \end{array} \right\}.$$

For those who prefer drinking to gambling, the Gambler's Ruin Problem can be reworded as a Random Walk Problem: A drunk stands somewhere between two lamp posts, one on each side of the road. Every second he takes a step (of length 1) to the left or to the right with probabilities p and q respectively. If he

reaches one of the lamp posts he stops there to wait for the pubs to open. He starts his walk a distance n from the right lamp post, and the distance between lamp posts is N. What is the probability that he ends up at the left lamp post?

Exercises

1. Read through Theorems 10C and 10E in [GW]. Do not ''study'' these theorems as pieces of theory which you are expected to memorise and reproduce---just consider them as worked examples. Theorem 10C is just the example we have worked above, and Theorem 10E is the case when B is infinitely rich and can never be ruined (for example, a casino). You might find Theorem 10E more difficult to follow. There is no need for you to go through all the details. It is sufficient for you to understand the setting up of the recurrence relation and, at least, the solution for the case $p \leq q$ (in which case, A is sure to be ruined, as is proved there). Note that in [GW] it is not expected that the reader is already familiar with recurrence relations (a brief discussion is given in their Appendix), therefore in examples involving recurrence relations, the methods used might not be the most straightforward.

Problems

1. A coin is tossed repeatedly. Each time there is a probability p of a head turning up. Let a_n denote the probability that an even number of heads has occurred after n tosses. By conditioning on the result of the first toss, show that

$$a_n = p + (1-2p)a_{n-1}.$$

Hence show that $a_n = \frac{1}{2}[1 + (1-2p)^n]$.

2. Do Exercise 27 in Chapter 1 of [GW].

3. Do Exercise 9 in Chapter 10 of [GW].

6 Random variables: The discrete case

Definition. Let $(\Omega, \mathcal{F}, \mathrm{P})$ be a probability space. A *random variable (r.v.)* on this space is a function $X : \Omega \to \mathbb{R}$ such that, for every $x \in \mathbb{R}$, the set $\{\omega \in \Omega : X(\omega) \leq x\}$ is one of the sets in the collection $\mathcal{F}$. □

The last condition in the definition can, at this stage, be considered to be a minor technicality. You will not go wrong if you simply think of a random variable as a function from Ω to $\mathbb{R}$. Note that the term "variable" is really a misnomer—a random variable is actually a function.

Random variables in general are too difficult to handle, so we impose extra conditions on the definition. The two main types of random variables are the discrete and the continuous ones. We shall study discrete r.v.'s in some detail, and we shall later take a brief look at continuous r.v.'s.

Definition. The *range* or *image* of a r.v X, denoted by $\mathrm{Im}X$, is the set of values which X takes, that is,

$$\mathrm{Im}X = \{x \in \mathbb{R} : X(\omega) = x, \omega \in \Omega\}.$$

Definition. A random variable is said to be *discrete* if its range $X(\Omega)$ is countable, that is, if it only takes on a countable number of values. □

The previous technical requirement would become, for a discrete r.v., that for any x in the range of X, the set $\{\omega \in \Omega : X(\omega) = x\}$ is one of the sets in the collection $\mathcal{F}$.

Note that a random variable whose range is finite is, of course, discrete.

Examples

1. A fair die is thrown. If the numbers 1, 2 or 3 come up I have to pay Lm1, if 5 or 6 comes up, I win Lm2, while if 4 comes up no money changes hands. Here $\Omega = \{1, 2, 3, 4, 5, 6\}$, and $X(1) = X(2) = X(3) = -1$, $X(4) = 0$ and $X(5) = X(6) = 2$. Clearly, X is a discrete r.v. We could ask questions like, "What is the probability that $X \geq 0$?" This would be the probability of the event $A = \{0, 5, 6\}$ in the original probability space. However, we are often only interested in the possible results of the lottery, that is, the possible values of the r.v., and not in the original probability space itself. We would therefore end up working in a different (and sometimes smaller) probability space. Here, for example, the probability space associated with the r.v. X has a sample space with only three elementary events, -1, 0 and 2, with probabilities $\frac{1}{2}$, $\frac{1}{6}$ and $\frac{1}{3}$, respectively.

2. A coin is tossed three times; heads comes up with probability p and tails with probability $q = 1 - p$. Then

$$\Omega = \{HHH, HHT, HTH, HTT, THH, THT, TTH, TTT\}.$$

Suppose we are only interested in the number of heads. We define a r.v. X as follows: $X(HHH) = 3$, $X(HHT) = X(HTH) = X(THH) = 2$, $X(HTT) = X(THT) = X(TTH) = 1$ and $X(TTT) = 0$. The probability space associated with X would then has four elementary events, 0, 1, 2 and 3. The probability, for example, of the event $X = 1$ would be the probability of the event $\{HTT, THT, TTH\}$ in the original space, which is equal to $3pq^2$. □

The following notation was implied in the above example. Let X be a discrete r.v. on a probability space $(\Omega, \mathcal{F}, \mathrm{P})$. Then, the notation $\mathrm{P}(X = x)$ is a short way of writing

$$\mathrm{P}(\{\omega \in \Omega : X(\omega) = x\}).$$

Example

3. Here is an example of a r.v. which is not discrete. This example also illustrates why we have only defined the above notation for discrete r.v.'s. A rod is pivoted freely at one end (not too freely, otherwise is would never stop spinning), and it is spun round in a horizontal plane. The random variable X is defined to be the angle θ which the rod makes with the line of its initial position when it comes to a stop. Clearly X is not discrete, since its range is $[0, 2\pi[$. What is the probability that X is $\pi/3$, for example? We may suppose that all angles are equally likely, therefore we could let this common probability be k. But if $k > 0$, then "summing" probabilities over $[0, 2\pi[$ would give ∞ instead of 1. Therefore k must be 0. This will always be the case with non-discrete r.v.'s

such as X here—the probability that $X = k$ is 0 for any k. The type of question which we should ask here is of the form "What is the probability that X lies between a and b?" The answer to this would be $(b-a)/2\pi$. We shall study in some more detail such r.v.'s when we consider continuous r.v.'s.

Note that even a probability space such as this (the spinning of the rod) with an uncountable number of elementary events, can lead to a discrete r.v. Suppose, for example, a lottery is associated with the spin. If $0 \leq \theta \leq \pi/2$, I win Lm1, if $\pi/2 \leq \theta \leq \pi$, I win Lm2, if $\pi \leq \theta \leq 3\pi/2$, I win Lm1, and if $3\pi/2 \leq \theta \leq 2\pi$, I win Lm4. This gives a random variable Y with range $\{1, 2, 4\}$. The probability $\mathrm{P}(Y = 1)$, for example, would be equal to $\frac{1}{2}$. □

Definition. The *probability mass function* or *mass function* of a discrete r.v. X is a function p_X defined on the range of X such that for any x in the range, $p_X(x)$ equals $\mathrm{P}(X = x)$. □

Therefore the mass function of X gives us the probability that X takes a particular value in its range. Sometimes we drop the suffix inp_X when it is clear that we are referring to the r.v. X. Although p_X is only defined for values in the range of X, we often assume that it is defined for all $x \in \mathbb{R}$, taking the value 0 for $x \notin \mathrm{Im}X$.

Examples

4. In Example 1, $p_X(-1) = 1/2$, $p_X(0) = 1/6$ and $p_X(2) = 1/3$.

5. In Example 2, $p_X(0) = q^3$, $p_X(1) = 3pq^3$, $p_X(2) = 3p^2q$, $p_X(3) = p^3$.

6. In Example 3, $p_Y(1) = 1/2$, $p_Y(2) = 1/4$, $p_Y(4) = 1/4$. □

Note the following elementary but important property of mass functions:

$$\begin{aligned}\sum_{x \in \mathrm{Im}X} p_X(x) &= \sum_{x \in \mathrm{Im}X} \mathrm{P}(\{\omega \in \Omega : X(\omega) = x\}) \\ &= \mathrm{P}\Big(\bigcup_{x \in \mathrm{Im}X} \{\omega \in \Omega : X(\omega) = x\}\Big) \\ &= \mathrm{P}(\Omega) = 1.\end{aligned}$$

That is, a mass function sums to 1 over all values in the range of X. Sometimes, this is written as

$$\sum_{x \in \mathbb{R}} p_X(x) = 1.$$

This, of course, does not mean that we are summing over all values in $\mathbb{R}$—those values not in the range of X simply give zero.

Exercises

1. `Read through Section 2.1 in [GW].`

2. `Do Exercise 5 in Chapter 2 of [GW].`

We have noted that when given a discrete r.v. on a probability space, we can often "forget" the original space and work with the r.v. and the probabilities given by its mass function. In fact, in most situations we actually start at this point, that is, we are just given a discrete r.v. and its mass function—we know

that, if required, we can go back and constructed the original probability space, but this is often not necessary.

Some mass functions arise so often and are so important that we give them names and we study their properties in particular detail. Here are a few of them. For each, we give the range of the r.v. X, its mass function and an example of a situation in which that type of r.v. arises.

Bernoulli distribution with parameter p

The range of X is $\{0,1\}$ and $p_X(0) = p, p_X(1) = q$ with $p + q = 1$.

Such a random variable occurs when considering an experiment with two possible outcomes. For example, a coin is tossed, the probability of tails is p, and here X would take the value 0, while the probability of heads is $q = 1 - p$, and in this case X would take the value 1.

Binomial distribution with parameters n, p

The range of X is $\{0, 1, 2, \ldots, n\}$ and, for $k \in \mathrm{Im}X$,

$$p_X(k) = \binom{n}{k} p^k q^{n-k}$$

where $q = 1 - p$.

It is clear that this does, in fact, satisfy the basic property of probability mass functions, namely that the sum of $p_X(k)$ over all the range of X is 1, because

$$\sum_{k=0}^{n} p_X(k) = \sum_{k=0}^{n} \binom{n}{k} p^k q^{n-k} = (p+q)^n = 1.$$

Binomial random variables arise in situations where we want to find the probability of obtaining k "successes" out of n "trials". For example, suppose a coin is tossed n times, and, on each occasion, the probability of heads is p while q is the probability of tails. Let X be the number of heads obtained. What is the probability that k heads are obtained? There are $\binom{n}{k}$ ways in which this event can happen, and for each the probability is $p^k q^{n-k}$, by independence of the tosses. Also, these $\binom{n}{k}$ events are disjoint, therefore the probability that one of these occurs is the sum of their probabilities, therefore $\mathrm{P}(X = k) = \binom{n}{k} p^k q^{n-k}$, that is, X has the binomial distribution. (See Example 2; there we had the binomial distribution with $n = 3$.)

We often write $b(k; n, p)$ for $\binom{n}{k} p^k q^{n-k}$.

Geometric distribution with parameter $p > 0$

The range of X is $\mathbb{N} = \{1, 2, 3 \ldots\}$, and, for any k in this range, $p_X(k) = pq^{k-1}$, where, as usual, $q = 1 - p$. Again we check that the probabilities sum to 1:

$$\sum_{k=1}^{\infty} pq^{k-1} = p.\frac{1}{1-q} = \frac{p}{p} = 1.$$

Geometric random variables arise in situations where we want to find the probability of having to carry out k "trials" until (and including) the first "success". Suppose the same coin as in the previous example is tossed until the first occurrence of heads (we have already encountered this situation). Each outcome

can be represented by T^iH. Define the random variable X by $X(T^iH) = i+1$, that is, X is the number of tosses up to and including the first heads. Then $\mathrm{P}(X = k) = \mathrm{P}(T^{k-1}H) = q^{k-1}p$, by independence of tosses.

Poisson distribution with parameter $\lambda > 0$

The range of X is $\{0, 1, 2, 3 \dots\}$ and, for any k in this range, $p_X(k) = \frac{1}{k!}\lambda^k e^{-\lambda}$. Again the usual check:

$$\sum_{k=0}^{\infty} p_X(k) = e^{-\lambda} \sum_{k=0}^{\infty} \frac{\lambda^k}{k!} = e^{-\lambda}.e^{\lambda} = 1.$$

It is not so easy in an elementary course in probability to give examples where the Poisson distribution arises naturally. One very important use of this distribution can however be explained, although we shall not do this completely rigorously. Calculating probabilities involving the binomial distribution can be quite awkward for large values of n, and a good approximation would therefore be very helpful. In fact, when n is large and p is small so that np is of "reasonable" size, we can let $\lambda = np$ and use the Poisson distribution with parameter λ to approximate the binomial distribution with parameters n, p — that is, $\binom{n}{k}p^k(1-p)^{n-k}$ is approximately equal to $\frac{\lambda^k}{k!}.e^{-\lambda}$, where $\lambda = np$, if n is large and p is small. The following is a sketch of the proof of this.

Let k be fixed. and let $\lambda = np$. Then,

$$\begin{aligned}
\binom{n}{k}p^k(1-p)^{n-k} &= \frac{n!}{k!(n-k)!}.\left(\frac{\lambda}{n}\right)^k.\left(1-\frac{\lambda}{n}\right)^n.\left(1-\frac{\lambda}{n}\right)^{-k} \\
&\simeq \frac{n^k}{k!}.\frac{\lambda^k}{n^k}.\left(1-\frac{\lambda}{n}\right)^n \\
&\simeq \frac{\lambda^k}{k!}.e^{-\lambda}.
\end{aligned}$$

Here, in the first approximating step, we used the approximation $n(n-1)\dots(n-k+1) \simeq n^k$, and the next step we used the result $\lim_{n\to\infty}\left(1+\frac{a}{n}\right) = e^a$.

The graphs in Figure 2 illustrate how the binomial distribution tends to the Poisson. (In these diagrams, $\frac{\lambda^k}{k!}.e^{-\lambda}$ is denoted by $p(k;\lambda)$.)

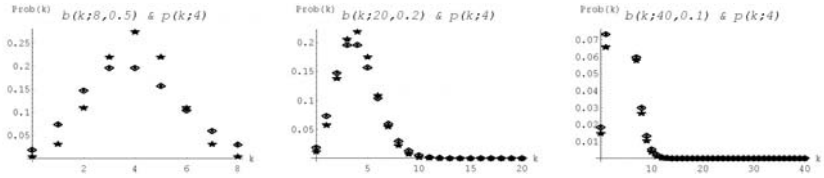

Figure 2: The Poisson approximation to the binomial distribution: Stars, diamonds denote probabilities calculated by the binomial, Poisson respectively

Example

7. A book contains $10,000$ characters, and it is estimated that any character could be misprinted with probability $1/500$. The probability that there are k

misprints in the book is therefore given by $b\left(k;10000,\frac{1}{500}\right) = \binom{10000}{k}\left(\frac{1}{500}\right)^k\left(\frac{499}{500}\right)^{n-k}$. However, a good approximation to this is given by the Poisson distribution: Let $\lambda = np = 10000.\frac{1}{500} = 20$; then the above probability is approximately given by $\frac{1}{k!}.20^k.e^{-20}$. □

Exercises

3. `Read through Section 2.2 in [GW].`

4. `Do Exercises 5 and 6 in Chapter 2 of [GW].`

5. `This exercise has nothing to do with probability really, but it is a good question. Calculate, to two decimal figures, the value of`

$$(1 + 5 \times 10^{-34})^{4\times 10^{33}}.$$

Problems

1. Airlines find that each passenger who reserves a seat fails to turn up with probability 0.1, independently of the other passengers. So Airline W&P (On a Wing and a Prayer) always sells ten tickets for its 9-seater aeroplane, while Airline TAL (Twice as Large) always sells twenty tickets for its 18-seater aeroplane. Which airline is more likely to be overbooked (that is, more passengers turn up than there are seats)? Use the binomial distribution with $p = 0.1$ and $n = 10$ or $n = 20$ and try also the Poisson distribution with $\lambda = 1$ or $\lambda = 2$.

2. Let X be a geometric random variable. Show that $\mathrm{P}(X > m+n|X > m) = \mathrm{P}(X > n)$. This property is usually called the "lack of memory" property. Can you see the connection with Example 3 in Section 3?

7 Expectation and variance

It is often very useful to have a few parameters which summarise the behaviour of a random variable. The notion of the "average value" of a random variable is well-known. The following definition captures the idea of what we normally understand by average.

Definition. Let X be a discrete r.v. on the probability space $(\Omega, \mathcal{F}, \mathrm{P})$. The *expectation* of X (also called the *expected value* or the *mean* of X), denoted by $\mathrm{E}(X)$, is defined by

$$\mathrm{E}(X) = \sum_{x\in \mathrm{Im}X} x p_X(x)$$

provided the sum converges absolutely. □

Examples

1. The range of the r.v. X is $\{2, 7, 12, 43\}$. The probabilities $p_X(x) = \mathrm{P}(X = x)$ are all equal to $\frac{1}{4}$. Then $\mathrm{E}(X) = 2 \times \frac{1}{4} + 7 \times \frac{1}{4} + 12 \times \frac{1}{4} + 43 \times \frac{1}{4}$ and this simply equals the average of all the values in the range of X.

2. The random variable X has the same range as in the previous example, but the probabilities $p_X(x)$ are $\frac{1}{2}, \frac{1}{12}, \frac{1}{6}, \frac{1}{12}$ for $x = 2, 7, 12, 43$, respectively. Then $\mathrm{E}(X) = 2 \times \frac{1}{2} + 7 \times \frac{1}{12} + 12 \times \frac{1}{6} + 43 \times \frac{1}{12}$. This is a "weighted average" of all the values in the range of X.

3. There do exist random variables which do not have an expected value. For example, let the $\mathrm{Im}X = \mathbb{N}$ and let $p_X(k) = 12/(k\pi)^2$. Then $\mathrm{E}(X)$ would involve the summation $\sum \frac{1}{k}$ which does not converge. □

You should be able to work out the expected value of X when X is a Bernoulli, binomial, geometric or Poisson random variable. These are given as problems below, and they are quite easy to obtain. You would need to sum series such as

$$\sum_{k=0}^{n} \binom{n}{k} k p^k q^{n-k} \qquad \sum_{k=1}^{\infty} k q^{k-1} \qquad \sum_{k=0}^{\infty} k \frac{\lambda^k}{k!} e^{-\lambda}.$$

There is a useful technique for working out such sums, and we describe it for the binomial case. Start with

$$\sum_{k=0}^{n} \binom{n}{k} p^k q^{n-k} = (p+q)^n$$

and differentiate with respect to p to give

$$\sum_{k=0}^{n} \binom{n}{k} k p^{k-1} q^{n-k} = n(p+q)^{n-1}.$$

Multiply both sides by p to give

$$\sum_{k=0}^{n} \binom{n}{k} k p^k q^{n-k} = np(p+q)^{n-1}$$

and *now* put $p + q = 1$ to give

$$\sum_{k=0}^{n} \binom{n}{k} k p^k q^{n-k} = np.$$

Problems

1. Let X be a Bernoulli r.v. with parameter p. Then $\mathrm{E}(X) = p$.

2. Let X be a binomial r.v. with parameters n, p. Then $\mathrm{E}(X) = np$.

3. Let X be a geometric r.v. with parameter p. Then $\mathrm{E}(X) = \frac{1}{p}$.

4. Let X be a Poisson r.v. with parameter λ. Then $\mathrm{E}(X) = \lambda$.

5. Do Problem 6 in Chapter 2 of [GW]. This deals with the unbiased die of Problem 1 in Section 3 of these notes. In that problem you have found $\mathrm{P}(N > k)$ (where N is as defined in *this* problem). Use this result.

As equally important as the expectation of a r.v. is its variance. While the expectation is an indication of the "centre" of the distribution of the r.v., the variance is an indication of the "spread" of the distribution about the mean. It is defined as follows.

Definition. Let X be a r.v. on a probability space $(\Omega, \mathcal{F}, \mathrm{P})$ such that the expected value of X exists, and let $\mu = \mathrm{E}(X)$. Then the *variance* of X, denoted by $\mathrm{var}(X)$ is defined by

$$\mathrm{var}(X) = \sum_{x \in \mathrm{Im}X} (x - \mu)^2 p_X(x)$$

provided this sum converges. □

Examples

4. Let X be the r.v. of Example 1. Then $\mu = 16$ and $\mathrm{var}(X)$ equals

$$(2-16)^2 \times \frac{1}{4} + (7 - 16^2) \times \frac{1}{4} + (12-16)^2 \times \frac{1}{4} + (43-16)^2 \times \frac{1}{4}.$$

5. Let X be the r.v. of Example 2. Then $\mu = 7.16$ and $\mathrm{var}(X)$ equals

$$(2-7.16)^2 \times \frac{1}{2} + (7 - 7.16^2) \times \frac{1}{12} + (12-7.16)^2 \times \frac{1}{6} + (43-7.16)^2 \times \frac{1}{12}.$$

□

Can use see why the differences between the possible values of X and μ are squared in the definition of $\mathrm{var}(X)$? What would $\mathrm{var}(X)$ be equal to if these differences were not squared?

At this point we should mention the idea of functions of random variables. Suppose X is a discrete r.v. and $g : \mathbb{R} \to \mathbb{R}$. Then $Y = g(X)$ (composition of function or a function of a function) is also a r.v. Knowing the probability mass function and expected value of X we could ask what the mass function and expected value of Y is.

Example

6. Let X take the values $-2, 2, 3$ with probabilities $\frac{1}{4}, \frac{1}{8}, \frac{5}{8}$, respectively. Let $Y = X^2$. What is $\mathrm{E}(Y)$? First we must find the range and the mass function of Y. Y takes the values $4, 9$ with probabilities $\frac{3}{8}, \frac{5}{8}$ respectively. Then $\mathrm{E}(Y) = 4(3/8) + 9(5/8) = 57/8$. □

However, there is a result which enables us to find $\mathrm{E}(Y)$ by working directly with the mass function of X without having to work it out for Y. This result (sometimes called the Law of the Unconscious Statistician) is proved as Theorem 2B in [GW]. We simply state it here.

Theorem. *If X is a discrete r.v. and $g : \mathbb{R} \to \mathbb{R}$ then*

$$E(g(X)) = \sum_{x \in Im X} g(x) P(X = x).$$

□

Example

7. Consider the r.v.'s X and Y of the previous example. Using the above theorem, $\mathrm{E}(Y)$ can be calculated as $(-2)^2(1/4)+(2)^2(1/8)+(3^2)(5/8) = (57/8)$. □

We emphasise one particularly important case of this theorem, namely when $Y = aX$, where a is a constant. The theorem then gives that $\mathrm{E}(Y) = a\mathrm{E}(X)$, but this is very easy to prove directly from the definition of expected value. When $Y = aX$ it is also very easy to give $\mathrm{var}(Y)$ in terms of $\mathrm{var}(X)$; in this case, $\mathrm{var}(Y) = a^2\mathrm{var}(X)$. These relationships between the expectation and variance of Y and X when Y is a linear function of X will be used in a later section.

Now back to the discussion of variance. Looking at the formula and using the theorem we see that $\mathrm{var}(X)$ can be seen as $\mathrm{E}([x-\mu]^2)$. The formula is not always the most convenient way to calculate variances. It is easy to show (see Section 2.4 of [GW]) that $\mathrm{var}(X)$, which is defined as $\sum_{x\in\mathrm{Im}X}(x-\mu)^2 p_X(x)$, turns out to be equal to

$$\sum_{x\in\mathrm{Im}X} x^2 p_X(x) - \mu^2$$

and this formula is often easier to apply. Note that, with this formula, $\mathrm{var}(X)$ can be seen as

$$\mathrm{E}(X^2) - \mu^2 = \mathrm{E}(X^2) - \mathrm{E}(X)^2$$

.

You should be able to work out the variance of X when X is a Bernoulli, binomial, geometric or Poisson random variable. These are given as problems below, and they are not too difficult to obtain. You would now need to sum series such as

$$\sum_{k=0}^{n}\binom{n}{k}k^2p^kq^{n-k} \qquad \sum_{k=1}^{\infty}k^2q^{k-1} \qquad \sum_{k=0}^{\infty}k^2\frac{\lambda^k}{k!}e^{-\lambda}.$$

The technique which we described above also works here, but this time you have to differentiate twice. Again we describe it for the binomial case. Start with

$$\sum_{k=0}^{n}\binom{n}{k}p^kq^{n-k} = (p+q)^n$$

and differentiate twice with respect to p to give

$$\sum_{k=0}^{n}\binom{n}{k}k(k-1)p^{k-2}q^{n-k} = n(n-1)(p+q)^{n-2}.$$

Multiply both sides by p^2 and separate the lefthand side into two summations, one involving k^2 and the other involving k. The latter has been previously found, so the one we require (involving k^2) can also be found.

Problems

6. Let X be a Bernoulli r.v. with parameter p. Then $\text{var}(X) = pq = p(1-p)$.

7. Let X be a binomial r.v. with parameters n, p. Then $\text{var}(X) = npq = np(1-p)$.

8. Let X be a geometric r.v. with parameter p. Then $\text{var}(X) = \frac{q}{p^2} = \frac{1-p}{p^2}$.

9. Let X be a Poisson r.v. with parameter λ. Then $\text{var}(X) = \lambda$.

Exercise

1. `Read through Sections 2.3 and 2.4 of [GW].`

8 Conditional expectation and the partition theorem

Just as we had conditional probabilities we can discuss conditional expectation. Basically, if we are given that the event B has occurred, then the calculation of the expected value of X given this condition will now be based only on those elementary events in B and not on all of Ω — that is, the probabilities $\mathrm{P}(X = x)$ will now be replaced by the probabilities

$$\begin{aligned} \mathrm{P}(X = x|B) &= \frac{\mathrm{P}(\{\omega \in \Omega : X(\omega) = x\} \cap B)}{\mathrm{P}(B)} \\ &= \frac{\mathrm{P}((X = x) \cap B)}{B} = \frac{\mathrm{P}(\{\omega \in B : X(\omega) = x\})}{\mathrm{P}(B)}. \end{aligned}$$

The definition of conditional expectation is therefore as follows.

Definition. Let X be a discrete r.v. on a probability space $(\Omega, \mathcal{F}, \mathrm{P})$ and let B be an event such that $\mathrm{P}(B) > 0$. The *conditional expectation of X given B*, denoted by $\mathrm{E}(X|B)$, is defined by

$$\mathrm{E}(X|B) = \sum_{x \in \mathrm{Im}X} x\mathrm{P}(X = x|B)$$

provided this sum converges absolutely. □

We have seen how useful the partition theorem is for calculating probabilities. Similar techniques can be used to calculate expectations. To do this we need a version of the partition theorem for expected values. This theorem is proved in Section 2.5 of [GW].

Theorem. The Partition Theorem for Expectations. *If X is a discrete r.v. and $\{B_1, B_2, \dots\}$ is a partition of the sample space such that $P(B_i) > 0$ for each i then*

$$E(X) = \sum_i E(X|B_i)P(B_i)$$

provided this sum converges absolutely. □

Using this theorem often leads to situations involving the setting up and solution of recurrence relations, as was the case for probabilities.

Exercises

1. `Read through Section 2.5 of [GW].`

2. `Go through Theorem 10D of [GW].`

Problems

1. Do Problem 2 in Chapter 2 of [GW].

2. Let N be the number of tosses of a coin up to and including the appearance of the first heads. Let the probability that a toss of the coin gives heads be p. By conditioning on the result of the first toss (and without using any knowledge about the geometric distribution), show that $\mathrm{E}(N) = \frac{1}{p}$.

3. Consider the two shooters Jack and Jill whom we first met in Problem 2 of Section 3. Given that Jack wins, show that the expected number of shots which he fires is $1/(1 - q_0 q_1)$.

3. An urn contains b blue balls and r red balls. Balls are removed at random until the first blue ball is drawn. Show that the expected number of balls drawn is $(b + r + 1)/(b + 1)$. [Let m_r be the expected number when there are r red balls. Condition on the colour of the first ball drawn to obtain that $m_r = 1 + \frac{r}{b+r}.m_{r-1}$.]

9 More than one random variable

Let X and Y be two discrete r.v.'s on the probability space $(\Omega, \mathcal{F}, \mathrm{P})$. The two mass function p_X and p_Y give separately the probabilities that X or Y take values in their ranges. However, we to have the probabilities that X takes the value x *and* Y takes the value y for all pairs $x \in \mathrm{Im}X, y \in \mathrm{Im}Y$. This would give a joint mass function $p_{X,Y}$, which is formally defined as follows.

Definition. Let X, Y be discrete r.v.'s on a probability space $(\Omega, \mathcal{F}, \mathrm{P})$. The *joint mass function* of X, Y is a function $p_{X,Y} : \mathrm{Im}X \times \mathrm{Im}Y \to [0, 1]$ defined by

$$p_{X,Y} = \mathrm{P}(\{\omega \in \Omega : X(\omega) = x \text{ and } Y(\omega) = y\}).$$

□

A shorthand way of writing $\mathrm{P}(\{\omega \in \Omega : X(\omega) = x \text{ and } Y(\omega) = y\})$ is $\mathrm{P}(X = x, Y = y)$.

Example

1. See the example in Section 3.1 of [GW]. □

The idea of joint mass functions leads to the notion of independent random variables. Independence of r.v.'s is closely related to independence of events as we have already defined them—in fact, X and Y are independent r.v.'s if the two events $(X = x)$ and $(Y = y)$ are independent for all $x \in \mathrm{Im}X$ and $y \in \mathrm{Im}Y$. The definition which captures all this is the following.

Definition. Two r.v.'s X, Y on a probability space are *independent* if

$$\mathrm{P}(X = x, Y = y) = \mathrm{P}(X = x)\mathrm{P}(Y = y)$$

for all $x \in \mathrm{Im}X$ and $y \in \mathrm{Im}Y$. □

Very often, a problem involving one r.v. can be simplified by writing the r.v. as a sum of two or more simpler r.v.'s. It is therefore very useful if we have results which help us to work with sums of r.v.'s. One such result is about expectations is of great importance because it is surprisingly very simple. It says that the expectation of the sum of two r.v.'s is the sum of the two expectations. This property is called the *linearity of expectation.*

Theorem. Linearity of Expectation. *Let X and Y be two r.v.'s on a probability space $(\Omega, \mathcal{F}, P)$ and let a, b be constants. Then $E(aX + bY) = aE(X) + bE(Y)$.* □

In [GW] this result is proved after Theorem 3A. It is proved as a special case of Theorem 3A which is a generalisation of Theorem 2B for two variables. However, a simpler direct proof of the linearity of expectation can be given. You do not need to do this, but you should be able to use the result in problems such as the example below and the problem following. Although we have given this result in terms of the sum of two r.v.'s, it can clearly be generalised to any finite number of r.v.'s by induction. Also, the linearity of expectation is very easily extended to the following useful form:

$$\mathrm{E}(aX + bY) = a\mathrm{E}(X) + b\mathrm{E}(Y)$$

where a and b are constants.

Example

2. Consider the following situation which is closely connected to the problem of derangements. A particularly inefficient secretary puts n letters at random into their envelopes. What is the expected number of letters which end up in their correct envelope?

The sample space Ω here consists of all $n!$ permutations of the set $\{1, 2, \ldots, n\}$ representing the way the letters are distributed amongst the envelopes. All permutations are equally likely. Let the r.v. X be defined as follows: Let π be a permutation, then $X(\pi)$ equals the number of fixed points of π, that is, the number of letters put into the correct envelope by the permutation π. For $i = 1 \ldots n$, let X_i be defined as follows: Given the permutation π, let $X_i(\pi)$ be equal to the number of times the element i is in the ith position (that is, the number of times the permutation π puts the ith letter in the correct envelope). Of course, X_i only takes on the values 0 and 1, that is, it is a Bernoulli variable. The probability $\mathrm{P}(X_i = 1)$ equals $\frac{(n-1)!}{n!} = \frac{1}{n}$, therefore $\mathrm{E}(X_i) = \frac{1}{n}$. But $X = X_1 + X_2 + \cdots + X_n$, therefore $\mathrm{E}(X) = 1$. □

Problem

1. THIS IS CALLED THE *coupon collecting problem.* EACH PACKET OF CEREAL CONTAINS ONE OF c POSSIBLE TYPES OF COUPON, AND EACH TIME YOU BUY A PACKET IT IS EQUALLY LIKELY TO CONTAIN A COUPON OF ANY ONE OF THE c TYPES. LET X BE THE NUMBER OF COUPONS YOU NEED TO BUY IN ORDER TO COLLECT A COMPLETE SET OF COUPONS. THE AIM OF THIS PROBLEM IS TO FIND $\mathrm{E}(X)$.

First let $X_i, i = 0, 1, \ldots, c-1$ denote the additional number of packets you need to buy in order to get any new coupon after i coupons have already been collected. Show that X_i is geometric with parameter $(c-i)/c$, and hence find $\mathrm{E}(X_i)$.

Now, using the fact that $X = X_0 + X_1 + \cdots + X_{c-1}$, show that $\mathrm{E}(X) = c\left(1 + \frac{1}{2} + \cdots + \frac{1}{c}\right)$.

Exercise

1. `Read quickly through Section 2.3 and Chapter 3 of [GW].`

10 Chebyshov's Inequality and the Law of Averages

Our intuitive understanding of probability is firmly rooted in what we would expect to happen in the long run if a random event were repeated very often. For example, suppose a fair die is cast six hundred times. We would certainly not expect that exactly one hundred of the throws give the number 4, but we would be very surprised if the number of 4's appearing were to be very different from one hundred, say five hundred, or twenty. This would be even more so if the die has been cast six thousand times—we would then more certainly expect that the number of 4's appearing be close to one thousand. Is there any real justification for how we believe that probabilities should behave? One of the supreme justifications of the mathematical machinery that we have set up is that we can prove such statements as mathematical theorems. We shall only prove here the very simplest of these results, called the Law of Averages or the Weak Law of Large Numbers.

First we prove an equally important result called Chebyshov's Inequality. It shows us that the definitions we gave for expectation and variance are very well justified. They do measure the "centrality" and "dispersion" of a r.v., in the sense that the smaller the variance is, the less likely is the r.v. to take values very different from the mean.

Theorem. Chebyshov's Inequality. *Let X be a discrete r.v. on a probability space $(\Omega, \mathcal{F}, E)$, and let $\mu = E(X)$ and $\sigma^2 = var(X)$. Then, for all $a > 0$,*

$$P(|X - \mu| > h\sigma) < \frac{1}{h^2}.$$

Proof. For simplicity we shall prove this for the case when Ω is countable.

Then

$$\begin{aligned}\sigma^2 &= \sum_{\omega\in\Omega}(X(\omega)-\mu)^2\mathrm{P}(\omega)\\ &\geq \sum_{\omega:[X(\omega)-\mu]^2>h^2\sigma^2}(X(\omega)-\mu)^2\mathrm{P}(\omega)\\ &> \sum_{\omega:[X(\omega)-\mu]^2>h^2\sigma^2}h^2\sigma^2\mathrm{P}(\omega)\\ &= h^2\sigma^2\mathrm{P}[(X-\mu)^2>h^2\sigma^2].\end{aligned}$$

Dividing by $h^2\sigma^2$ gives $\mathrm{P}[(X-\mu)^2 > h^2\sigma^2] < 1/h^2$. But the condition $(X-\mu)^2 > h^2\sigma^2$ is equivalent to $|X-\mu| > h\sigma$, and this finishes the proof. □

Example

1. The probability that X lies within 2σ of μ is at least 0.75 and the probability that it lies within 10σ of μ is at least 0.99. □

When the distribution of X is known, it is generally possible to obtain sharper bounds than those given by Chebyshov's Inequality. However, this inequality is remarkable because it does not depend on the distribution of X. It simply requires that the expectation and variance of X exist.

Theorem. The Law of Averages. *Let X be a binomial r.v. with parameters n, p. Let $Y = X/n$, that is, Y denotes the average number of "successes" out of the n trials. Let $a > 0$. Then, given any $\epsilon > 0$ there exists N such that, for all $n > N$,*

$$P(|Y-p| > a) < \epsilon,$$

that is, $\lim_{n\to\infty} P(|Y-p| > a) = 0$.

Proof. Note that $\mathrm{E}(X) = np$ and $\mathrm{var}(X) = npq$. Therefore $\mathrm{E}(Y) = p$ and $\mathrm{var}(Y) = pq/n^2$. We shall apply Chebyshov's Inequality to Y with $h = a/\sigma$, $\mu = p$ and $\sigma^2 = pq/n$. Let $N > pq/a^2\epsilon$ and let $n > N$. Then,

$$\mathrm{P}(|Y-p| > a) < \frac{\sigma^2}{a^2} = \frac{pq}{a^2 n} < \frac{pq}{a^2 N} < \epsilon.$$

□

Exercises

1. Read quickly through Sections 8.1 and 8.2 of [GW], but do not worry too much if you do not understand these two sections well.

2. A poor understanding of the Law of Averages sometimes leads to the following type of misconception: A gambler is betting on the outcomes of successive tosses of a fair coin. He notices that the first seven tosses are all heads. Since the Law of Averages guarantees that, in the long run, the number of heads and tails should even out, he reckons that on the next throw tails is more likely to come up, or, at least, that the next tosses should show more tails than heads in order to make up for the initial run of heads. Why is this reasoning incorrect?

We might not make such false judgements when predicting outcomes of tosses of coins, but one can think of various more complex situations in everyday life when we reason in this fashion.

Problem

THERE ARE NO PROBLEMS FOR THIS SECTION. TAKE A REST.

10. Continuous random variables

We shall finally take a brief look at continuous r.v.'s. We have already seen that understanding the behaviour of non-discrete r.v.'s can be very difficult. In particular, it is useless to work with a probability mass function since this could take on the value 0 for all elements in the sample space. We shall therefore define another function which measures probabilities for r.v.'s.

Definition. Let X be a r.v. defined on the probability space $(\Omega, \mathcal{F}, \mathrm{P})$. The *distribution function* of X is the function $F_X : \mathbb{R} \to [0,1]$ defined by

$$F_X(x) = \mathrm{P}(X \leq x) = \mathrm{P}(\{\omega \in \Omega : X(\omega) \leq x\}).$$

□

A continuous r.v. is then defined to be one whose distribution function can be written as an integral.

Definition. The r.v. X is said to be a *continuous random variable* if there is a non-negative function f_X such that

$$F_X(x) = \int_{-\infty}^{x} f_X(u)du.$$

The function f_X is called the *(probability) density function* of X. □

We shall not go into great detail on continuous r.v.'s. A few important types of continuous r.v.'s are given in Section 5.4 of [GW], and you should at least go through this section in detail.

The expected value of a continuous r.v. X is defined by

$$\mathrm{E}(X) = \int_{-\infty}^{\infty} x f_X(x)dx$$

provided the integral converges absolutely. The variance is defined by

$$\mathrm{var}(X) = \int_{-\infty}^{\infty} (x-\mu)^2 f_X(x)dx = \int_{-\infty}^{\infty} x^2 f_X(x)dx - \mu^2$$

provided the integral converges.

Exercise

1. Read through Chapter 5 of [GW], particularly Section 5.4.

Problems

1. FROM A POINT ON THE CIRCUMFERENCE OF A CIRCLE OF RADIUS a, A CHORD IS DRAWN IN A RANDOM DIRECTION. SHOW THAT THE MEAN LENGTH OF THE CHORD IS $4a/\pi$ AND THE VARIANCE OF THE LENGTH IS $2a^2(1 - \frac{8}{\pi^2})$. SHOW ALSO THAT $\mathrm{P}(l > a\sqrt{3}) = 1/3$.

2. A chord of a circle of radius a is drawn parallel to a given straight line. All distances from the centre of the circle are equally likely. Show that the mean of the length of the chord is $\pi a/2$ and its variance is $(8a^2/3) - (\pi^2 a^2/4)$.

3. The quadratic equation $x^2 - ax + b = 0$ is known to have two real distinct roots α, β $(\alpha > \beta)$. The coefficient a is known but the coefficient b is a positive r.v. with uniform distribution in the permissible range of values. Find the mean and the variance of α and β.