Notes on Probability

Peter J. Cameron

Preface

Here are the course lecture notes for the course MAS108, Probability I, at Queen Mary, University of London, taken by most Mathematics students and some others in the first semester.

The description of the course is as follows:

> This course introduces the basic notions of probability theory and develops them to the stage where one can begin to use probabilistic ideas in statistical inference and modelling, and the study of stochastic processes. Probability axioms. Conditional probability and independence. Discrete random variables and their distributions. Continuous distributions. Joint distributions. Independence. Expectations. Mean, variance, covariance, correlation. Limiting distributions.

The syllabus is as follows:

1. Basic notions of probability. Sample spaces, events, relative frequency, probability axioms.
2. Finite sample spaces. Methods of enumeration. Combinatorial probability.
3. Conditional probability. Theorem of total probability. Bayes theorem.
4. Independence of two events. Mutual independence of n events. Sampling with and without replacement.
5. Random variables. Univariate distributions - discrete, continuous, mixed. Standard distributions - hypergeometric, binomial, geometric, Poisson, uniform, normal, exponential. Probability mass function, density function, distribution function. Probabilities of events in terms of random variables.
6. Transformations of a single random variable. Mean, variance, median, quantiles.
7. Joint distribution of two random variables. Marginal and conditional distributions. Independence.

8. Covariance, correlation. Means and variances of linear functions of random variables.

9. Limiting distributions in the Binomial case.

These course notes explain the naterial in the syllabus. They have been "field-tested" on the class of 2000. Many of the examples are taken from the course homework sheets or past exam papers.

Set books The notes cover only material in the Probability I course. The textbooks listed below will be useful for other courses on probability and statistics. You need *at most one* of the three textbooks listed below, but you will need the statistical tables.

- *Probability and Statistics for Engineering and the Sciences* by Jay L. Devore (fifth edition), published by Wadsworth.

Chapters 2–5 of this book are very close to the material in the notes, both in order and notation. However, the lectures go into more detail at several points, especially proofs. If you find the course difficult then you are advised to buy this book, read the corresponding sections straight after the lectures, and do extra exercises from it.

Other books which you can use instead are:

- *Probability and Statistics in Engineering and Management Science* by W. W. Hines and D. C. Montgomery, published by Wiley, Chapters 2–8.

- *Mathematical Statistics and Data Analysis* by John A. Rice, published by Wadsworth, Chapters 1–4.

You should also buy a copy of

- *New Cambridge Statistical Tables* by D. V. Lindley and W. F. Scott, published by Cambridge University Press.

You need to become familiar with the tables in this book, which will be provided for you in examinations. All of these books will also be useful to you in the courses Statistics I and Statistical Inference.

The next book is not compulsory but introduces the ideas in a friendly way:

- *Taking Chances: Winning with Probability*, by John Haigh, published by Oxford University Press.

Web resources Course material for the MAS108 course is kept on the Web at the address

`http://www.maths.qmw.ac.uk/~pjc/MAS108/`

This includes a preliminary version of these notes, together with coursework sheets, test and past exam papers, and some solutions.

Other web pages of interest include

`http://www.dartmouth.edu/~chance/teaching_aids/books_articles/probability_book/pdf.html`

A textbook *Introduction to Probability*, by Charles M. Grinstead and J. Laurie Snell, available free, with many exercises.

`http://www.math.uah.edu/stat/`

The Virtual Laboratories in Probability and Statistics, a set of web-based resources for students and teachers of probability and statistics, where you can run simulations etc.

`http://www.newton.cam.ac.uk/wmy2kposters/july/`

The Birthday Paradox (poster in the London Underground, July 2000).

`http://www.combinatorics.org/Surveys/ds5/VennEJC.html`

An article on Venn diagrams by Frank Ruskey, with history and many nice pictures.

Web pages for other Queen Mary maths courses can be found from the on-line version of the Maths Undergraduate Handbook.

Peter J. Cameron
December 2000

Contents

1 Basic ideas **1**

1.1 Sample space, events 1
1.2 What is probability? 3
1.3 Kolmogorov's Axioms 3
1.4 Proving things from the axioms 4
1.5 Inclusion-Exclusion Principle 6
1.6 Other results about sets 7
1.7 Sampling 8
1.8 Stopping rules 12
1.9 Questionnaire results 13
1.10 Independence 14
1.11 Mutual independence 16
1.12 Properties of independence 17
1.13 Worked examples 20

2 Conditional probability **23**

2.1 What is conditional probability? 23
2.2 Genetics 25
2.3 The Theorem of Total Probability 26
2.4 Sampling revisited 28
2.5 Bayes' Theorem 29
2.6 Iterated conditional probability 31
2.7 Worked examples 34

3 Random variables **39**

3.1 What are random variables? 39
3.2 Probability mass function 40
3.3 Expected value and variance 41
3.4 Joint p.m.f. of two random variables 43
3.5 Some discrete random variables 47
3.6 Continuous random variables 55

3.7 Median, quartiles, percentiles 57
3.8 Some continuous random variables 58
3.9 On using tables 61
3.10 Worked examples 63

4 More on joint distribution **67**
4.1 Covariance and correlation 67
4.2 Conditional random variables 70
4.3 Joint distribution of continuous r.v.s 73
4.4 Transformation of random variables 74
4.5 Worked examples 77

A Mathematical notation **79**

B Probability and random variables **83**

Chapter 1

Basic ideas

In this chapter, we don't really answer the question 'What is probability?' Nobody has a really good answer to this question. We take a mathematical approach, writing down some basic axioms which probability must satisfy, and making deductions from these. We also look at different kinds of sampling, and examine what it means for events to be independent.

1.1 Sample space, events

The general setting is: We perform an experiment which can have a number of different outcomes. The *sample space* is the set of all possible outcomes of the experiment. We usually call it $\mathcal{S}$.

It is important to be able to list the outcomes clearly. For example, if I plant ten bean seeds and count the number that germinate, the sample space is

$$\mathcal{S} = \{0,1,2,3,4,5,6,7,8,9,10\}.$$

If I toss a coin three times and record the result, the sample space is

$$\mathcal{S} = \{HHH, HHT, HTH, HTT, THH, THT, TTH, TTT\},$$

where (for example) HTH means 'heads on the first toss, then tails, then heads again'.

Sometimes we can assume that *all the outcomes are equally likely*. (Don't assume this unless either you are told to, or there is some physical reason for assuming it. In the beans example, it is most unlikely. In the coins example, the assumption will hold if the coin is 'fair': this means that there is no physical reason for it to favour one side over the other.) If all outcomes are equally likely, then each has probability $1/|\mathcal{S}|$. (Remember that $|\mathcal{S}|$ is the number of elements in the set $\mathcal{S}$).

On this point, Albert Einstein wrote, in his 1905 paper *On a heuristic point of view concerning the production and transformation of light* (for which he was awarded the Nobel Prize),

> In calculating entropy by molecular-theoretic methods, the word "probability" is often used in a sense differing from the way the word is defined in probability theory. In particular, "cases of equal probability" are often hypothetically stipulated when the theoretical methods employed are definite enough to permit a deduction rather than a stipulation.

In other words: Don't just assume that all outcomes are equally likely, *especially* when you are given enough information to calculate their probabilities!

An *event* is a subset of $\mathcal{S}$. We can specify an event by listing all the outcomes that make it up. In the above example, let A be the event 'more heads than tails' and B the event 'heads on last throw'. Then

$$\begin{aligned} A &= \{HHH, HHT, HTH, THH\}, \\ B &= \{HHH, HTH, THH, TTH\}. \end{aligned}$$

The probability of an event is calculated by adding up the probabilities of all the outcomes comprising that event. So, *if all outcomes are equally likely*, we have

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

In our example, both A and B have probability $4/8 = 1/2$.

An event is *simple* if it consists of just a single outcome, and is *compound* otherwise. In the example, A and B are compound events, while the event 'heads on every throw' is simple (as a set, it is $\{HHH\}$). If $A = \{a\}$ is a simple event, then the probability of A is just the probability of the outcome a, and we usually write $P(a)$, which is simpler to write than $P(\{a\})$. (Note that a is an *outcome*, while $\{a\}$ is an *event*, indeed a simple event.)

We can build new events from old ones:

- $A \cup B$ (read 'A union B') consists of all the outcomes in A or in B (or both!)
- $A \cap B$ (read 'A intersection B') consists of all the outcomes in both A and B;
- $A \setminus B$ (read 'A minus B') consists of all the outcomes in A but not in B;
- A' (read 'A complement') consists of all outcomes not in A (that is, $\mathcal{S} \setminus A$);
- $\emptyset$ (read 'empty set') for the event which doesn't contain any outcomes.

$A' \cap A = \emptyset$

Note the backward-sloping slash; this is not the same as either a vertical slash | or a forward slash /.

In the example, A' is the event 'more tails than heads', and $A \cap B$ is the event $\{HHH, THH, HTH\}$. Note that $P(A \cap B) = 3/8$; this is not equal to $P(A) \cdot P(B)$, despite what you read in some books!

1.2 What is probability?

There is really no answer to this question.

Some people think of it as 'limiting frequency'. That is, to say that the probability of getting heads when a coin is tossed means that, if the coin is tossed many times, it is likely to come down heads about half the time. But if you toss a coin 1000 times, you are not likely to get exactly 500 heads. You wouldn't be surprised to get only 495. But what about 450, or 100?

Some people would say that you can work out probability by physical arguments, like the one we used for a fair coin. But this argument doesn't work in all cases, and it doesn't explain what probability means.

Some people say it is subjective. You say that the probability of heads in a coin toss is $1/2$ because you have no reason for thinking either heads or tails more likely; you might change your view if you knew that the owner of the coin was a magician or a con man. But we can't build a theory on something subjective.

We regard probability as a mathematical construction satisfying some axioms (devised by the Russian mathematician A. N. Kolmogorov). We develop ways of doing calculations with probability, so that (for example) we can calculate how unlikely it is to get 480 or fewer heads in 1000 tosses of a fair coin. The answer agrees well with experiment.

1.3 Kolmogorov's Axioms

Remember that an event is a subset of the sample space $\mathcal{S}$. A number of events, say $A_1, A_2, \ldots$, are called *mutually disjoint* or *pairwise disjoint* if $A_i \cap A_j = \emptyset$ for any two of the events A_i and A_j; that is, no two of the events overlap.

According to Kolmogorov's axioms, each event A has a probability $P(A)$, which is a number. These numbers satisfy three axioms:

Axiom 1: For any event A, we have $P(A) \geq 0$.

Axiom 2: $P(\mathcal{S}) = 1$.

Axiom 3: If the events $A_1, A_2, \ldots$ are pairwise disjoint, then

$$P(A_1 \cup A_2 \cup \cdots) = P(A_1) + P(A_2) + \cdots$$

Note that in Axiom 3, we have the union of events and the sum of numbers. Don't mix these up; never write $P(A_1) \cup P(A_2)$, for example. Sometimes we separate Axiom 3 into two parts: Axiom 3a if there are only finitely many events $A_1, A_2, \ldots, A_n$, so that we have

$$P(A_1 \cup \cdots \cup A_n) = \sum_{i=1}^{n} P(A_i),$$

and Axiom 3b for infinitely many. We will only use Axiom 3a, but 3b is important later on.

Notice that we write

$$\sum_{i=1}^{n} P(A_i)$$

for

$$P(A_1) + P(A_2) + \cdots + P(A_n).$$

1.4 Proving things from the axioms

You can prove simple properties of probability from the axioms. That means, every step must be justified by appealing to an axiom. These properties seem obvious, just as obvious as the axioms; but the point of this game is that we assume only the axioms, and build everything else from that.

Here are some examples of things proved from the axioms. There is really no difference between a theorem, a proposition, and a corollary; they all have to be proved. Usually, a theorem is a big, important statement; a proposition a rather smaller statement; and a corollary is something that follows quite easily from a theorem or proposition that came before.

Proposition 1.1 *If the event A contains only a finite number of outcomes, say $A = \{a_1, a_2, \ldots, a_n\}$, then*

$$P(A) = P(a_1) + P(a_2) + \cdots + P(a_n).$$

To prove the proposition, we define a new event A_i containing only the outcome a_i, that is, $A_i = \{a_i\}$, for $i = 1, \ldots, n$. Then $A_1, \ldots, A_n$ are mutually disjoint

(each contains only one element which is in none of the others), and $A_1 \cup A_2 \cup \cdots \cup A_n = A$; so by Axiom 3a, we have

$$P(A) = P(a_1) + P(a_2) + \cdots + P(a_n).$$

Corollary 1.2 *If the sample space $\mathcal{S}$ is finite, say $\mathcal{S} = \{a_1, \ldots, a_n\}$, then*

$$P(a_1) + P(a_2) + \cdots + P(a_n) = 1.$$

For $P(a_1) + P(a_2) + \cdots + P(a_n) = P(\mathcal{S})$ by Proposition 1.1, and $P(\mathcal{S}) = 1$ by Axiom 2. Notice that once we have proved something, we can use it on the same basis as an axiom to prove further facts.

Now we see that, if all the n outcomes are equally likely, and their probabilities sum to 1, then each has probability $1/n$, that is, $1/|\mathcal{S}|$. Now going back to Proposition 1.1, we see that, *if all outcomes are equally likely*, then

$$P(A) = \frac{|A|}{|\mathcal{S}|}$$

for any event A, justifying the principle we used earlier.

Proposition 1.3 *$P(A') = 1 - P(A)$ for any event A.*

Let $A_1 = A$ and $A_2 = A'$ (the complement of A). Then $A_1 \cap A_2 = \emptyset$ (that is, the events A_1 and A_2 are disjoint), and $A_1 \cup A_2 = \mathcal{S}$. So

$$\begin{array}{rcll} P(A_1) + P(A_2) & = & P(A_1 \cup A_2) & \text{(Axiom 3)} \\ & = & P(\mathcal{S}) & \\ & = & 1 & \text{(Axiom 2).} \end{array}$$

So $P(A) = P(A_1) = 1 - P(A_2)$.

Corollary 1.4 *$P(A) \leq 1$ for any event A.*

For $1 - P(A) = P(A')$ by Proposition 1.3, and $P(A') \geq 0$ by Axiom 1; so $1 - P(A) \geq 0$, from which we get $P(A) \leq 1$.

Remember that if you ever calculate a probability to be less than 0 or more than 1, you have made a mistake!

Corollary 1.5 *$P(\emptyset) = 0$.*

For $\emptyset = \mathcal{S}'$, so $P(\emptyset) = 1 - P(\mathcal{S})$ by Proposition 1.3; and $P(\mathcal{S}) = 1$ by Axiom 2, so $P(\emptyset) = 0$.

Here is another result. The notation $A \subseteq B$ means that A is contained in B, that is, every outcome in A also belongs to B.

Proposition 1.6 *If $A \subseteq B$, then $P(A) \leq P(B)$.*

This time, take $A_1 = A$, $A_2 = B \setminus A$. Again we have $A_1 \cap A_2 = \emptyset$ (since the elements of $B \setminus A$ are, by definition, not in A), and $A_1 \cup A_2 = B$. So by Axiom 3,

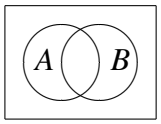
$$P(A_1) + P(A_2) = P(A_1 \cup A_2) = P(B).$$

In other words, $P(A) + P(B \setminus A) = P(B)$. Now $P(B \setminus A) \geq 0$ by Axiom 1; so

$$P(A) \leq P(B),$$

as we had to show.

1.5 Inclusion-Exclusion Principle



A Venn diagram for two sets A and B suggests that, to find the size of $A \cup B$, we add the size of A and the size of B, but then we have included the size of $A \cap B$ twice, so we have to take it off. In terms of probability:

Proposition 1.7

$$P(A \cup B) = P(A) + P(B) - P(A \cap B).$$

We now prove this from the axioms, using the Venn diagram as a guide. We see that $A \cup B$ is made up of three parts, namely

$$A_1 = A \cap B, \qquad A_2 = A \setminus B, \qquad A_3 = B \setminus A.$$

Indeed we do have $A \cup B = A_1 \cup A_2 \cup A_3$, since anything in $A \cup B$ is in both these sets or just the first or just the second. Similarly we have $A_1 \cup A_2 = A$ and $A_1 \cup A_3 = B$.

The sets A_1, A_2, A_3 are mutually disjoint. (We have three pairs of sets to check. Now $A_1 \cap A_2 = \emptyset$, since all elements of A_1 belong to B but no elements of A_2 do. The arguments for the other two pairs are similar – you should do them yourself.)

So, by Axiom 3, we have

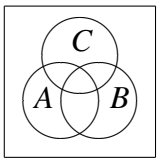
$$\begin{aligned} P(A) &= P(A_1)+P(A_2), \\ P(B) &= P(A_1)+P(A_3), \\ P(A\cup B) &= P(A_1)+P(A_2)+P(A_3). \end{aligned}$$

From this we obtain

$$\begin{aligned} P(A)+P(B)-P(A\cap B) &= (P(A_1)+P(A_2))+(P(A_1)+P(A_3))-P(A_1) \\ &= P(A_1)+P(A_2)+P(A_3) \\ &= P(A\cup B) \end{aligned}$$

as required.

The Inclusion-Exclusion Principle extends to more than two events, but gets more complicated. Here it is for three events; try to prove it yourself.



To calculate $P(A\cup B\cup C)$, we first add up $P(A)$, $P(B)$, and $P(C)$. The parts in common have been counted twice, so we subtract $P(A\cap B)$, $P(A\cap C)$ and $P(B\cap C)$. But then we find that the outcomes lying in all three sets have been taken off completely, so must be put back, that is, we add $P(A\cap B\cap C)$.

Proposition 1.8 *For any three events A,B,C, we have*

$$P(A\cup B\cup C)=P(A)+P(B)+P(C)-P(A\cap B)-P(A\cap C)-P(B\cap C)+P(A\cap B\cap C).$$

Can you extend this to any number of events?

1.6 Other results about sets

There are other standard results about sets which are often useful in probability theory. Here are some examples.

Proposition 1.9 *Let A,B,C be subsets of $\mathcal{S}$.*

Distributive laws: $(A\cap B)\cup C=(A\cup C)\cap(B\cup C)$ *and* $(A\cup B)\cap C=(A\cap C)\cup(B\cap C)$.

De Morgan's Laws: $(A\cup B)'=A'\cap B'$ *and* $(A\cap B)'=A'\cup B'$.

We will not give formal proofs of these. You should draw Venn diagrams and convince yourself that they work.

1.7 Sampling

I have four pens in my desk drawer; they are red, green, blue, and purple. I draw a pen; each pen has the same chance of being selected. In this case, $\mathcal{S} = \{R, G, B, P\}$, where R means 'red pen chosen' and so on. In this case, if A is the event 'red or green pen chosen', then

$$P(A) = \frac{|A|}{|\mathcal{S}|} = \frac{2}{4} = \frac{1}{2}.$$

More generally, if I have a set of n objects and choose one, with each one equally likely to be chosen, then each of the n outcomes has probability $1/n$, and an event consisting of m of the outcomes has probability m/n.

What if we choose more than one pen? We have to be more careful to specify the sample space.

First, we have to say whether we are

- *sampling with replacement*, or
- *sampling without replacement*.

Sampling with replacement means that we choose a pen, note its colour, put it back and shake the drawer, then choose a pen again (which may be the same pen as before or a different one), and so on until the required number of pens have been chosen. If we choose two pens with replacement, the sample space is

$$\begin{array}{cccc} \{RR, & RG, & RB, & RP, \\ GR, & GG, & GB, & GP, \\ BR, & BG, & BB, & BP, \\ PR, & PG, & PB, & PP\} \end{array}$$

The event 'at least one red pen' is $\{RR, RG, RB, RP, GR, BR, PR\}$, and has probability $7/16$.

Sampling without replacement means that we choose a pen but do not put it back, so that our final selection cannot include two pens of the same colour. In this case, the sample space for choosing two pens is

$$\begin{array}{cccc} \{ & RG, & RB, & RP, \\ GR, & & GB, & GP, \\ BR, & BG, & & BP, \\ PR, & PG, & PB & \} \end{array}$$

and the event 'at least one red pen' is $\{RG, RB, RP, GR, BR, PR\}$, with probability $6/12 = 1/2$.

Now there is another issue, depending on whether we care about the order in which the pens are chosen. We will only consider this in the case of sampling without replacement. It doesn't really matter in this case whether we choose the pens one at a time or simply take two pens out of the drawer; and we are not interested in which pen was chosen first. So in this case the sample space is

$$\{\{R,G\},\{R,B\},\{R,P\},\{G,B\},\{G,P\},\{B,P\}\},$$

containing six elements. (Each element is written as a set since, in a set, we don't care which element is first, only which elements are actually present. So the sample space is a set of sets!) The event 'at least one red pen' is $\{\{R,G\},\{R,B\},\{R,P\}\}$, with probability $3/6 = 1/2$. We should not be surprised that this is the same as in the previous case.

There are formulae for the sample space size in these three cases. These involve the following functions:

$$\begin{aligned} n! &= n(n-1)(n-2)\cdots 1 \\ {}^nP_k &= n(n-1)(n-2)\cdots(n-k+1) \\ {}^nC_k &= {}^nP_k/k! \end{aligned}$$

Note that $n!$ is the product of all the whole numbers from 1 to n; and

$${}^nP_k = \frac{n!}{(n-k)!},$$

so that

$${}^nC_k = \frac{n!}{k!(n-k)!}.$$

Theorem 1.10 *The number of selections of k objects from a set of n objects is given in the following table.*

	with replacement	*without replacement*
ordered sample	n^k	${}^nP_k$
unordered sample		${}^nC_k$

In fact the number that goes in the empty box is ${}^{n+k-1}C_k$, but this is much harder to prove than the others, and you are very unlikely to need it.

Here are the proofs of the other three cases. First, for sampling with replacement and ordered sample, there are n choices for the first object, and n choices for the second, and so on; we multiply the choices for different objects. (Think of the choices as being described by a branching tree.) The product of k factors each equal to n is n^k.

For sampling without replacement and ordered sample, there are still n choices for the first object, but now only $n-1$ choices for the second (since we do not replace the first), and $n-2$ for the third, and so on; there are $n-k+1$ choices for the kth object, since $k-1$ have previously been removed and $n-(k-1)$ remain. As before, we multiply. This product is the formula for ${}^nP_k$.

For sampling without replacement and unordered sample, think first of choosing an ordered sample, which we can do in ${}^nP_k$ ways. But each unordered sample could be obtained by drawing it in $k!$ different orders. So we divide by $k!$, obtaining ${}^nP_k/k! = {}^nC_k$ choices.

In our example with the pens, the numbers in the three boxes are $4^2 = 16$, ${}^4P_2 = 12$, and ${}^4C_2 = 6$, in agreement with what we got when we wrote them all out.

Note that, if we use the phrase 'sampling without replacement, ordered sample', or any other combination, we are assuming that *all outcomes are equally likely*.

Example The names of the seven days of the week are placed in a hat. Three names are drawn out; these will be the days of the Probability I lectures. What is the probability that no lecture is scheduled at the weekend?

Here the sampling is without replacement, and we can take it to be either ordered or unordered; the answers will be the same. For ordered samples, the size of the sample space is ${}^7P_3 = 7 \cdot 6 \cdot 5 = 210$. If A is the event 'no lectures at weekends', then A occurs precisely when all three days drawn are weekdays; so $|A| = {}^5P_3 = 5 \cdot 4 \cdot 3 = 60$. Thus, $P(A) = 60/210 = 2/7$.

If we decided to use unordered samples instead, the answer would be ${}^5C_3/{}^7C_3$, which is once again $2/7$.

Example A six-sided die is rolled twice. What is the probability that the sum of the numbers is at least 10?

This time we are sampling with replacement, since the two numbers may be the same or different. So the number of elements in the sample space is $6^2 = 36$.

To obtain a sum of 10 or more, the possibilities for the two numbers are $(4,6)$, $(5,5)$, $(6,4)$, $(5,6)$, $(6,5)$ or $(6,6)$. So the probability of the event is $6/36 = 1/6$.

Example A box contains 20 balls, of which 10 are red and 10 are blue. We draw ten balls from the box, and we are interested in the event that exactly 5 of the balls are red and 5 are blue. Do you think that this is more likely to occur if the draws are made with or without replacement?

Let $\mathcal{S}$ be the sample space, and A the event that five balls are red and five are blue.

Consider sampling with replacement. Then $|\mathcal{S}| = 20^{10}$. What is $|A|$? The number of ways in which we can choose first five red balls and then five blue ones (that is, *RRRRRBBBBB*), is $10^5 \cdot 10^5 = 10^{10}$. But there are many other ways to get five red and five blue balls. In fact, the five red balls could appear in any five of the ten draws. This means that there are ${}^{10}C_5 = 252$ different patterns of five *R*s and five *B*s. So we have

$$|A| = 252 \cdot 10^{10},$$

and so

$$P(A) = \frac{252 \cdot 10^{10}}{20^{10}} = 0.246\ldots$$

Now consider sampling without replacement. If we regard the sample as being ordered, then $|\mathcal{S}| = {}^{20}P_{10}$. There are ${}^{10}P_5$ ways of choosing five of the ten red balls, and the same for the ten blue balls, and as in the previous case there are ${}^{10}C_5$ patterns of red and blue balls. So

$$|A| = ({}^{10}P_5)^2 \cdot {}^{10}C_5,$$

and

$$P(A) = \frac{({}^{10}P_5)^2 \cdot {}^{10}C_5}{{}^{20}P_{10}} = 0.343\ldots$$

If we regard the sample as being unordered, then $|\mathcal{S}| = {}^{20}C_{10}$. There are ${}^{10}C_5$ choices of the five red balls and the same for the blue balls. We no longer have to count patterns since we don't care about the order of the selection. So

$$|A| = ({}^{10}C_5)^2,$$

and

$$P(A) = \frac{({}^{10}C_5)^2}{{}^{20}C_{10}} = 0.343\ldots$$

This is the same answer as in the case before, as it should be; the question doesn't care about order of choices!

So the event is more likely if we sample with replacement.

Example I have 6 gold coins, 4 silver coins and 3 bronze coins in my pocket. I take out three coins at random. What is the probability that they are all of different material? What is the probability that they are all of the same material?

In this case the sampling is without replacement and the sample is unordered. So $|\mathcal{S}| = {}^{13}C_3 = 286$. The event that the three coins are all of different material can occur in $6 \cdot 4 \cdot 3 = 72$ ways, since we must have one of the six gold coins, and so on. So the probability is $72/286 = 0.252\ldots$

The event that the three coins are of the same material can occur in

$$^6C_3 + {}^4C_3 + {}^3C_3 = 20 + 4 + 1 = 25$$

ways, and the probability is $25/286 = 0.087\ldots$

In a sampling problem, you should first read the question carefully and decide whether the sampling is with or without replacement. If it is without replacement, decide whether the sample is ordered (e.g. does the question say anything about the first object drawn?). If so, then use the formula for ordered samples. If not, then you can use either ordered or unordered samples, whichever is convenient; they should give the same answer. If the sample is with replacement, or if it involves throwing a die or coin several times, then use the formula for sampling with replacement.

1.8 Stopping rules

Suppose that you take a typing proficiency test. You are allowed to take the test up to three times. Of course, if you pass the test, you don't need to take it again. So the sample space is

$$\mathcal{S} = \{p, fp, ffp, fff\},$$

where for example ffp denotes the outcome that you fail twice and pass on your third attempt.

If all outcomes were equally likely, then your chance of eventually passing the test and getting the certificate would be $3/4$.

But it is unreasonable here to assume that all the outcomes are equally likely. For example, you may be very likely to pass on the first attempt. Let us assume that the probability that you pass the test is 0.8. (By Proposition 3, your chance of failing is 0.2.) Let us further assume that, no matter how many times you have failed, your chance of passing at the next attempt is still 0.8. Then we have

$$\begin{aligned} P(p) &= 0.8, \\ P(fp) &= 0.2 \cdot 0.8 = 0.16, \\ P(ffp) &= 0.2^2 \cdot 0.8 = 0.032, \\ P(fff) &= 0.2^3 = 0.008. \end{aligned}$$

Thus the probability that you eventually get the certificate is $P(\{p, fp, ffp\}) = 0.8 + 0.16 + 0.032 = 0.992$. Alternatively, you eventually get the certificate *unless* you fail three times, so the probability is $1 - 0.008 = 0.992$.

A *stopping rule* is a rule of the type described here, namely, continue the experiment until some specified occurrence happens. The experiment may potentially be infinite.

For example, if you toss a coin repeatedly until you obtain heads, the sample space is

$$\mathcal{S} = \{H, TH, TTH, TTTH, \ldots\}$$

since in principle you may get arbitrarily large numbers of tails before the first head. (We have to allow all possible outcomes.)

In the typing test, the rule is 'stop if either you pass or you have taken the test three times'. This ensures that the sample space is finite.

In the next chapter, we will have more to say about the 'multiplication rule' we used for calculating the probabilities. In the meantime you might like to consider whether it is a reasonable assumption for tossing a coin, or for someone taking a series of tests.

Other kinds of stopping rules are possible. For example, the number of coin tosses might be determined by some other random process such as the roll of a die; or we might toss a coin until we have obtained heads twice; and so on. We will not deal with these.

1.9 Questionnaire results

The students in the Probability I class in Autumn 2000 filled in the following questionnaire:

> 1. I have a hat containing 20 balls, 10 red and 10 blue. I draw 10 balls from the hat. I am interested in the event that I draw exactly five red and five blue balls. Do you think that this is more likely if I note the colour of each ball I draw and replace it in the hat, or if I don't replace the balls in the hat after drawing?
>
> *More likely with replacement* □ *More likely without replacement* □
>
> 2. What colour are your eyes?
>
> *Blue* □ *Brown* □ *Green* □ *Other* □
>
> 3. Do you own a mobile phone? *Yes* □ *No* □

After discarding incomplete questionnaires, the results were as follows:

Answer to question	"More likely with replacement"		"More likely without replacement"	
Eyes	Brown	Other	Brown	Other
Mobile phone	35	4	35	9
No mobile phone	10	3	7	1

What can we conclude?

Half the class thought that, in the experiment with the coloured balls, sampling with replacement make the result more likely. In fact, as we saw in Chapter 1, actually it is more likely if we sample without replacement. (This doesn't matter, since the students were instructed not to think too hard about it!)

You might expect that eye colour and mobile phone ownership would have no influence on your answer. Let's test this. If true, then of the 87 people with brown eyes, half of them (i.e. 43 or 44) would answer "with replacement", whereas in fact 45 did. Also, of the 83 people with mobile phones, we would expect half (that is, 41 or 42) would answer "with replacement", whereas in fact 39 of them did. So perhaps we have demonstrated that people who own mobile phones are slightly smarter than average, whereas people with brown eyes are slightly less smart!

In fact we have shown no such thing, since our results refer only to the people who filled out the questionnaire. But they do show that these events are not independent, in a sense we will come to soon.

On the other hand, since 83 out of 104 people have mobile phones, if we think that phone ownership and eye colour are independent, we would expect that the same fraction $83/104$ of the 87 brown-eyed people would have phones, i.e. $(83 \cdot 87)/104 = 69.4$ people. In fact the number is 70, or as near as we can expect. So indeed it seems that eye colour and phone ownership are more-or-less independent.

1.10 Independence

Two events A and B are said to be *independent* if

$$P(A \cap B) = P(A) \cdot P(B).$$

This is the definition of independence of events. If you are asked in an exam to define independence of events, this is the correct answer. Do not say that two events are independent if one has no influence on the other; and *under no circumstances* say that A and B are independent if $A \cap B = \emptyset$ (this is the statement that A and B are disjoint, which is quite a different thing!) Also, do not ever say that $P(A \cap B) = P(A) \cdot P(B)$ unless you have some good reason for assuming that A and B are independent (either because this is given in the question, or as in the next-but-one paragraph).

Let us return to the questionnaire example. Suppose that a student is chosen at random from those who filled out the questionnaire. Let A be the event that this student thought that the event was more likely if we sample with replacement; B the event that the student has brown eyes; and C the event that the student has a

mobile phone. Then

$$\begin{aligned} P(A) &= 52/104 = 0.5, \\ P(B) &= 87/104 = 0.8365, \\ P(C) &= 83/104 = 0.7981. \end{aligned}$$

Furthermore,

$$\begin{aligned} P(A \cap B) = 45/104 = 0.4327, \quad & P(A) \cdot P(B) = 0.4183, \\ P(A \cap C) = 39/104 = 0.375, \quad & P(A) \cdot P(C) = 0.3990, \\ P(B \cap C) = 70/104 = 0.6731, \quad & P(B) \cap P(C) = 0.6676. \end{aligned}$$

So none of the three pairs is independent, but in a sense B and C 'come closer' than either of the others, as we noted.

In practice, if it is the case that the event A has no effect on the outcome of event B, then A and B are independent. But this does not apply in the other direction. There might be a very definite connection between A and B, but still it could happen that $P(A \cap B) = P(A) \cdot P(B)$, so that A and B are independent. We will see an example shortly.

Example If we toss a coin more than once, or roll a die more than once, then you may assume that different tosses or rolls are independent. More precisely, if we roll a fair six-sided die twice, then the probability of getting 4 on the first throw and 5 on the second is $1/36$, since we assume that all 36 combinations of the two throws are equally likely. But $(1/36) = (1/6) \cdot (1/6)$, and the separate probabilities of getting 4 on the first throw and of getting 5 on the second are both equal to $1/6$. So the two events are independent. This would work just as well for any other combination.

In general, it is always OK to assume that the outcomes of different tosses of a coin, or different throws of a die, are independent. This holds even if the examples are not all equally likely. We will see an example later.

Example I have two red pens, one green pen, and one blue pen. I choose two pens without replacement. Let A be the event that I choose exactly one red pen, and B the event that I choose exactly one green pen.

If the pens are called R_1, R_2, G, B, then

$$\begin{aligned} \mathcal{S} &= \{R_1R_2, R_1G, R_1B, R_2G, R_2B, GB\}, \\ A &= \{R_1G, R_1B, R_2G, R_2B\}, \\ B &= \{R_1G, R_2G, GB\} \end{aligned}$$

We have $P(A) = 4/6 = 2/3$, $P(B) = 3/6 = 1/2$, $P(A \cap B) = 2/6 = 1/3 = P(A)P(B)$, so A and B are independent.

But before you say 'that's obvious', suppose that I have also a purple pen, and I do the same experiment. This time, if you write down the sample space and the two events and do the calculations, you will find that $P(A) = 6/10 = 3/5$, $P(B) = 4/10 = 2/5$, $P(A \cap B) = 2/10 = 1/5 \neq P(A)P(B)$, so adding one more pen has made the events non-independent!

We see that it is very difficult to tell whether events are independent or not. In practice, assume that events are independent only if either you are told to assume it, or the events are the outcomes of different throws of a coin or die. (There is one other case where you can assume independence: this is the result of different draws, with replacement, from a set of objects.)

Example Consider the experiment where we toss a fair coin three times and note the results. Each of the eight possible outcomes has probability $1/8$. Let A be the event 'there are more heads than tails', and B the event 'the results of the first two tosses are the same'. Then

- $A = \{HHH, HHT, HTH, THH\}$, $P(A) = 1/2$,
- $B = \{HHH, HHT, TTH, TTT\}$, $P(B) = 1/2$,
- $A \cap B = \{HHH, HHT\}$, $P(A \cap B) = 1/4$;

so A and B are independent. However, both A and B clearly involve the results of the first two tosses and it is not possible to make a convincing argument that one of these events has no influence or effect on the other. For example, let C be the event 'heads on the last toss'. Then, as we saw in Part 1,

- $C = \{HHH, HTH, THH, TTH\}$, $P(C) = 1/2$,
- $A \cap C = \{HHH, HTH, THH\}$, $P(A \cap C) = 3/8$;

so A and C are not independent.

Are B and C independent?

1.11 Mutual independence

This section is a bit technical. You will need to know the conclusions, though the arguments we use to reach them are not so important.

We saw in the coin-tossing example above that it is possible to have three events A, B, C so that A and B are independent, B and C are independent, but A and C are not independent.

If all three pairs of events happen to be independent, can we then conclude that $P(A \cap B \cap C) = P(A) \cdot P(B) \cdot P(C)$? At first sight this seems very reasonable; in Axiom 3, we only required all pairs of events to be exclusive in order to justify our conclusion. Unfortunately it is not true...

Example In the coin-tossing example, let A be the event 'first and second tosses have same result', B the event 'first and third tosses have the same result, and C the event 'second and third tosses have same result'. You should check that $P(A) = P(B) = P(C) = 1/2$, and that the events $A \cap B$, $B \cap C$, $A \cap C$, and $A \cap B \cap C$ are all equal to $\{HHH, TTT\}$, with probability $1/4$. Thus any pair of the three events are independent, but

$$\begin{aligned} P(A \cap B \cap C) &= 1/4, \\ P(A) \cdot P(B) \cdot P(C) &= 1/8. \end{aligned}$$

So A, B, C are not mutually independent.

The correct definition and proposition run as follows.

Let $A_1, \ldots, A_n$ be events. We say that these events are *mutually independent* if, given any distinct indices $i_1, i_2, \ldots, i_k$ with $k \geq 1$, the events

$$A_{i_1} \cap A_{i_2} \cap \cdots \cap A_{i_{k-1}} \quad \text{and} \quad A_{i_k}$$

are independent. In other words, any one of the events is independent of the intersection of any number of the other events in the set.

Proposition 1.11 *Let $A_1, \ldots, A_n$ be mutually independent. Then*

$$P(A_1 \cap A_2 \cap \cdots \cap A_n) = P(A_1) \cdot P(A_2) \cdots P(A_n).$$

Now all you really need to know is that the same 'physical' arguments that justify that two events (such as two tosses of a coin, or two throws of a die) are independent, also justify that any number of such events are mutually independent.

So, for example, if we toss a fair coin six times, the probability of getting the sequence $HHTHHT$ is $(1/2)^6 = 1/64$, and the same would apply for any other sequence. In other words, all 64 possible outcomes are equally likely.

1.12 Properties of independence

Proposition 1.12 *If A and B are independent, then A and B' are independent.*

We are given that $P(A \cap B) = P(A) \cdot P(B)$, and asked to prove that $P(A \cap B') = P(A) \cdot P(B')$.

From Corollary 4, we know that $P(B') = 1 - P(B)$. Also, the events $A \cap B$ and $A \cap B'$ are *disjoint* (since no outcome can be both in B and B'), and their union is A (since every event in A is either in B or in B'); so by Axiom 3, we have that $P(A) = P(A \cap B) + P(A \cap B')$. Thus,

$$\begin{aligned} P(A \cap B') &= P(A) - P(A \cap B) \\ &= P(A) - P(A) \cdot P(B) \\ &\qquad \text{(since } A \text{ and } B \text{ are independent)} \\ &= P(A)(1 - P(B)) \\ &= P(A) \cdot P(B'), \end{aligned}$$

which is what we were required to prove.

Corollary 1.13 *If A and B are independent, so are A' and B'.*

Apply the Proposition twice, first to A and B (to show that A and B' are independent), and then to B' and A (to show that B' and A' are independent).

More generally, if events $A_1, \ldots, A_n$ are mutually independent, and we replace some of them by their complements, then the resulting events are mutually independent. We have to be a bit careful though. For example, A and A' are not usually independent!

Results like the following are also true.

Proposition 1.14 *Let events A, B, C be mutually independent. Then A and $B \cap C$ are independent, and A and $B \cup C$ are independent.*

Example Consider the example of the typing proficiency test that we looked at earlier. You are allowed up to three attempts to pass the test.

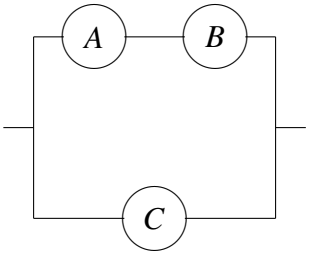
Suppose that your chance of passing the test is 0.8. Suppose also that the events of passing the test on any number of different occasions are mutually independent. Then, by Proposition 1.11, the probability of any sequence of passes and fails is the product of the probabilities of the terms in the sequence. That is,

$$P(p) = 0.8,\ P(fp) = (0.2) \cdot (0.8),\ P(ffp) = (0.2)^2 \cdot (0.8),\ P(fff) = (0.2)^3,$$

as we claimed in the earlier example.

In other words, mutual independence is the condition we need to justify the argument we used in that example.

Example
The electrical apparatus in the diagram works so long as current can flow from left to right. The three components are independent. The probability that component A works is 0.8; the probability that component B works is 0.9; and the probability that component C works is 0.75.
Find the probability that the apparatus works.



At risk of some confusion, we use the *letters* A, B and C for the *events* 'component A works', 'component B works', and 'component C works', respectively. Now the apparatus will work if *either* A and B are working, *or* C is working (or possibly both). Thus the event we are interested in is $(A \cap B) \cup C$.

Now

$$\begin{aligned} P((A \cap B) \cup C)) &= P(A \cap B) + P(C) - P(A \cap B \cap C) \\ &\qquad \text{(by Inclusion–Exclusion)} \\ &= P(A) \cdot P(B) + P(C) - P(A) \cdot P(B) \cdot P(C) \\ &\qquad \text{(by mutual independence)} \\ &= (0.8) \cdot (0.9) + (0.75) - (0.8) \cdot (0.9) \cdot (0.75) \\ &= 0.93. \end{aligned}$$

The problem can also be analysed in a different way. The apparatus will not work if both paths are blocked, that is, if C is not working and one of A and B is also not working. Thus, the event that the apparatus does not work is $(A' \cup B') \cap C'$. By the Distributive Law, this is equal to $(A' \cap C') \cup (B' \cap C')$. We have

$$\begin{aligned} P((A' \cap C') \cup (B' \cap C') &= P(A' \cap C') + P(B' \cap C') - P(A' \cap B' \cap C') \\ &\qquad \text{(by Inclusion–Exclusion)} \\ &= P(A') \cdot P(C') + P(B') \cdot P(C') - P(A') \cdot P(B') \cdot P(C') \\ &\qquad \text{(by mutual independence of } A', B', C') \\ &= (0.2) \cdot (0.25) + (0.1) \cdot (0.25) - (0.2) \cdot (0.1) \cdot (0.25) \\ &= 0.07, \end{aligned}$$

so the apparatus works with probability $1 - 0.07 = 0.93$.

There is a trap here which you should take care to avoid. You might be tempted to say $P(A' \cap C') = (0.2) \cdot (0.25) = 0.05$, and $P(B' \cap C') = (0.1) \cdot (0.25) = 0.025$; and conclude that

$$P((A' \cap C') \cup (B' \cap C')) = 0.05 + 0.025 - (0.05) \cdot (0.025) = 0.07375$$

by the Principle of Inclusion and Exclusion. But this is not correct, since the events $A' \cap C'$ and $B' \cap C'$ are *not* independent!

Example We can always assume that successive tosses of a coin are mutually independent, even if it is not a fair coin. Suppose that I have a coin which has probability 0.6 of coming down heads. I toss the coin three times. What are the probabilities of getting three heads, two heads, one head, or no heads?

For three heads, since successive tosses are mutually independent, the probability is $(0.6)^3 = 0.216$.

The probability of tails on any toss is $1 - 0.6 = 0.4$. Now the event 'two heads' can occur in three possible ways, as HHT, HTH, or THH. Each outcome has probability $(0.6) \cdot (0.6) \cdot (0.4) = 0.144$. So the probability of two heads is $3 \cdot (0.144) = 0.432$.

Similarly the probability of one head is $3 \cdot (0.6) \cdot (0.4)^2 = 0.288$, and the probability of no heads is $(0.4)^3 = 0.064$.

As a check, we have

$$0.216 + 0.432 + 0.288 + 0.064 = 1.$$

1.13 Worked examples

Question

(a) You go to the shop to buy a toothbrush. The toothbrushes there are red, blue, green, purple and white. The probability that you buy a red toothbrush is three times the probability that you buy a green one; the probability that you buy a blue one is twice the probability that you buy a green one; the probabilities of buying green, purple, and white are all equal. You are certain to buy exactly one toothbrush. For each colour, find the probability that you buy a toothbrush of that colour.

(b) James and Simon share a flat, so it would be confusing if their toothbrushes were the same colour. On the first day of term they both go to the shop to buy a toothbrush. For each of James and Simon, the probability of buying various colours of toothbrush is as calculated in (a), and their choices are independent. Find the probability that they buy toothbrushes of the same colour.

(c) James and Simon live together for three terms. On the first day of each term they buy new toothbrushes, with probabilities as in (b), independently of what they had bought before. This is the only time that they change their toothbrushes. Find the probablity that James and Simon have differently coloured toothbrushes from each other for all three terms. Is it more likely that they will have differently coloured toothbrushes from each other for

all three terms or that they will sometimes have toothbrushes of the same colour?

Solution

(a) Let R, B, G, P, W be the events that you buy a red, blue, green, purple and white toothbrush respectively. Let $x = P(G)$. We are given that

$$P(R) = 3x, \quad P(B) = 2x, \quad P(P) = P(W) = x.$$

Since these outcomes comprise the whole sample space, Corollary 2 gives

$$3x + 2x + x + x + x = 1,$$

so $x = 1/8$. Thus, the probabilities are $3/8$, $1/4$, $1/8$, $1/8$, $1/8$ respectively.

(b) Let *RB* denote the event 'James buys a red toothbrush and Simon buys a blue toothbrush', etc. By independence (given), we have, for example,

$$P(RR) = (3/8) \cdot (3/8) = 9/64.$$

The event that the toothbrushes have the same colour consists of the five outcomes *RR*, *BB*, *GG*, *PP*, *WW*, so its probability is

$$\begin{aligned} & P(RR) + P(BB) + P(GG) + P(PP) + P(WW) \\ = \; & \frac{9}{64} + \frac{1}{16} + \frac{1}{64} + \frac{1}{64} + \frac{1}{64} = \frac{1}{4}. \end{aligned}$$

(c) The event 'different coloured toothbrushes in the *i*th term' has probability $3/4$ (from part (b)), and these events are independent. So the event 'different coloured toothbrushes in all three terms' has probability

$$\frac{3}{4} \cdot \frac{3}{4} \cdot \frac{3}{4} = \frac{27}{64}.$$

The event 'same coloured toothbrushes in at least one term' is the complement of the above, so has probability $1 - (27/64) = (37)/(64)$. So it is more likely that they will have the same colour in at least one term.

Question There are 24 elephants in a game reserve. The warden tags six of the elephants with small radio transmitters and returns them to the reserve. The next month, he randomly selects five elephants from the reserve. He counts how many of these elephants are tagged. Assume that no elephants leave or enter the reserve, or die or give birth, between the tagging and the selection; and that all outcomes of the selection are equally likely. Find the probability that exactly two of the selected elephants are tagged, giving the answer correct to 3 decimal places.

Solution The experiment consists of picking the five elephants, *not* the original choice of six elephants for tagging. Let $\mathcal{S}$ be the sample space. Then $|\mathcal{S}| = {}^{24}C_5$.

Let A be the event that two of the selected elephants are tagged. This involves choosing two of the six tagged elephants and three of the eighteen untagged ones, so $|A| = {}^{6}C_2 \cdot {}^{18}C_3$. Thus

$$P(A) = \frac{{}^{6}C_2 \cdot {}^{18}C_3}{{}^{24}C_5} = 0.288$$

to 3 d.p.

Note: Should the sample should be ordered or unordered? Since the answer doesn't depend on the order in which the elephants are caught, an unordered sample is preferable. If you want to use an ordered sample, the calculation is

$$P(A) = \frac{{}^{6}P_2 \cdot {}^{18}P_3 \cdot {}^{5}C_2}{{}^{24}P_5} = 0.288,$$

since it is necessary to multiply by the ${}^{5}C_2$ possible patterns of tagged and untagged elephants in a sample of five with two tagged.

Question A couple are planning to have a family. They decide to stop having children *either* when they have two boys *or* when they have four children. Suppose that they are successful in their plan.

(a) Write down the sample space.

(b) Assume that, each time that they have a child, the probability that it is a boy is $1/2$, independent of all other times. Find $P(E)$ and $P(F)$ where $E =$ "there are at least two girls", $F =$ "there are more girls than boys".

Solution (a) $\mathcal{S} = \{BB, BGB, GBB, BGGB, GBGB, GGBB, BGGG, GBGG, GGBG, GGGB, GGGG\}$.

(b) $E = \{BGGB, GBGB, GGBB, BGGG, GBGG, GGBG, GGGB, GGGG\}$, $F = \{BGGG, GBGG, GGBG, GGGB, GGGG\}$.

Now we have $P(BB) = 1/4$, $P(BGB) = 1/8$, $P(BGGB) = 1/16$, and similarly for the other outcomes. So $P(E) = 8/16 = 1/2$, $P(F) = 5/16$.

Chapter 2

Conditional probability

In this chapter we develop the technique of conditional probability to deal with cases where events are not independent.

2.1 What is conditional probability?

Alice and Bob are going out to dinner. They toss a fair coin 'best of three' to decide who pays: if there are more heads than tails in the three tosses then Alice pays, otherwise Bob pays.

Clearly each has a 50% chance of paying. The sample space is

$$\mathcal{S} = \{HHH, HHT, HTH, HTT, THH, THT, TTH, TTT\},$$

and the events 'Alice pays' and 'Bob pays' are respectively

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

They toss the coin once and the result is heads; call this event E. How should we now reassess their chances? We have

$$E = \{HHH, HHT, HTH, HTT\},$$

and if we are given the information that the result of the first toss is heads, then E now becomes the sample space of the experiment, since the outcomes not in E are no longer possible. In the new experiment, the outcomes 'Alice pays' and 'Bob pays' are

$$A \cap E = \{HHH, HHT, HTH\},$$
$$B \cap E = \{HTT\}.$$

Thus the new probabilities that Alice and Bob pay for dinner are $3/4$ and $1/4$ respectively.

In general, suppose that we are given that an event E has occurred, and we want to compute the probability that another event A occurs. In general, we can no longer count, since the outcomes may not be equally likely. The correct definition is as follows.

Let E be an event with non-zero probability, and let A be any event. The *conditional probability of A given E* is defined as

$$P(A \mid E) = \frac{P(A \cap E)}{P(E)}.$$

Again I emphasise that this is the definition. If you are asked for the definition of conditional probability, it is not enough to say "the probability of A given that E has occurred", although this is the best way to understand it. There is no reason why event E should occur before event A!

Note the *vertical* bar in the notation. This is $P(A \mid E)$, not $P(A/E)$ or $P(A \setminus E)$.

Note also that the definition only applies in the case where $P(E)$ is not equal to zero, since we have to divide by it, and this would make no sense if $P(E) = 0$.

To check the formula in our example:

$$P(A \mid E) = \frac{P(A \cap E)}{P(E)} = \frac{3/8}{1/2} = \frac{3}{4},$$
$$P(B \mid E) = \frac{P(B \cap E)}{P(E)} = \frac{1/8}{1/2} = \frac{1}{4}.$$

It may seem like a small matter, but you should be familiar enough with this formula that you can write it down without stopping to think about the names of the events. Thus, for example,

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

if $P(B) \neq 0$.

Example A random car is chosen among all those passing through Trafalgar Square on a certain day. The probability that the car is yellow is $3/100$: the probability that the driver is blonde is $1/5$; and the probability that the car is yellow and the driver is blonde is $1/50$.

Find the conditional probability that the driver is blonde given that the car is yellow.

Solution: If Y is the event 'the car is yellow' and B the event 'the driver is blonde', then we are given that $P(Y) = 0.03$, $P(B) = 0.2$, and $P(Y \cap B) = 0.02$. So

$$P(B \mid Y) = \frac{P(B \cap Y)}{P(Y)} = \frac{0.02}{0.03} = 0.667$$

to 3 d.p. Note that we haven't used all the information given.

There is a connection between conditional probability and independence:

Proposition 2.1 *Let A and B be events with $P(B) \neq 0$. Then A and B are independent if and only if $P(A \mid B) = P(A)$.*

Proof The words 'if and only if' tell us that we have two jobs to do: we have to show that if A and B are independent, then $P(A \mid B) = P(A)$; and that if $P(A \mid B) = P(A)$, then A and B are independent.

So first suppose that A and B are independent. Remember that this means that $P(A \cap B) = P(A) \cdot P(B)$. Then

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)} = \frac{P(A) \cdot P(B)}{P(B)} = P(A),$$

that is, $P(A \mid B) = P(A)$, as we had to prove.

Now suppose that $P(A \mid B) = P(A)$. In other words,

$$\frac{P(A \cap B)}{P(B)} = P(A),$$

using the definition of conditional probability. Now clearing fractions gives

$$P(A \cap B) = P(A) \cdot P(B),$$

which is just what the statement 'A and B are independent' means.

This proposition is most likely what people have in mind when they say 'A and B are independent means that B has no effect on A'.

2.2 Genetics

Here is a simplified version of how genes code eye colour, assuming only two colours of eyes.

Each person has two genes for eye colour. Each gene is either B or b. A child receives one gene from each of its parents. The gene it receives from its father is one of its father's two genes, each with probability $1/2$; and similarly for its mother. The genes received from father and mother are independent.

If your genes are BB or Bb or bB, you have brown eyes; if your genes are bb, you have blue eyes.

Example Suppose that John has brown eyes. So do both of John's parents. His sister has blue eyes. What is the probability that John's genes are BB?

Solution John's sister has genes bb, so one b must have come from each parent. Thus each of John's parents is Bb or bB; we may assume Bb. So the possibilities for John are (writing the gene from his father first)

$$\mathrm{BB}, \mathrm{Bb}, \mathrm{bB}, \mathrm{bb}$$

each with probability $1/4$. (For example, John gets his father's B gene with probability $1/2$ and his mother's B gene with probability $1/2$, and these are independent, so the probability that he gets BB is $1/4$. Similarly for the other combinations.)

Let X be the event 'John has BB genes' and Y the event 'John has brown eyes'. Then $X = \{\mathrm{BB}\}$ and $Y = \{\mathrm{BB}, \mathrm{Bb}, \mathrm{bB}\}$. The question asks us to calculate $P(X \mid Y)$. This is given by

$$P(X \mid Y) = \frac{P(X \cap Y)}{P(Y)} = \frac{1/4}{3/4} = 1/3.$$

2.3 The Theorem of Total Probability

Sometimes we are faced with a situation where we do not know the probability of an event B, but we know what its probability would be if we were sure that some other event had occurred.

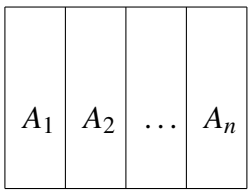
Example An ice-cream seller has to decide whether to order more stock for the Bank Holiday weekend. He estimates that, if the weather is sunny, he has a 90% chance of selling all his stock; if it is cloudy, his chance is 60%; and if it rains, his chance is only 20%. According to the weather forecast, the probability of sunshine is 30%, the probability of cloud is 45%, and the probability of rain is 25%. (We assume that these are all the possible outcomes, so that their probabilities must add up to 100%.) What is the overall probability that the salesman will sell all his stock?

This problem is answered by the *Theorem of Total Probability*, which we now state. First we need a definition. The events $A_1, A_2, \ldots, A_n$ form a *partition* of the sample space if the following two conditions hold:

(a) the events are pairwise disjoint, that is, $A_i \cap A_j = \emptyset$ for any pair of events A_i and A_j;

(b) $A_1 \cup A_2 \cup \cdots \cup A_n = \mathcal{S}$.

Another way of saying the same thing is that every outcome in the sample space lies in exactly one of the events $A_1, A_2, \ldots, A_n$. The picture shows the idea of a partition.



Now we state and prove the Theorem of Total Probability.

Theorem 2.2 *Let $A_1, A_2, \ldots, A_n$ form a partition of the sample space with $P(A_i) \neq 0$ for all i, and let B be any event. Then*

$$P(B) = \sum_{i=1}^{n} P(B \mid A_i) \cdot P(A_i).$$

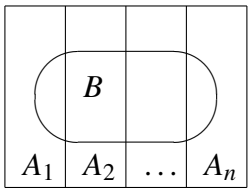
Proof By definition, $P(B \mid A_i) = P(B \cap A_i)/P(A_i)$. Multiplying up, we find that

$$P(B \cap A_i) = P(B \mid A_i) \cdot P(A_i).$$

Now consider the events $B \cap A_1, B \cap A_2, \ldots, B \cap A_n$. These events are pairwise disjoint; for any outcome lying in both $B \cap A_i$ and $B \cap A_j$ would lie in both A_i and A_j, and by assumption there are no such outcomes. Moreover, the union of all these events is B, since every outcome lies in one of the A_i. So, by Axiom 3, we conclude that

$$\sum_{i=1}^{n} P(B \cap A_i) = P(B).$$

Substituting our expression for $P(B \cap A_i)$ gives the result.



Consider the ice-cream salesman at the start of this section. Let A_1 be the event 'it is sunny', A_2 the event 'it is cloudy', and A_3 the event 'it is rainy'. Then A_1, A_2 and A_3 form a partition of the sample space, and we are given that

$$P(A_1) = 0.3, \quad P(A_2) = 0.45, \quad P(A_3) = 0.25.$$

Let B be the event 'the salesman sells all his stock'. The other information we are given is that

$$P(B \mid A_1) = 0.9, \quad P(B \mid A_2) = 0.6, \quad P(B \mid A_3) = 0.2.$$

By the Theorem of Total Probability,

$$P(B) = (0.9 \times 0.3) + (0.6 \times 0.45) + (0.2 \times 0.25) = 0.59.$$

You will now realise that the Theorem of Total Probability is really being used when you calculate probabilities by tree diagrams. It is better to get into the habit of using it directly, since it avoids any accidental assumptions of independence.

One special case of the Theorem of Total Probability is very commonly used, and is worth stating in its own right. For any event A, the events A and A' form a partition of $\mathcal{S}$. To say that both A and A' have non-zero probability is just to say that $P(A) \neq 0, 1$. Thus we have the following corollary:

Corollary 2.3 *Let A and B be events, and suppose that $P(A) \neq 0, 1$. Then*

$$P(B) = P(B \mid A) \cdot P(A) + P(B \mid A') \cdot P(A').$$

2.4 Sampling revisited

We can use the notion of conditional probability to treat sampling problems involving ordered samples.

Example I have two red pens, one green pen, and one blue pen. I select two pens without replacement.

(a) What is the probability that the first pen chosen is red?

(b) What is the probability that the second pen chosen is red?

For the first pen, there are four pens of which two are red, so the chance of selecting a red pen is $2/4 = 1/2$.

For the second pen, we must separate cases. Let A_1 be the event 'first pen red', A_2 the event 'first pen green' and A_3 the event 'first pen blue'. Then $P(A_1) = 1/2$, $P(A_2) = P(A_3) = 1/4$ (arguing as above). Let B be the event 'second pen red'.

If the first pen is red, then only one of the three remaining pens is red, so that $P(B \mid A_1) = 1/3$. On the other hand, if the first pen is green or blue, then two of the remaining pens are red, so $P(B \mid A_2) = P(B \mid A_3) = 2/3$.

By the Theorem of Total Probability,

$$\begin{aligned} P(B) &= P(B \mid A_1)P(A_1) + P(B \mid A_2)P(A_2) + P(B \mid A_3)P(A_3) \\ &= (1/3) \times (1/2) + (2/3) \times (1/4) + (2/3) \times (1/4) \\ &= 1/2. \end{aligned}$$

We have reached by a roundabout argument a conclusion which you might think to be obvious. If we have no information about the first pen, then the second pen is equally likely to be any one of the four, and the probability should be $1/2$, just as for the first pen. This argument happens to be correct. But, until your ability to distinguish between correct arguments and plausible-looking false ones is very well developed, you may be safer to stick to the calculation that we did. Beware of obvious-looking arguments in probability! Many clever people have been caught out.

2.5 Bayes' Theorem

There is a very big difference between $P(A \mid B)$ and $P(B \mid A)$.

Suppose that a new test is developed to identify people who are liable to suffer from some genetic disease in later life. Of course, no test is perfect; there will be some carriers of the defective gene who test negative, and some non-carriers who test positive. So, for example, let A be the event 'the patient is a carrier', and B the event 'the test result is positive'.

The scientists who develop the test are concerned with the probabilities that the test result is wrong, that is, with $P(B \mid A')$ and $P(B' \mid A)$. However, a patient who has taken the test has different concerns. If I tested positive, what is the chance that I have the disease? If I tested negative, how sure can I be that I am not a carrier? In other words, $P(A \mid B)$ and $P(A' \mid B')$.

These conditional probabilities are related by *Bayes' Theorem*:

Theorem 2.4 *Let A and B be events with non-zero probability. Then*

$$P(A \mid B) = \frac{P(B \mid A) \cdot P(A)}{P(B)}.$$

The proof is not hard. We have

$$P(A \mid B) \cdot P(B) = P(A \cap B) = P(B \mid A) \cdot P(A),$$

using the definition of conditional probability twice. (Note that we need both A and B to have non-zero probability here.) Now divide this equation by $P(B)$ to get the result.

If $P(A) \neq 0, 1$ and $P(B) \neq 0$, then we can use Corollary 17 to write this as

$$P(A \mid B) = \frac{P(B \mid A) \cdot P(A)}{P(B \mid A) \cdot P(A) + P(B \mid A') \cdot P(A')}.$$

Bayes' Theorem is often stated in this form.

Example Consider the ice-cream salesman from Section 2.3. Given that he sold all his stock of ice-cream, what is the probability that the weather was sunny? (This question might be asked by the warehouse manager who doesn't know what the weather was actually like.) Using the same notation that we used before, A_1 is the event 'it is sunny' and B the event 'the salesman sells all his stock'. We are asked for $P(A_1 \mid B)$. We were given that $P(B \mid A_1) = 0.9$ and that $P(A_1) = 0.3$, and we calculated that $P(B) = 0.59$. So by Bayes' Theorem,

$$P(A_1 \mid B) = \frac{P(B \mid A_1)P(A_1)}{P(B)} = \frac{0.9 \times 0.3}{0.59} = 0.46$$

to 2 d.p.

Example Consider the clinical test described at the start of this section. Suppose that 1 in 1000 of the population is a carrier of the disease. Suppose also that the probability that a carrier tests negative is 1%, while the probability that a non-carrier tests positive is 5%. (A test achieving these values would be regarded as very successful.) Let A be the event 'the patient is a carrier', and B the event 'the test result is positive'. We are given that $P(A) = 0.001$ (so that $P(A') = 0.999$), and that

$$P(B \mid A) = 0.99, \qquad P(B \mid A') = 0.05.$$

(a) A patient has just had a positive test result. What is the probability that the patient is a carrier? The answer is

$$\begin{aligned} P(A \mid B) &= \frac{P(B \mid A)P(A)}{P(B \mid A)P(A) + P(B \mid A')P(A')} \\ &= \frac{0.99 \times 0.001}{(0.99 \times 0.001) + (0.05 \times 0.999)} \\ &= \frac{0.00099}{0.05094} = 0.0194. \end{aligned}$$

(b) A patient has just had a negative test result. What is the probability that the patient is a carrier? The answer is

$$P(A \mid B') = \frac{P(B' \mid A)P(A)}{P(B' \mid A)P(A) + P(B' \mid A')P(A')}$$

$$= \frac{0.01 \times 0.001}{(0.01 \times 0.001) + (0.95 \times 0.999)}$$
$$= \frac{0.00001}{0.94095} = 0.00001.$$

So a patient with a negative test result can be reassured; but a patient with a positive test result still has less than 2% chance of being a carrier, so is likely to worry unnecessarily.

Of course, these calculations assume that the patient has been selected at random from the population. If the patient has a family history of the disease, the calculations would be quite different.

Example 2% of the population have a certain blood disease in a serious form; 10% have it in a mild form; and 88% don't have it at all. A new blood test is developed; the probability of testing positive is 9/10 if the subject has the serious form, 6/10 if the subject has the mild form, and 1/10 if the subject doesn't have the disease.

I have just tested positive. What is the probability that I have the serious form of the disease?

Let A_1 be 'has disease in serious form', A_2 be 'has disease in mild form', and A_3 be 'doesn't have disease'. Let B be 'test positive'. Then we are given that A_1, A_2, A_3 form a partition and

$$P(A_1) = 0.02 \quad P(A_2) = 0.1 \quad P(A_3) = 0.88$$
$$P(B \mid A_1) = 0.9 \quad P(B \mid A_2) = 0.6 \quad P(B \mid A_3) = 0.1$$

Thus, by the Theorem of Total Probability,

$$P(B) = 0.9 \times 0.02 + 0.6 \times 0.1 + 0.1 \times 0.88 = 0.166,$$

and then by Bayes' Theorem,

$$P(A_1 \mid B) = \frac{P(B \mid A_1)P(A_1)}{P(B)} = \frac{0.9 \times 0.02}{0.166} = 0.108$$

to 3 d.p.

2.6 Iterated conditional probability

The conditional probability of C, given that both A and B have occurred, is just $P(C \mid A \cap B)$. Sometimes instead we just write $P(C \mid A, B)$. It is given by

$$P(C \mid A, B) = \frac{P(C \cap A \cap B)}{P(A \cap B)},$$

so

$$P(A \cap B \cap C) = P(C \mid A,B)P(A \cap B).$$

Now we also have

$$P(A \cap B) = P(B \mid A)P(A),$$

so finally (assuming that $P(A \cap B) \neq 0$), we have

$$P(A \cap B \cap C) = P(C \mid A,B)P(B \mid A)P(A).$$

This generalises to any number of events:

Proposition 2.5 *Let $A_1, \ldots, A_n$ be events. Suppose that $P(A_1 \cap \cdots \cap A_{n-1}) \neq 0$. Then*

$$P(A_1 \cap A_2 \cap \cdots \cap A_n) = P(A_n \mid A_1, \ldots, A_{n-1}) \cdots P(A_2 \mid A_1)P(A_1).$$

We apply this to the birthday paradox.

The *birthday paradox* is the following statement:

> If there are 23 or more people in a room, then the chances are better than even that two of them have the same birthday.

To simplify the analysis, we ignore 29 February, and assume that the other 365 days are all equally likely as birthdays of a random person. (This is not quite true but not inaccurate enough to have much effect on the conclusion.) Suppose that we have n people $p_1, p_2, \ldots, p_n$. Let A_2 be the event 'p_2 has a different birthday from p_1'. Then $P(A_2) = 1 - \frac{1}{365}$, since whatever p_1's birthday is, there is a 1 in 365 chance that p_2 will have the same birthday.

Let A_3 be the event 'p_3 has a different birthday from p_1 and p_2'. It is not straightforward to evaluate $P(A_3)$, since we have to consider whether p_1 and p_2 have the same birthday or not. (See below). But we can calculate that $P(A_3 \mid A_2) = 1 - \frac{2}{365}$, since if A_2 occurs then p_1 and p_2 have birthdays on different days, and A_3 will occur only if p_3's birthday is on neither of these days. So

$$P(A_2 \cap A_3) = P(A_2)P(A_3 \mid A_2) = (1 - \tfrac{1}{365})(1 - \tfrac{2}{365}).$$

What is $A_2 \cap A_3$? It is simply the event that all three people have birthdays on different days.

Now this process extends. If A_i denotes the event 'p_i's birthday is not on the same day as any of $p_1, \ldots, p_{i-1}$', then

$$P(A_i \mid A_1, \ldots, A_{i-1}) = 1 - \tfrac{i-1}{365},$$

and so by Proposition 2.5,

$$P(A_1 \cap \cdots \cap A_i) = (1 - \tfrac{1}{365})(1 - \tfrac{2}{365}) \cdots (1 - \tfrac{i-1}{365}).$$

Call this number q_i; it is the probability that all of the people $p_1, \ldots, p_i$ have their birthdays on different days.

The numbers q_i decrease, since at each step we multiply by a factor less than 1. So there will be some value of n such that

$$q_{n-1} > 0.5, \qquad q_n \leq 0.5,$$

that is, n is the smallest number of people for which the probability that they all have different birthdays is less than $1/2$, that is, the probability of at least one coincidence is greater than $1/2$.

By calculation, we find that $q_{22} = 0.5243$, $q_{23} = 0.4927$ (to 4 d.p.); so 23 people are enough for the probability of coincidence to be greater than $1/2$.

Now return to a question we left open before. What is the probability of the event A_3? (This is the event that p_3 has a different birthday from both p_1 and p_2.)

If p_1 and p_2 have different birthdays, the probability is $1 - \frac{2}{365}$: this is the calculation we already did. On the other hand, if p_1 and p_2 have the same birthday, then the probability is $1 - \frac{1}{365}$. These two numbers are $P(A_3 \mid A_2)$ and $P(A_3 \mid A_2')$ respectively. So, by the Theorem of Total Probability,

$$\begin{aligned} P(A_3) &= P(A_3 \mid A_2)P(A_2) + P(A_3 \mid A_2')P(A_2') \\ &= (1 - \tfrac{2}{365})(1 - \tfrac{1}{365}) + (1 - \tfrac{1}{365})\tfrac{1}{365} \\ &= 0.9945 \end{aligned}$$

to 4 d.p.

Problem How many people would you need to pick at random to ensure that the chance of two of them being born in the same month are better than even?

Assuming all months equally likely, if B_i is the event that p_i is born in a different month from any of $p_1, \ldots, p_{i-1}$, then as before we find that

$$P(B_i \mid B_1, \cdots, B_{i-1}) = 1 - \tfrac{i-1}{12},$$

so

$$P(B_1 \cap \cdots \cap B_i) = (1 - \tfrac{1}{12})(1 - \tfrac{2}{12})(1 - \tfrac{i-1}{12}).$$

We calculate that this probability is

$$(11/12) \times (10/12) \times (9/12) = 0.5729$$

for $i = 4$ and

$$(11/12) \times (10/12) \times (9/12) \times (8/12) = 0.3819$$

for $i = 5$. So, with five people, it is more likely that two will have the same birth month.

A true story. Some years ago, in a probability class with only ten students, the lecturer started discussing the Birthday Paradox. He said to the class, "I bet that no two people in the room have the same birthday". He should have been on safe ground, since $q_{11} = 0.859$. (Remember that there are eleven people in the room!) However, a student in the back said "I'll take the bet", and after a moment all the other students realised that the lecturer would certainly lose his wager. Why?

(Answer in the next chapter.)

2.7 Worked examples

Question Each person has two genes for cystic fibrosis. Each gene is either N or C. Each child receives one gene from each parent. If your genes are NN or NC or CN then you are normal; if they are CC then you have cystic fibrosis.

(a) Neither of Sally's parents has cystic fibrosis. Nor does she. However, Sally's sister Hannah does have cystic fibrosis. Find the probability that Sally has at least one C gene (given that she does not have cystic fibrosis).

(b) In the general population the ratio of N genes to C genes is about 49 to 1. You can assume that the two genes in a person are independent. Harry does not have cystic fibrosis. Find the probability that he has at least one C gene (given that he does not have cystic fibrosis).

(c) Harry and Sally plan to have a child. Find the probability that the child will have cystic fibrosis (given that neither Harry nor Sally has it).

Solution During this solution, we will use a number of times the following principle. Let A and B be events with $A \subseteq B$. Then $A \cap B = A$, and so

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

(a) This is the same as the eye colour example discussed earlier. We are given that Sally's sister has genes CC, and one gene must come from each parent. But

neither parent is CC, so each parent is CN or NC. Now by the basic rules of genetics, all the four combinations of genes for a child of these parents, namely CC, CN, NC, NN, will have probability $1/4$.

If S_1 is the event 'Sally has at least one C gene', then $S_1 = \{CN, NC, CC\}$; and if S_2 is the event 'Sally does not have cystic fibrosis', then $S_2 = \{CN, NC, NN\}$. Then

$$P(S_1 \mid S_2) = \frac{P(S_1 \cap S_2)}{P(S_2)} = \frac{2/4}{3/4} = \frac{2}{3}.$$

(b) We know nothing specific about Harry, so we assume that his genes are randomly and independently selected from the population. We are given that the probability of a random gene being C or N is $1/50$ and $49/50$ respectively. Then the probabilities of Harry having genes CC, CN, NC, NN are respectively $(1/50)^2$, $(1/50)\cdot(49/50)$, $(49/50)\cdot(1/50)$, and $(49/50)^2$, respectively. So, if H_1 is the event 'Harry has at least one C gene', and H_2 is the event 'Harry does not have cystic fibrosis', then

$$P(H_1 \mid H_2) = \frac{P(H_1 \cap H_2)}{P(H_2)} = \frac{(49/2500) + (49/2500)}{(49/2500) + (49/2500) + (2401/2500)} = \frac{2}{51}.$$

(c) Let X be the event that Harry's and Sally's child has cystic fibrosis. As in (a), this can only occur if Harry and Sally both have CN or NC genes. That is, $X \subseteq S_3 \cap H_3$, where $S_3 = S_1 \cap S_2$ and $H_3 = H_1 \cap H_2$. Now if Harry and Sally are both CN or NC, these genes pass independently to the baby, and so

$$P(X \mid S_3 \cap H_3) = \frac{P(X)}{P(S_3 \cap H_3)} = \frac{1}{4}.$$

(Remember the principle that we started with!)

We are asked to find $P(X \mid S_2 \cap H_2)$, in other words (since $X \subseteq S_3 \cap H_3 \subseteq S_2 \cap H_2$),

$$\frac{P(X)}{P(S_2 \cap H_2)}.$$

Now Harry's and Sally's genes are independent, so

$$\begin{aligned} P(S_3 \cap H_3) &= P(S_3) \cdot P(H_3), \\ P(S_2 \cap H_2) &= P(S_2) \cdot P(H_2). \end{aligned}$$

Thus,

$$\frac{P(X)}{P(S_2 \cap H_2)} = \frac{P(X)}{P(S_3 \cap H_3)} \cdot \frac{P(S_3 \cap H_3)}{P(S_2 \cap H_2)}$$

$$\begin{aligned} &= \frac{1}{4} \cdot \frac{P(S_1 \cap S_2)}{P(S_2)} \cdot \frac{P(H_1 \cap H_2)}{P(H_2)} \\ &= \frac{1}{4} \cdot P(S_1 \mid S_2) \cdot P(H_1 \mid H_2) \\ &= \frac{1}{4} \cdot \frac{2}{3} \cdot \frac{2}{51} \\ &= \frac{1}{153}. \end{aligned}$$

I thank Eduardo Mendes for pointing out a mistake in my previous solution to this problem.

Question The Land of Nod lies in the monsoon zone, and has just two seasons, Wet and Dry. The Wet season lasts for $1/3$ of the year, and the Dry season for $2/3$ of the year. During the Wet season, the probability that it is raining is $3/4$; during the Dry season, the probability that it is raining is $1/6$.

(a) I visit the capital city, Oneirabad, on a random day of the year. What is the probability that it is raining when I arrive?

(b) I visit Oneirabad on a random day, and it is raining when I arrive. *Given this information*, what is the probability that my visit is during the Wet season?

(c) I visit Oneirabad on a random day, and it is raining when I arrive. *Given this information*, what is the probability that it will be raining when I return to Oneirabad in a year's time?

(You may assume that in a year's time the season will be the same as today but, given the season, whether or not it is raining is independent of today's weather.)

Solution (a) Let W be the event 'it is the wet season', D the event 'it is the dry season', and R the event 'it is raining when I arrive'. We are given that $P(W) = 1/3$, $P(D) = 2/3$, $P(R \mid W) = 3/4$, $P(R \mid D) = 1/6$. By the ToTP,

$$\begin{aligned} P(R) &= P(R \mid W)P(W) + P(R \mid D)P(D) \\ &= (3/4) \cdot (1/3) + (1/6) \cdot (2/3) = 13/36. \end{aligned}$$

(b) By Bayes' Theorem,

$$P(W \mid R) = \frac{P(R \mid W)P(W)}{P(R)} = \frac{(3/4) \cdot (1/3)}{13/36} = \frac{9}{13}.$$

(c) Let R' be the event 'it is raining in a year's time'. The information we are given is that $P(R \cap R' \mid W) = P(R \mid W)P(R' \mid W)$ and similarly for D. Thus

$$\begin{aligned} P(R \cap R') &= P(R \cap R' \mid W)P(W) + P(R \cap R' \mid D)P(D) \\ &= (3/4)^2 \cdot (1/3) + (1/6)^2 \cdot (2/3) = \frac{89}{432}, \end{aligned}$$

and so

$$P(R' \mid R) = \frac{P(R \cap R')}{P(R)} = \frac{89/432}{13/36} = \frac{89}{156}.$$

Chapter 3

Random variables

In this chapter we define random variables and some related concepts such as probability mass function, expected value, variance, and median; and look at some particularly important types of random variables including the binomial, Poisson, and normal.

3.1 What are random variables?

The Holy Roman Empire was, in the words of the historian Voltaire, "neither holy, nor Roman, nor an empire". Similarly, a random variable is neither random nor a variable:

A *random variable* is a function defined on a sample space.

The values of the function can be anything at all, but for us they will always be numbers. The standard abbreviation for 'random variable' is r.v.

Example I select at random a student from the class and measure his or her height in centimetres.

Here, the sample space is the set of students; the random variable is 'height', which is a function from the set of students to the real numbers: $h(S)$ is the height of student S in centimetres. (Remember that a function is nothing but a rule for associating with each element of its domain set an element of its target or range set. Here the domain set is the sample space $\mathcal{S}$, the set of students in the class, and the target space is the set of real numbers.)

Example I throw a six-sided die twice; I am interested in the sum of the two numbers. Here the sample space is

$$\mathcal{S} = \{(i, j) : 1 \leq i, j \leq 6\},$$

and the random variable F is given by $F(i,j) = i + j$. The target set is the set $\{2,3,\ldots,12\}$.

The two random variables in the above examples are representatives of the two types of random variables that we will consider. These definitions are not quite precise, but more examples should make the idea clearer.

A random variable F is *discrete* if the values it can take are separated by gaps. For example, F is discrete if it can take only finitely many values (as in the second example above, where the values are the integers from 2 to 12), or if the values of F are integers (for example, the number of nuclear decays which take place in a second in a sample of radioactive material – the number is an integer but we can't easily put an upper limit on it.)

A random variable is *continuous* if there are no gaps between its possible values. In the first example, the height of a student could in principle be any real number between certain extreme limits. A random variable whose values range over an interval of real numbers, or even over all real numbers, is continuous.

One could concoct random variables which are neither discrete nor continuous (e.g. the possible, values could be 1, 2, 3, or any real number between 4 and 5), but we will not consider such random variables.

We begin by considering discrete random variables.

3.2 Probability mass function

Let F be a discrete random variable. The most basic question we can ask is: given any value a in the target set of F, what is the probability that F takes the value a? In other words, if we consider the event

$$A = \{x \in \mathcal{S} : F(x) = a\}$$

what is $P(A)$? (Remember that an event is a subset of the sample space.) Since events of this kind are so important, we simplify the notation: we write

$$P(F = a)$$

in place of

$$P(\{x \in \mathcal{S} : F(x) = a\}).$$

(There is a fairly common convention in probability and statistics that random variables are denoted by capital letters and their values by lower-case letters. In fact, it is quite common to use the same letter in lower case for a value of the random variable; thus, we would write $P(F = f)$ in the above example. But remember that this is only a convention, and you are not bound to it.)

The *probability mass function* of a discrete random variable F is the function, formula or table which gives the value of $P(F = a)$ for each element a in the target set of F. If F takes only a few values, it is convenient to list it in a table; otherwise we should give a formula if possible. The standard abbreviation for 'probability mass function' is p.m.f.

Example I toss a fair coin three times. The random variable X gives the number of heads recorded. The possible values of X are $0, 1, 2, 3$, and its p.m.f. is

a	0	1	2	3
$P(X = a)$	$\frac{1}{8}$	$\frac{3}{8}$	$\frac{3}{8}$	$\frac{1}{8}$

For the sample space is $\{HHH, HHT, HTH, HTT, THH, THT, TTH, TTT\}$, and each outcome is equally likely. The event $X = 1$, for example, when written as a set of outcomes, is equal to $\{HTT, THT, TTH\}$, and has probability $3/8$.

Two random variables X and Y are said to have *the same distribution* if the values they take and their probability mass functions are equal. We write $X \sim Y$ in this case.

In the above example, if Y is the number of tails recorded during the experiment, then X and Y have the same distribution, even though their actual values are different (indeed, $Y = 3 - X$).

3.3 Expected value and variance

Let X be a discrete random variable which takes the values $a_1, \ldots, a_n$. The *expected value* or *mean* of X is the number $E(X)$ given by the formula

$$E(X) = \sum_{i=1}^{n} a_i P(X = a_i).$$

That is, we multiply each value of X by the probability that X takes that value, and sum these terms. The expected value is a kind of 'generalised average': if each of the values is equally likely, so that each has probability $1/n$, then $E(X) = (a_1 + \cdots + a_n)/n$, which is just the average of the values.

There is an interpretation of the expected value in terms of mechanics. If we put a mass p_i on the axis at position a_i for $i = 1, \ldots, n$, where $p_i = P(X = a_i)$, then the centre of mass of all these masses is at the point $E(X)$.

If the random variable X takes infinitely many values, say $a_1, a_2, a_3, \ldots$, then we define the expected value of X to be the infinite sum

$$E(X) = \sum_{i=1}^{\infty} a_i P(X = a_i).$$

Of course, now we have to worry about whether this means anything, that is, whether this infinite series is convergent. This is a question which is discussed at great length in analysis. We won't worry about it too much. Usually, discrete random variables will only have finitely many values; in the few examples we consider where there are infinitely many values, the series will usually be a geometric series or something similar, which we know how to sum. In the proofs below, we assume that the number of values is finite.

The *variance* of X is the number $\mathrm{Var}(X)$ given by

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

Here, X^2 is just the random variable whose values are the squares of the values of X. Thus

$$E(X^2) = \sum_{i=1}^{n} a_i^2 P(X = a_i)$$

(or an infinite sum, if necessary). The next theorem shows that, if $E(X)$ is a kind of average of the values of X, then $\mathrm{Var}(X)$ is a measure of how spread-out the values are around their average.

Proposition 3.1 *Let X be a discrete random variable with $E(X) = \mu$. Then*

$$\mathrm{Var}(X) = E((X-\mu)^2) = \sum_{i=1}^{n} (a_i - \mu)^2 P(X = a_i).$$

For the second term is equal to the third by definition, and the third is

$$\begin{aligned} & \sum_{i=1}^{n} (a_i - \mu)^2 P(X = a_i) \\ = & \sum_{i=1}^{n} (a_i^2 - 2\mu a_i + \mu^2) P(X = a_i) \\ = & \left(\sum_{i=1}^{n} a_i^2 P(X = a_i) \right) - 2\mu \left(\sum_{i=1}^{n} a_i P(X = a_i) \right) + \mu^2 \left(\sum_{i=1}^{n} P(X = a_i) \right). \end{aligned}$$

(What is happening here is that the entire sum consists of n rows with three terms in each row. We add it up by columns instead of by rows, getting three parts with n terms in each part.) Continuing, we find

$$\begin{aligned} E((X-\mu)^2) &= E(X^2) - 2\mu E(X) + \mu^2 \\ &= E(X^2) - E(X)^2, \end{aligned}$$

and we are done. (Remember that $E(X) = \mu$, and that $\sum_{i=1}^{n} P(X = a_i) = 1$ since the events $X = a_i$ form a partition.)

Some people take the conclusion of this proposition as the definition of variance.

Example I toss a fair coin three times; X is the number of heads. What are the expected value and variance of X?

$$E(X) = 0 \times (1/8) + 1 \times (3/8) + 2 \times (3/8) + 3 \times (1/8) = 3/2,$$

$$\mathrm{Var}(X) = 0^2 \times (1/8) + 1^2 \times (3/8) + 2^2 \times (3/8) + 3^2 \times (1/8) - (3/2)^2 = 3/4.$$

If we calculate the variance using Proposition 3.1, we get

$$\mathrm{Var}(X) = \left(-\frac{3}{2}\right)^2 \times \frac{1}{8} + \left(-\frac{1}{2}\right)^2 \times \frac{3}{8} + \left(\frac{1}{2}\right)^2 \times \frac{3}{8} + \left(\frac{3}{2}\right)^2 \times \frac{1}{8} = \frac{3}{4}.$$

Two properties of expected value and variance can be used as a check on your calculations.

- The expected value of X always lies between the smallest and largest values of X.
- The variance of X is never negative. (For the formula in Proposition 3.1 is a sum of terms, each of the form $(a_i - \mu)^2$ (a square, hence non-negative) times $P(X = a_i)$ (a probability, hence non-negative).

3.4 Joint p.m.f. of two random variables

Let X be a random variable taking the values $a_1, \ldots, a_n$, and let Y be a random variable taking the values $b_1, \ldots, b_m$. We say that X and Y are *independent* if, for any possible values i and j, we have

$$P(X = a_i, Y = b_j) = P(X = a_i) \cdot P(Y = b_j).$$

Here $P(X = a_i, Y = b_j)$ means the probability of the event that X takes the value a_i and Y takes the value b_j. So we could re-state the definition as follows:

> The random variables X and Y are *independent* if, for any value a_i of X and any value b_j of Y, the events $X = a_i$ and $Y = b_j$ are independent (events).

Note the difference between 'independent events' and 'independent random variables'.

Example In Chapter 2, we saw the following: I have two red pens, one green pen, and one blue pen. I select two pens without replacement. Then the events 'exactly one red pen selected' and 'exactly one green pen selected' turned out to be independent. Let X be the number of red pens selected, and Y the number of green pens selected. Then

$$P(X = 1, Y = 1) = P(X = 1) \cdot P(Y = 1).$$

Are X and Y independent random variables?

No, because $P(X = 2) = 1/6$, $P(Y = 1) = 1/2$, but $P(X = 2, Y = 1) = 0$ (it is impossible to have two red and one green in a sample of two).

On the other hand, if I roll a die twice, and X and Y are the numbers that come up on the first and second throws, then X and Y will be independent, even if the die is not fair (so that the outcomes are not all equally likely).

If we have more than two random variables (for example X, Y, Z), we say that they are *mutually independent* if the events that the random variables take specific values (for example, $X = a$, $Y = b$, $Z = c$) are mutually independent. (You may want to revise the material on mutually independent events.)

What about the expected values of random variables? For expected value, it is easy, but for variance it helps if the variables are independent:

Theorem 3.2 *Let X and Y be random variables.*

(a) $E(X+Y) = E(X) + E(Y)$.

(b) If X and Y are independent, then $\mathrm{Var}(X+Y) = \mathrm{Var}(X) + \mathrm{Var}(Y)$.

We will see the proof later.

If two random variables X and Y are not independent, then knowing the p.m.f. of each variable does not tell the whole story. The *joint probability mass function* (or *joint p.m.f.*) of X and Y is the table giving, for each value a_i of X and each value b_j of Y, the probability that $X = a_i$ and $Y = b_j$. We arrange the table so that the rows correspond to the values of X and the columns to the values of Y. Note that summing the entries in the row corresponding to the value a_i gives the probability that $X = a_i$; that is, the row sums form the p.m.f. of X. Similarly the column sums form the p.m.f. of Y. (The row and column sums are sometimes called the *marginal distributions* or *marginals*.)

In particular, X and Y are independent r.v.s if and only if each entry of the table is equal to the product of its row sum and its column sum.

Example I have two red pens, one green pen, and one blue pen, and I choose two pens without replacement. Let X be the number of red pens that I choose and Y the number of green pens. Then the joint p.m.f. of X and Y is given by the following table:

		Y	
		0	1
X	0	0	$\frac{1}{6}$
	1	$\frac{1}{3}$	$\frac{1}{3}$
	2	$\frac{1}{6}$	0

The row and column sums give us the p.m.f.s for X and Y:

a	0	1	2
$P(X=a)$	$\frac{1}{6}$	$\frac{2}{3}$	$\frac{1}{6}$

b	0	1
$P(Y=b)$	$\frac{1}{2}$	$\frac{1}{2}$

Now we give the proof of Theorem 3.2.

We consider the joint p.m.f. of X and Y. The random variable $X+Y$ takes the values a_i+b_j for $i=1,\ldots,n$ and $j=1,\ldots,m$. Now the probability that it takes a given value c_k is the sum of the probabilities $P(X=a_i,Y=b_j)$ over all i and j such that $a_i+b_j=c_k$. Thus,

$$\begin{aligned} E(X+Y) &= \sum_k c_k P(X+Y=c_k) \\ &= \sum_{i=1}^{n}\sum_{j=1}^{m}(a_i+b_j)P(X=a_i,Y=b_j) \\ &= \left(\sum_{i=1}^{n} a_i \sum_{j=1}^{m} P(X=a_i,Y=b_j)\right) + \left(\sum_{j=1}^{m} b_j \sum_{i=1}^{n} P(X=a_i,Y=b_j)\right). \end{aligned}$$

Now $\sum_{j=1}^{m} P(X=a_i,Y=b_j)$ is a row sum of the joint p.m.f. table, so is equal to $P(X=a_i)$, and similarly $\sum_{i=1}^{n} P(X=a_i,Y=b_j)$ is a column sum and is equal to $P(Y=b_j)$. So

$$\begin{aligned} E(X+Y) &= \sum_{i=1}^{n} a_i P(X=a_i) + \sum_{j=1}^{m} b_j P(Y=b_j) \\ &= E(X)+E(Y). \end{aligned}$$

The variance is a bit trickier. First we calculate

$$E((X+Y)^2) = E(X^2+2XY+Y^2) = E(X^2)+2E(XY)+E(Y^2),$$

using part (a) of the Theorem. We have to consider the term $E(XY)$. For this, we have to make the assumption that X and Y are independent, that is,

$$P(X = a_1, Y = b_j) = P(X = a_i) \cdot P(Y = b_j).$$

As before, we have

$$\begin{aligned} E(XY) &= \sum_{i=1}^{n} \sum_{j=1}^{m} a_i b_j P(X = a_i, Y = b_j) \\ &= \sum_{i=1}^{n} \sum_{j=1}^{n} a_i b_j P(X = a_i) P(Y = b_j) \\ &= \left(\sum_{i=1}^{n} a_i P(X = a_i) \right) \cdot \left(\sum_{j=1}^{m} b_j P(Y = b_j) \right) \\ &= E(X) \cdot E(Y). \end{aligned}$$

So

$$\begin{aligned} \mathrm{Var}(X+Y) &= E((X+Y)^2) - (E(X+Y))^2 \\ &= (E(X^2) + 2E(XY) + E(Y^2)) - (E(X)^2 + 2E(X)E(Y) + E(Y)^2) \\ &= (E(X^2) - E(X)^2) + 2(E(XY) - E(X)E(Y)) + (E(Y^2) - E(Y)^2) \\ &= \mathrm{Var}(X) + \mathrm{Var}(Y). \end{aligned}$$

To finish this section, we consider constant random variables. (If the thought of a 'constant variable' worries you, remember that a random variable is not a variable at all but a function, and there is nothing amiss with a constant function.)

Proposition 3.3 *Let C be a constant random variable with value c. Let X be any random variable.*

(a) $E(C) = c$, $\mathrm{Var}(C) = 0$.

(b) $E(X + c) = E(X) + c$, $\mathrm{Var}(X + c) = \mathrm{Var}(X)$.

(c) $E(cX) = cE(X)$, $\mathrm{Var}(cX) = c^2 \mathrm{Var}(X)$.

Proof (a) The random variable C takes the single value c with $P(C = c) = 1$. So $E(C) = c \cdot 1 = c$. Also,

$$\mathrm{Var}(C) = E(C^2) - E(C)^2 = c^2 - c^2 = 0.$$

(For C^2 is a constant random variable with value c^2.)

(b) This follows immediately from Theorem 3.2, once we observe that the constant random variable C and any random variable X are independent. (This is true because $P(X = a, C = c) = P(X = a) \cdot 1$.) Then

$$E(X + c) = E(X) + E(C) = E(X) + c,$$
$$\mathrm{Var}(X + c) = \mathrm{Var}(X) + \mathrm{Var}(C) = \mathrm{Var}(X).$$

(c) If $a_1, \ldots, a_n$ are the values of X, then $ca_1, \ldots, ca_n$ are the values of cX, and $P(cX = ca_i) = P(x = a_i)$. So

$$\begin{aligned} E(cX) &= \sum_{i=1}^{n} ca_i P(cX = ca_i) \\ &= c \sum_{i=1}^{n} a_i P(X = a_i) \\ &= cE(X). \end{aligned}$$

Then

$$\begin{aligned} \mathrm{Var}(cX) &= E(c^2X^2) - E(cX)^2 \\ &= c^2 E(X^2) - (cE(X))^2 \\ &= c^2 (E(X^2) - E(X)^2) \\ &= c^2 \mathrm{Var}(X). \end{aligned}$$

3.5 Some discrete random variables

We now look at five types of discrete random variables, each depending on one or more parameters. We describe for each type the situations in which it arises, and give the p.m.f., the expected value, and the variance. If the variable is tabulated in the *New Cambridge Statistical Tables*, we give the table number, and some examples of using the tables. You should have a copy of the tables to follow the examples.

A summary of this information is given in Appendix B.

Before we begin, a comment on the *New Cambridge Statistical Tables*. They don't give the probability mass function (or p.m.f.), but a closely related function called the *cumulative distribution function*. It is defined for a discrete random variable as follows.

Let X be a random variable taking values $a_1, a_2, \ldots, a_n$. We assume that these are arranged in ascending order: $a_1 < a_2 < \cdots < a_n$. The *cumulative distribution function*, or c.d.f., of X is given by

$$F_X(a_i) = P(X \leq a_i).$$

We see that it can be expressed in terms of the p.m.f. of X as follows:

$$F_X(a_i) = P(X = a_1) + \cdots + P(X = a_i) = \sum_{j=1}^{i} P(X = a_j).$$

In the other direction, we cn recover the p.m.f. from the c.d.f.:

$$P(X = a_i) = F_X(a_i) - F_X(a_{i-1}).$$

We won't use the c.d.f. of a discrete random variable except for looking up the tables. It is much more important for continuous random variables!

Bernoulli random variable Bernoulli(p)

A Bernoulli random variable is the simplest type of all. It only takes two values, 0 and 1. So its p.m.f. looks as follows:

x	0	1
$P(X = x)$	q	p

Here, p is the probability that $X = 1$; it can be any number between 0 and 1. Necessarily q (the probability that $X = 0$) is equal to $1 - p$. So p determines everything.

For a Bernoulli random variable X, we sometimes describe the experiment as a 'trial', the event $X = 1$ as'success', and the event $X = 0$ as 'failure'.

For example, if a biased coin has probability p of coming down heads, then the number of heads that we get when we toss the coin once is a Bernoulli(p) random variable.

More generally, let A be any event in a probability space $\mathcal{S}$. With A, we associate a random variable I_A (remember that a random variable is just a function on $\mathcal{S}$) by the rule

$$I_A(s) = \begin{cases} 1 & \text{if } s \in A; \\ 0 & \text{if } s \notin A. \end{cases}$$

The random variable I_A is called the *indicator variable* of A, because its value indicates whether or not A occurred. It is a Bernoulli(p) random variable, where $p = P(A)$. (The event $I_A = 1$ is just the event A.) Some people write $\mathbf{1}_A$ instead of I_A.

Calculation of the expected value and variance of a Bernoulli random variable is easy. Let $X \sim$ Bernoulli(p). (Remember that $\sim$ means "has the same p.m.f. as".)

$E(X) = 0 \cdot q + 1 \cdot p = p;$

$\mathrm{Var}(X) = 0^2 \cdot q + 1^2 \cdot p - p^2 = p - p^2 = pq.$

(Remember that $q = 1 - p$.)

Binomial random variable $\mathrm{Bin}(n,p)$

Remember that for a Bernoulli random variable, we describe the event $X=1$ as a 'success'. Now a *binomial random variable* counts the number of successes in n independent trials each associated with a Bernoulli(p) random variable.

For example, suppose that we have a biased coin for which the probability of heads is p. We toss the coin n times and count the number of heads obtained. This number is a $\mathrm{Bin}(n,p)$ random variable.

A $\mathrm{Bin}(n,p)$ random variable X takes the values $0,1,2,\ldots,n$, and the p.m.f. of X is given by

$$P(X=k) = {}^nC_k q^{n-k} p^k$$

for $k=0,1,2,\ldots,n$, where $q=1-p$. This is because there are ${}^nC_k$ different ways of obtaining k heads in a sequence of n throws (the number of choices of the k positions in which the heads occur), and the probability of getting k heads and $n-k$ tails in a particular order is $q^{n-k}p^k$.

Note that we have given a formula rather than a table here. For small values we could tabulate the results; for example, for $\mathrm{Bin}(4,p)$:

k	0	1	2	3	4
$P(X=k)$	q^4	$4q^3p$	$6q^2p^2$	$4qp^3$	p^4

Note: when we add up all the probabilities in the table, we get

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

as it should be: here we used the *binomial theorem*

$$(x+y)^n = \sum_{k=0}^{n} {}^nC_k x^{n-k} y^k.$$

(This argument explains the name of the binomial random variable!)

If $X \sim \mathrm{Bin}(n,p)$, then

$$E(X) = np, \qquad \mathrm{Var}(X) = npq.$$

There are two ways to prove this, an easy way and a harder way. The easy way only works for the binomial, but the harder way is useful for many random variables. However, you can skip it if you wish: I have set it in smaller type for this reason.

Here is the easy method. We have a coin with probability p of coming down heads, and we toss it n times and count the number X of heads. Then X is our $\mathrm{Bin}(n,p)$ random variable. Let X_k be the random variable defined by

$$X_k = \begin{cases} 1 & \text{if we get heads on the } k\text{th toss,} \\ 0 & \text{if we get tails on the } k\text{th toss.} \end{cases}$$

In other words, X_i is the indicator variable of the event 'heads on the kth toss'. Now we have

$$X = X_1 + X_2 + \cdots + X_n$$

(can you see why?), and $X_1, \ldots, X_n$ are *independent* Bernoulli(p) random variables (since they are defined by different tosses of a coin). So, as we saw earlier, $E(X_i) = p$, $\mathrm{Var}(X_i) = pq$. Then, by Theorem 21, since the variables are independent, we have

$$\begin{aligned} E(X) &= p + p + \cdots + p = np, \\ \mathrm{Var}(X) &= pq + pq + \cdots + pq = npq. \end{aligned}$$

The other method uses a gadget called the *probability generating function*. We only use it here for calculating expected values and variances, but if you learn more probability theory you will see other uses for it. Let X be a random variable whose values are non-negative integers. (We don't insist that it takes all possible values; this method is fine for the binomial $\mathrm{Bin}(n,p)$, which takes values between 0 and n. To save space, we write p_k for the probability $P(X = k)$. Now the *probability generating function* of X is the power series

$$G_X(x) = \sum p_k x^k.$$

(The sum is over all values k taken by X.)

We use the notation $[F(x)]_{x=1}$ for the result of substituting $x = 1$ in the series $F(x)$.

Proposition 3.4 *Let $G_X(x)$ be the probability generating function of a random variable X. Then*

(a) $[G_X(x)]_{x=1} = 1$;

(b) $E(X) = \left[\frac{\mathrm{d}}{\mathrm{d}x} G_X(x)\right]_{x=1}$;

(c) $\mathrm{Var}(X) = \left[\frac{\mathrm{d}^2}{\mathrm{d}x^2} G_X(x)\right]_{x=1} + E(X) - E(X)^2$.

Part (a) is just the statement that probabilities add up to 1: when we substitute $x = 1$ in the power series for $G_X(x)$ we just get $\sum p_k$.

For part (b), when we differentiate the series term-by-term (you will learn later in Analysis that this is OK), we get

$$\frac{\mathrm{d}}{\mathrm{d}x} G_X(x) = \sum k p_k x^{k-1}.$$

Now putting $x = 1$ in this series we get

$$\sum k p_k = E(X).$$

For part (c), differentiating twice gives

$$\frac{\mathrm{d}^2}{\mathrm{d}x^2} G_X(x) = \sum k(k-1) p_k x^{k-2}.$$

Now putting $x = 1$ in this series we get

$$\sum k(k-1) p_k = \sum k^2 p_k - \sum k p_k = E(X^2) - E(X).$$

Adding $E(X)$ and subtracting $E(X)^2$ gives $E(X^2) - E(X)^2$, which by definition is $\mathrm{Var}(X)$.

Now let us appply this to the binomial random variable $X \sim \mathrm{Bin}(n,p)$. We have

$$p_k = P(X=k) = {}^nC_k q^{n-k} p^k,$$

so the probability generating function is

$$\sum_{k=0}^{n} {}^nC_k q^{n-k} p^k x^k = (q+px)^n,$$

by the Binomial Theorem. Putting $x=1$ gives $(q+p)^n = 1$, in agreement with Proposition 3.4(a).

Differentiating once, using the Chain Rule, we get $np(q+px)^{n-1}$. Putting $x=1$ we find that

$$E(X) = np.$$

Differentiating again, we get $n(n-1)p^2(q+px)^{n-2}$. Putting $x=1$ gives $n(n-1)p^2$. Now adding $E(X)-E(X)^2$, we get

$$\mathrm{Var}(X) = n(n-1)p^2 + np - n^2p^2 = np - np^2 = npq.$$

The binomial random variable is tabulated in Table 1 of the *Cambridge Statistical Tables* [1]. As explained earlier, the tables give the cumulative distribution function.

For example, suppose that the probability that a certain coin comes down heads is 0.45. If the coin is tossed 15 times, what is the probability of five or fewer heads? Turning to the page $n=15$ in Table 1 and looking at the row 0.45, you read off the answer 0.2608. What is the probability of exactly five heads? This is $P(5 \text{ or fewer}) - P(4 \text{ or fewer})$, and from tables the answer is $0.2608 - 0.1204 = 0.1404$.

The tables only go up to $p = 0.5$. For larger values of p, use the fact that the number of failures in $\mathrm{Bin}(n,p)$ is equal to the number of successes in $\mathrm{Bin}(n, 1-p)$. So the probability of five heads in 15 tosses of a coin with $p = 0.55$ is $0.9745 - 0.9231 = 0.0514$.

Another interpretation of the binomial random variable concerns sampling. Suppose that we have N balls in a box, of which M are red. We sample n balls from the box with replacement; let the random variable X be the number of red balls in the sample. What is the distribution of X? Since each ball has probability M/N of being red, and different choices are independent, $X \sim \mathrm{Bin}(n,p)$, where $p = M/N$ is the proportion of red balls in the sample.

What about sampling without replacement? This leads us to our next random variable:

Hypergeometric random variable $\mathrm{Hg}(n,M,N)$

Suppose that we have N balls in a box, of which M are red. We sample n balls from the box *without replacement*. Let the random variable X be the number of

red balls in the sample. Such an X is called a *hypergeometric* random variable $\mathrm{Hg}(n,M,N)$.

The random variable X can take any of the values $0,1,2,\ldots,n$. Its p.m.f. is given by the formula

$$P(X=k) = \frac{{}^{M}C_k \cdot {}^{N-M}C_{n-k}}{{}^{N}Cn}.$$

For the number of samples of n balls from N is ${}^{N}C_n$; the number of ways of choosing k of the M red balls and $n-k$ of the $N-M$ others is ${}^{M}C_k \cdot {}^{N-M}C_{n-k}$; and all choices are equally likely.

The expected value and variance of a hypergeometric random variable are as follows (we won't go into the proofs):

$$E(X) = n\left(\frac{M}{N}\right), \qquad \mathrm{Var}(X) = n\left(\frac{M}{N}\right)\left(\frac{N-M}{N}\right)\left(\frac{N-n}{N-1}\right).$$

You should compare these to the values for a binomial random variable. If we let $p = M/N$ be the proportion of red balls in the hat, then $E(X) = np$, and $\mathrm{Var}(X)$ is equal to npq multiplied by a 'correction factor' $(N-n)/(N-1)$.

In particular, if the numbers M and $N-M$ of red and non-red balls in the hat are both very large compared to the size n of the sample, then the difference between sampling with and without replacement is very small, and indeed the 'correction factor' is close to 1. So we can say that $\mathrm{Hg}(n,M,N)$ is approximately $\mathrm{Bin}(n,M/N)$ if n is small compared to M and $N-M$.

Consider our example of choosing two pens from four, where two pens are red, one green, and one blue. The number X of red pens is a $\mathrm{Hg}(2,2,4)$ random variable. We calculated earlier that $P(X=0) = 1/6$, $P(X=1) = 2/3$ and $P(X=2) = 1/6$. From this we find by direct calculation that $E(X) = 1$ and $\mathrm{Var}(X) = 1/3$. These agree with the formulae above.

Geometric random variable $\mathrm{Geom}(p)$

The geometric random variable is like the binomial but with a different stopping rule. We have again a coin whose probability of heads is p. Now, instead of tossing it a fixed number of times and counting the heads, we toss it until it comes down heads for the first time, and count the number of times we have tossed the coin. Thus, the values of the variable are the positive integers $1,,2,3,\ldots$ (In theory we might never get a head and toss the coin infinitely often, but if $p > 0$ this possibility is 'infinitely unlikely', i.e. has probability zero, as we will see.) We always assume that $0 < p < 1$.

More generally, the number of independent Bernoulli trials required until the first success is obtained is a geometric random variable.

The p.m.f of a Geom(p) random variable is given by

$$P(X = k) = q^{k-1}p,$$

where $q = 1 - p$. For the event $X = k$ means that we get tails on the first $k - 1$ tosses and heads on the kth, and this event has probability $q^{k-1}p$, since 'tails' has probability q and different tosses are independent.

Let's add up these probabilities:

$$\sum_{k=1}^{\infty} q^{k-1}p = p + qp + q^2p + \cdots = \frac{p}{1-q} = 1,$$

since the series is a geometric progression with first term p and common ratio q, where $q < 1$. (Just as the binomial theorem shows that probabilities sum to 1 for a binomial random variable, and gives its name to the random variable, so the geometric progression does for the geometric random variable.)

We calculate the expected value and the variance using the probability generating function. If $X \sim \text{Geom}(p)$, the result will be that

$$E(X) = 1/p, \qquad \text{Var}(X) = q/p^2.$$

We have

$$G_X(x) = \sum_{k=1}^{\infty} q^{k-1}px^k = \frac{px}{1-qx},$$

again by summing a geometric progression. Differentiating, we get

$$\frac{\mathrm{d}}{\mathrm{d}x}G_X(x) = \frac{(1-qx)p + pxq}{(1-qx)^2} = \frac{p}{(1-qx)^2}.$$

Putting $x = 1$, we obtain

$$E(X) = \frac{p}{(1-q)^2} = \frac{1}{p}.$$

Differentiating again gives $2pq/(1-qx)^3$, so

$$\text{Var}(X) = \frac{2pq}{p^3} + \frac{1}{p} - \frac{1}{p^2} = \frac{q}{p^2}.$$

For example, if we toss a fair coin until heads is obtained, the expected number of tosses until the first head is 2 (so the expected number of tails is 1); and the variance of this number is also 2.

Poisson random variable Poisson(λ)

The Poisson random variable, unlike the ones we have seen before, is very closely connected with continuous things.

Suppose that 'incidents' occur at random times, but at a steady rate overall. The best example is radioactive decay: atomic nuclei decay randomly, but the average number λ which will decay in a given interval is constant. The Poisson random variabe X counts the number of 'incidents' which occur in a given interval. So if, on average, there are 2.4 nuclear decays per second, then the number of decays in one second starting now is a Poisson(2.4) random variable.

Another example might be the number of telephone calls a minute to a busy telephone number.

Although we will not prove it, the p.m.f. for a Poisson(λ) variable X is given by the formula

$$P(X = k) = \frac{\lambda^k}{k!} e^{-\lambda}.$$

Let's check that these probabilities add up to one. We get

$$\left(\sum_{k=0}^{\infty} \frac{\lambda^k}{k!} \right) e^{-\lambda} = e^{\lambda} \cdot e^{-\lambda} = 1,$$

since the expression in brackets is the sum of the exponential series.

By analogy with what happened for the binomial and geometric random variables, you might have expected that this random variable would be called 'exponential'. Unfortunately, this name has been given to a closely-related continuous random variable which we will meet later. However, if you speak a little French, you might use as a mnemonic the fact that if I go fishing, and the fish are biting at the rate of λ per hour on average, then the number of fish I will catch in the next hour is a Poisson(λ) random variable.

The expected value and variance of a Poisson(λ) random variable X are given by

$$E(X) = \mathrm{Var}(X) = \lambda.$$

Again we use the probability generating function. If $X \sim \mathrm{Poisson}(\lambda)$, then

$$G_X(x) = \sum_{k=0}^{\infty} \frac{(\lambda x)^k}{k!} e^{-\lambda} = e^{\lambda(x-1)},$$

again using the series for the exponential function.

Differentiation gives $\lambda e^{\lambda(x-1)}$, so $E(X) = \lambda$. Differentiating again gives $\lambda^2 e^{\lambda(x-1)}$, so

$$\mathrm{Var}(X) = \lambda^2 + \lambda - \lambda^2 = \lambda.$$

The cumulative distribution function of a Poisson random variable is tabulated in Table 2 of the *New Cambridge Statistical Tables*. So, for example, we find from the tables that, if 2.4 fish bite per hour on average, then the probability that I will catch no fish in the next hour is 0.0907, while the probability that I catch at five or fewer is 0.9643 (so that the probability that I catch six or more is 0.0357).

There is another situation in which the Poisson distribution arises. Suppose I am looking for some very rare event which only occurs once in 1000 trials on average. So I conduct 1000 independent trials. How many occurrences of the event do I see? This number is really a binomial random variable $\mathrm{Bin}(1000, 1/1000)$. But it turns out to be $\mathrm{Poisson}(1)$, to a very good approximation. So, for example, the probability that the event doesn't occur is about $1/\mathrm{e}$.

The general rule is:

> If n is large, p is small, and $np = \lambda$, then $\mathrm{Bin}(n, p)$ can be approximated by $\mathrm{Poisson}(\lambda)$.

3.6 Continuous random variables

We haven't so far really explained what a continuous random variable is. Its target set is the set of real numbers, or perhaps the non-negative real numbers or just an interval. The crucial property is that, for any real number a, we have $(X = a) = 0$; that is, the probability that the height of a random student, or the time I have to wait for a bus, is *precisely* a, is zero. So we can't use the probability mass function for continuous random variables; it would always be zero and give no information.

We use the *cumulative distribution function* or c.d.f. instead. Remember from last week that the c.d.f. of the random variable X is the function F_X defined by

$$F_X(x) = P(X \leq x).$$

Note: The name of the function is F_X; the lower case x refers to the argument of the function, the number which is substituted into the function. It is common but not universal to use as the argument the lower-case version of the name of the random variable, as here. Note that $F_X(y)$ is the same function written in terms of the variable y instead of x, whereas $F_Y(x)$ is the c.d.f. of the random variable Y (which might be a completely different function.)

Now let X be a continuous random variable. Then, since the probability that X takes the precise value x is zero, there is no difference between $P(X \leq x)$ and $P(X < x)$.

Proposition 3.5 *The c.d.f. is an increasing function (this means that $F_X(x) \leq F_X(y)$ if $x < y$), and approaches the limits 0 as $x \to -\infty$ and 1 as $x \to \infty$.*

The function is increasing because, if $x < y$, then

$$F_X(y) - F_X(x) = P(X \leq y) - P(X \leq x) = P(x < X \leq y) \geq 0.$$

Also $F_X(\infty) = 1$ because X must certainly take some finite value; and $F_X(-\infty) = 0$ because no value is smaller than $-\infty$!

Another important function is the *probability density function* f_X. It is obtained by differentiating the c.d.f.:

$$f_X(x) = \frac{\mathrm{d}}{\mathrm{d}x} F_X(x).$$

Now $f_X(x)$ is non-negative, since it is the derivative of an increasing function. If we know $f_X(x)$, then F_X is obtained by integrating. Because $F_X(-\infty) = 0$, we have

$$F_X(x) = \int_{-\infty}^{x} f_X(t)\mathrm{d}t.$$

Note the use of the "dummy variable" t in this integral. Note also that

$$P(a \leq X \leq b) = F_X(b) - F_X(a) = \int_a^b f_X(t)\mathrm{d}t.$$

You can think of the p.d.f. like this: the probability that the value of X lies in a very small interval from x to $x+h$ is approximately $f_X(x) \cdot h$. So, although the probability of getting exactly the value x is zero, the probability of being close to x is proportional to $f_X(x)$.

There is a mechanical analogy which you may find helpful. Remember that we modelled a discrete random variable X by placing at each value a of X a mass equal to $P(X = a)$. Then the total mass is one, and the expected value of X is the centre of mass. For a continuous random variable, imagine instead a wire of variable thickness, so that the density of the wire (mass per unit length) at the point x is equal to $f_X(x)$. Then again the total mass is one; the mass to the left of x is $F_X(x)$; and again it will hold that the centre of mass is at $E(X)$.

Most facts about continuous random variables are obtained by replacing the p.m.f. by the p.d.f. and replacing sums by integrals. Thus, the *expected value* of X is given by

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

and the *variance* is (as before)

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

where

$$E(X^2) = \int_{-\infty}^{\infty} x^2 f_X(x)\mathrm{d}x.$$

It is also true that $\mathrm{Var}(X) = E((X-\mu)^2)$, where $\mu = E(X)$.

We will see examples of these calculations shortly. But here is a small example to show the ideas. The *support* of a continuous random variable is the smallest interval containing all values of x where $f_X(x) > 0$.

Suppose that the random variable X has p.d.f. given by

$$f_X(x) = \begin{cases} 2x & \text{if } 0 \le x \le 1, \\ 0 & \text{otherwise.} \end{cases}$$

The support of $X$is the interval $[0,1]$. We check the integral:

$$\int_{-\infty}^{\infty} f_X(x)\mathrm{d}x = \int_0^1 2x\,\mathrm{d}x = \left[x^2\right]_{x=0}^{x=1} = 1.$$

The cumulative distribution function of X is

$$F_X(x) = \int_{-\infty}^{x} f_X(t)\mathrm{d}t = \begin{cases} 0 & \text{if } x < 0, \\ x^2 & \text{if } 0 \le x \le 1, \\ 1 & \text{if } x > 1. \end{cases}$$

(Study this carefully to see how it works.) We have

$$\begin{aligned} E(X) &= \int_{-\infty}^{\infty} x f_X(x)\mathrm{d}x = \int_0^1 2x^2\mathrm{d}x = \frac{2}{3}, \\ E(X^2) &= \int_{-\infty}^{\infty} x^2 f_X(x)\mathrm{d}x = \int_0^1 2x^3\mathrm{d}x = \frac{1}{2}, \\ \mathrm{Var}(X) &= \frac{1}{2} - \left(\frac{2}{3}\right)^2 = \frac{1}{18}. \end{aligned}$$

3.7 Median, quartiles, percentiles

Another measure commonly used for continuous random variables is the *median*; this is the value m such that "half of the distribution lies to the left of m and half to the right". More formally, m should satisfy $F_X(m) = 1/2$. It is not the same as the mean or expected value. In the example at the end of the last section, we saw that $E(X) = 2/3$. The median of X is the value of m for which $F_X(m) = 1/2$. Since $F_X(x) = x^2$ for $0 \le x \le 1$, we see that $m = 1/\sqrt{2}$.

If there is a value m such that the graph of $y = f_X(x)$ is symmetric about $x = m$, then both the expected value and the median of X are equal to m.

The *lower quartile* l and the *upper quartile* u are similarly defined by

$$F_X(l) = 1/4, \qquad F_X(u) = 3/4.$$

Thus, the probability that X lies between l and u is $3/4 - 1/4 = 1/2$, so the quartiles give an estimate of how spread-out the distribution is. More generally, we define the nth *percentile* of X to be the value of x_n such that

$$F_X(x_n) = n/100,$$

that is, the probability that X is smaller than x_n is $n\%$.

Reminder If the c.d.f. of X is $F_X(x)$ and the p.d.f. is $f_X(x)$, then

- differentiate F_X to get f_X, and integrate f_X to get F_X;
- use f_X to calculate $E(X)$ and $\mathrm{Var}(X)$;
- use F_X to calculate $P(a \leq X \leq b)$ (this is $F_X(b) - F_X(a)$), and the median and percentiles of X.

3.8 Some continuous random variables

In this section we examine three important continuous random variables: the uniform, exponential, and normal. The details are summarised in Appendix B.

Uniform random variable $U(a,b)$

Let a and b be real numbers with $a < b$. A uniform random variable on the interval $[a,b]$ is, roughly speaking, "equally likely to be anywhere in the interval". In other words, its probability density function is constant on the interval $[a,b]$ (and zero outside the interval). What should the constant value c be? The integral of the p.d.f. is the area of a rectangle of height c and base $b-a$; this must be 1, so $c = 1/(b-a)$. Thus, the p.d.f. of the random variable $X \sim U(a,b)$ is given by

$$f_X(x) = \begin{cases} 1/(b-a) & \text{if } a \leq x \leq b, \\ 0 & \text{otherwise.} \end{cases}$$

By integration, we find that the c.d.f. is

$$F_X(x) = \begin{cases} 0 & \text{if } x < a, \\ (x-a)/(b-a) & \text{if } a \leq x \leq b, \\ 1 & \text{if } x > b. \end{cases}$$

Further calculation (or the symmetry of the p.d.f.) shows that the expected value and the median of X are both given by $(a+b)/2$ (the midpoint of the interval), while $\mathrm{Var}(X) = (b-a)^2/12$.

The uniform random variable doesn't really arise in practical situations. However, it is very useful for simulations. Most computer systems include a *random number generator*, which apparently produces independent values of a uniform random variable on the interval $[0,1]$. Of course, they are not really random, since the computer is a deterministic machine; but there should be no obvious pattern to

the numbers produced, and in a large number of trials they should be distributed uniformly over the interval.

You will learn in the Statistics course how to use a uniform random variable to construct values of other types of discrete or continuous random variables. Its great simplicity makes it the best choice for this purpose.

Exponential random variable $\mathrm{Exp}(\lambda)$

The exponential random variable arises in the same situation as the Poisson: be careful not to confuse them! We have events which occur randomly but at a constant average rate of λ per unit time (e.g. radioactive decays, fish biting). The Poisson random variable, which is discrete, counts how many events will occur in the next unit of time. The exponential random variable, which is continuous, measures exactly how long from now it is until the next event occurs. Not that it takes non-negative real numbers as values.

If $X \sim \mathrm{Exp}(\lambda)$, the p.d.f. of X is

$$f_X(x) = \begin{cases} 0 & \text{if } x < 0, \\ \lambda \mathrm{e}^{-\lambda x} & \text{if } x \geq 0. \end{cases}$$

By integration, we find the c.d.f. to be

$$F_X(x) = \begin{cases} 0 & \text{if } x < 0, \\ 1 - \mathrm{e}^{-\lambda x} & \text{if } x \geq 0. \end{cases}$$

Further calculation gives

$$E(X) = 1/\lambda, \qquad \mathrm{Var}(X) = 1/\lambda^2.$$

The median m satisfies $1 - \mathrm{e}^{-\lambda m} = 1/2$, so that $m = \log 2/\lambda$. (The logarithm is to base e, so that $\log 2 = 0.69314718056$ approximately.

Normal random variable $N(\mu, \sigma^2)$

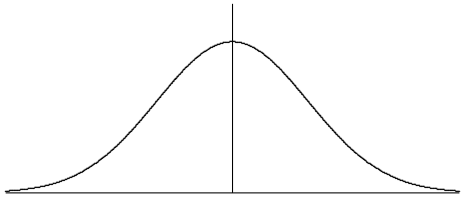
The normal random variable is the commonest of all in applications, and the most important. There is a theorem called the *central limit theorem* which says that, for virtually any random variable X which is not too bizarre, if you take the sum (or the average) of n independent random variables with the same distribution as X, the result will be approximately normal, and will become more and more like a normal variable as n grows. This partly explains why a random variable affected by many independent factors, like a man's height, has an approximately normal distribution.

More precisely, if n is large, then a $\mathrm{Bin}(n,p)$ random variable is well approximated by a normal random variable with the same expected value np and the same variance npq. (If you are approximating any discrete random variable by a continuous one, you should make a "continuity correction" – see the next section for details and an example.)

The p.d.f. of the random variable $X \sim N(\mu, \sigma^2)$ is given by the formula

$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}}\, \mathrm{e}^{-(x-\mu)^2/2\sigma^2}.$$

We have $E(X) = \mu$ and $\mathrm{Var}(X) = \sigma^2$. The picture below shows the graph of this function for $\mu = 0$, the familiar 'bell-shaped curve'.



The c.d.f. of X is obtained as usual by integrating the p.d.f. However, it is not possible to write the integral of this function (which, stripped of its constants, is e^{-x^2}) in terms of 'standard' functions. So there is no alternative but to make tables of its values.

The crucial fact that means that we don't have to tabulate the function for all values of μ and σ is the following:

Proposition 3.6 *If $X \sim N(\mu, \sigma^2)$, and $Y = (X-\mu)/\sigma$, then $Y \sim N(0,1)$.*

So we only need tables of the c.d.f. for $N(0,1)$ – this is the so-called *standard normal random variable* – and we can find the c.d.f. of any normal random variable. The c.d.f. of the standard normal is given in Table 4 of the *New Cambridge Statistical Tables* [1]. The function is called Φ in the tables.

For example, suppose that $X \sim N(6,25)$.What is the probability that $X \leq 8$? Putting $Y = (X-6)/5$, so that $Y \sim N(0,1)$, we find that $X \leq 8$ if and only if $Y \leq (8-6)/5 = 0.4$. From the tables, the probability of this is $\Phi(0.4) = 0.6554$.

The p.d.f. of a standard normal r.v. Y is symmetric about zero. This means that, for any positive number c,

$$\Phi(-c) = P(Y \leq -c) = P(Y \geq c) = 1 - P(Y \leq c) = 1 - \Phi(c).$$

So it is only necessary to tabulate the function for positive values of its argument.

So, if $X \sim N(6,25)$ and $Y = (X-6)/5$ as before, then

$$P(X \leq 3) = P(Y \leq -0.6) = 1 - P(Y \leq 0.6) = 1 - 0.7257 = 0.2743.$$

3.9 On using tables

We end this section with a few comments about using tables, not tied particularly to the normal distribution (though most of the examples will come from there).

Interpolation

Any table is limited in the number of entries it contains. Tabulating something with the input given to one extra decimal place would make the table ten times as bulky! Interpolation can be used to extend the range of values tabulated.

Suppose that some function F is tabulated with the input given to three places of decimals. It is probably true that F is changing at a roughly constant rate between, say, 0.28 and 0.29. So $F(0.283)$ will be about three-tenths of the way between $F(0.28)$ and $F(0.29)$.

For example, if Φ is the c.d.f. of the normal distribution, then $\Phi(0.28) = 0.6103$ and $\Phi(0.29) = 0.6141$, so $\Phi(0.283) = 0.6114$. (Three-tenths of 0.0038 is 0.0011.)

Using tables in reverse

This means, if you have a table of values of F, use it to find x such that $F(x)$ is a given value c. Usually, c won't be in the table and we have to interpolate between values x_1 and x_2, where $F(x_1)$ is just less than c and $F(x_2)$ is just greater.

For example, if Φ is the c.d.f. of the normal distribution, and we want the upper quartile, then we find from tables $\Phi(0.67) = 0.7486$ and $\Phi(0.68) = 0.7517$, so the required value is about 0.6745 (since $0.0014/0.0031 = 0.45$).

In this case, the percentile points of the standard normal r.v. are given in Table 5 of the *New Cambridge Statistical Tables* [1], so you don't need to do this. But you will find it necessary in other cases.

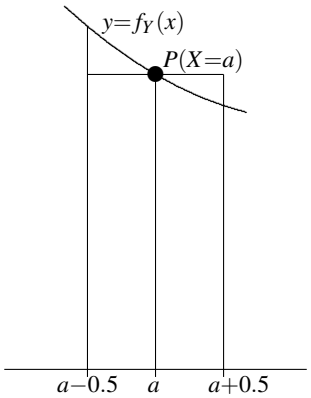
Continuity correction

Suppose we know that a discrete random variable X is well approximated by a continuous random variable Y. We are given a table of the c.d.f. of Y and want to find information about X. For example, suppose that X takes integer values and we want to find $P(a \leq X \leq b)$, where a and b are integers. This probability is equal to

$$P(X = a) + P(x = a+1) + \cdots + P(X = b).$$

To say that X can be approximated by Y means that, for example, $P(X = a)$ is approximately equal to $f_Y(a)$, where f_Y is the p.d.f. of Y. This is equal to the area

of a rectangle of height $f_Y(a)$ and base 1 (from $a-0.5$ to $a+0.5$). This in turn is, to a good approximation, the area under the curve $y = f_Y(x)$ from $x = a-0.5$ to $x = a+0.5$, since the pieces of the curve above and below the rectangle on either side of $x = a$ will approximately cancel. Similarly for the other values.



Adding all these pieces. we find that $P(a \leq X \leq b)$ is approximately equal to the area under the curve $y = f_Y(x)$ from $x = a-0.5$ to $x = b+0.5$. This area is given by $F_Y(b+0.5) - F_Y(a-0.5)$, since F_Y is the integral of f_Y. Said otherwise, this is $P(a-0.5 \leq Y \leq b+0.5)$.

We summarise the *continuity correction*:

> Suppose that the discrete random variable X, taking integer values, is approximated by the continuous random variable Y. Then
>
> $$P(a \leq X \leq b) \approx P(a-0.5 \leq Y \leq b+0.5) = F_Y(b+0.5) - F_Y(a-0.5).$$

(Here, $\approx$ means "approximately equal".) Similarly, for example, $P(X \leq b) \approx P(Y \leq b+0.5)$, and $P(X \geq a) \approx P(Y \geq a-0.5)$.

Example The probability that a light bulb will fail in a year is 0.75, and light bulbs fail independently. If 192 bulbs are installed, what is the probability that the number which fail in a year lies between 140 and 150 inclusive?

Solution Let X be the number of light bulbs which fail in a year. Then $X \sim \text{Bin}(192, 3/4)$, and so $E(X) = 144$, $\text{Var}(X) = 36$. So X is approximated by $Y \sim N(144, 36)$, and

$$P(140 \leq X \leq 150) \approx P(139.5 \leq Y \leq 150.5)$$

by the continuity correction.

Let $Z = (Y - 144)/6$. Then $Z \sim N(0,1)$, and

$$\begin{aligned} P(139.5 \leq Y \leq 150.5) &= P\left(\frac{139.5-144}{6} \leq Z \leq \frac{150.5-144}{6}\right) \\ &= P(-0.75 \leq Z \leq 1.083) \\ &= 0.8606 - 0.2268 \qquad \text{(from tables)} \\ &= 0.6338. \end{aligned}$$

3.10 Worked examples

Question I roll a fair die twice. Let the random variable X be the maximum of the two numbers obtained, and let Y be the modulus of their difference (that is, the value of Y is the larger number minus the smaller number).

(a) Write down the joint p.m.f. of (X,Y).

(b) Write down the p.m.f. of X, and calculate its expected value and its variance.

(c) Write down the p.m.f. of Y, and calculate its expected value and its variance.

(d) Are the random variables X and Y independent?

Solution (a)

		Y					
		0	1	2	3	4	5
X	1	$\frac{1}{36}$	0	0	0	0	0
	2	$\frac{1}{36}$	$\frac{2}{36}$	0	0	0	0
	3	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{2}{36}$	0	0	0
	4	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{2}{36}$	$\frac{2}{36}$	0	0
	5	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{2}{36}$	$\frac{2}{36}$	$\frac{2}{36}$	0
	6	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{2}{36}$	$\frac{2}{36}$	$\frac{2}{36}$	$\frac{2}{36}$

The best way to produce this is to write out a 6×6 table giving all possible values for the two throws, work out for each cell what the values of X and Y are, and then count the number of occurrences of each pair. For example: $X = 5$, $Y = 2$ can occur in two ways: the numbers thrown must be $(5,3)$ or $(3,5)$.

(b) Take row sums:

x	1	2	3	4	5	6
$P(X = x)$	$\frac{1}{36}$	$\frac{3}{36}$	$\frac{5}{36}$	$\frac{7}{36}$	$\frac{9}{36}$	$\frac{11}{36}$

Hence in the usual way

$$E(X) = \frac{161}{36}, \qquad \mathrm{Var}(X) = \frac{2555}{1296}.$$

(c) Take column sums:

y	0	1	2	3	4	5
$P(Y=y)$	$\frac{6}{36}$	$\frac{10}{36}$	$\frac{8}{36}$	$\frac{6}{36}$	$\frac{4}{36}$	$\frac{2}{36}$

and so

$$E(Y) = \frac{35}{18}, \qquad \mathrm{Var}(Y) = \frac{665}{324}.$$

(d) No: e.g. $P(X=1,Y=2)=0$ but $P(X=1)\cdot P(Y=2) = \frac{8}{1296}$.

Question An archer shoots an arrow at a target. The distance of the arrow from the centre of the target is a random variable X whose p.d.f. is given by

$$f_X(x) = \begin{cases} (3+2x-x^2)/9 & \text{if } x \le 3, \\ 0 & \text{if } x > 3. \end{cases}$$

The archer's score is determined as follows:

Distance	$X < 0.5$	$0.5 \le X < 1$	$1 \le X < 1.5$	$1.5 \le X < 2$	$X \ge 2$
Score	10	7	4	1	0

Construct the probability mass function for the archer's score, and find the archer's expected score.

Solution First we work out the probability of the arrow being in each of the given bands:

$$\begin{aligned} P(X<0.5) = F_X(0.5) - F_X(0) &= \int_0^{0.5} \frac{3+2x-x^2}{9}\,\mathrm{d}x \\ &= \left[\frac{9x+3x^2-x^3}{27}\right]_0^{1/2} \\ &= \frac{41}{216}. \end{aligned}$$

Similarly we find that $P(0.5 \le X < 1) = 47/216$, $P(1 \le X < 1.5) = 47/216$, $P(1.5 \le X < 2) = 41/216$, and $P(X \ge 2) = 40/216$. So the p.m.f. fot the archer's score S is

s	0	1	4	7	10
$P(S=s)$	$\frac{40}{216}$	$\frac{41}{216}$	$\frac{47}{216}$	$\frac{47}{216}$	$\frac{41}{216}$

Hence

$$E(S) = \frac{41 + 47 \cdot 4 + 47 \cdot 7 + 41 \cdot 10}{216} = \frac{121}{27}.$$

Question Let T be the lifetime in years of new bus engines. Suppose that T is continuous with probability density function

$$f_T(x) = \begin{cases} 0 & \text{for } x < 1 \\ \dfrac{d}{x^3} & \text{for } x > 1 \end{cases}$$

for some constant d.

(a) Find the value of d.

(b) Find the mean and median of T.

(c) Suppose that 240 new bus engines are installed at the same time, and that their lifetimes are independent. By making an appropriate approximation, find the probability that at most 10 of the engines last for 4 years or more.

Solution (a) The integral of $f_T(x)$, over the support of T, must be 1. That is,

$$\begin{aligned} 1 &= \int_1^\infty \frac{d}{x^3}\,\mathrm{d}x \\ &= \left[\frac{-d}{2x^2}\right]_1^\infty \\ &= d/2, \end{aligned}$$

so $d = 2$.

(b) The c.d.f. of T is obtained by integrating the p.d.f.; that is, it is

$$F_T(x) = \begin{cases} 0 & \text{for } x < 1 \\ 1 - \dfrac{1}{x^2} & \text{for } x > 1 \end{cases}$$

The mean of T is

$$\int_1^\infty x f_T(x)\,\mathrm{d}x = \int_1^\infty \frac{2}{x^2}\,\mathrm{d}x = 2.$$

The median is the value m such that $F_T(m) = 1/2$. That is, $1 - 1/m^2 = 1/2$, or $m = \sqrt{2}$.

(c) The probability that an engine lasts for four years or more is

$$1 - F_T(4) = 1 - \left(1 - \left(\frac{1}{4}\right)^2\right) = \frac{1}{16}.$$

So, if 240 engines are installed, the number which last for four years or more is a binomial random variable $X \sim \text{Bin}(240, 1/16)$, with expected value $240 \times (1/16) = 15$ and variance $240 \times (1/16) \times (15/16) = 225/16$.

We approximate X by $Y \sim N(15, (15/4)^2)$. Using the continuity correction, $P(X \leq 10) \approx P(Y \leq 10.5)$.

Now, if $Z = (Y - 15)/(15/4)$, then $Z \sim N(0,1)$, and

$$\begin{aligned} P(Y \leq 10.5) &= P(Z \leq -1.2) \\ &= 1 - P(Z \leq 1.2) \\ &= 0.1151 \end{aligned}$$

using the table of the standard normal distribution.

Note that we start with the continuous random variable T, move to the discrete random variable X, and then move on to the continuous random variables Y and Z, where finally Z is standard normal and so is in the tables.

A true story The answer to the question at the end of the last chapter: As the students in the class obviously knew, the class included a pair of twins! (The twins were Leo and Willy Moser, who both had successful careers as mathematicians.)

But what went wrong with our argument for the Birthday Paradox? We assumed (without saying so) that the birthdays of the people in the room were independent; but of course the birthdays of twins are clearly not independent!

Chapter 4

More on joint distribution

We have seen the joint p.m.f. of two discrete random variables X and Y, and we have learned what it means for X and Y to be independent. Now we examine this further to see measures of non-independence and conditional distributions of random variables.

4.1 Covariance and correlation

In this section we consider a pair of discrete random variables X and Y. Remember that X and Y are independent if

$$P(X = a_i, Y = b_j) = P(X = a_i) \cdot P(Y = b_j)$$

holds for any pair (a_i, b_j) of values of X and Y. We introduce a number (called the covariance of X and Y) which gives a measure of how far they are from being independent.

Look back at the proof of Theorem 21(b), where we showed that if X and Y are independent then $\mathrm{Var}(X+Y) = \mathrm{Var}(X) + \mathrm{Var}(Y)$. We found that, in any case,

$$\mathrm{Var}(X+Y) = \mathrm{Var}(X) + \mathrm{Var}(Y) + 2(E(XY) - E(X)E(Y)),$$

and then proved that if X and Y are independent then $E(XY) = E(X)E(Y)$, so that the last term is zero.

Now we define the *covariance* of X and Y to be $E(XY) - E(X)E(Y)$. We write $\mathrm{Cov}(X,Y)$ for this quantity. Then the argument we had earlier shows the following:

Theorem 4.1 *(a)* $\mathrm{Var}(X+Y) = \mathrm{Var}(X) + \mathrm{Var}(Y) + 2\,\mathrm{Cov}(X,Y)$.

(b) If X and Y are independent, then $\mathrm{Cov}(X,Y) = 0$.

In fact, a more general version of (a), proved by the same argument, says that

$$\mathrm{Var}(aX+bY) = a^2\,\mathrm{Var}(X) + b^2\,\mathrm{Var}(Y) + 2ab\,\mathrm{Cov}(X,Y). \tag{4.1}$$

Another quantity closely related to covariance is the *correlation coefficient*, $\mathrm{corr}(X,Y)$, which is just a "normalised" version of the covariance. It is defined as follows:

$$\mathrm{corr}(X,Y) = \frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{Var}(X)\,\mathrm{Var}(Y)}}.$$

The point of this is the first part of the following theorem.

Theorem 4.2 *Let X and Y be random variables. Then*

(a) $-1 \le \mathrm{corr}(X,Y) \le 1$*;*

(b) if X and Y are independent, then $\mathrm{corr}(X,Y) = 0$*;*

(c) if $Y = mX + c$ for some constants $m \neq 0$ and c, then $\mathrm{corr}(X,Y) = 1$ *if* $m > 0$*, and* $\mathrm{corr}(X,Y) = -1$ *if* $m < 0$*.*

The proof of the first part is optional: see the end of this section. But note that this is another check on your calculations: if you calculate a correlation coefficient which is bigger than 1 or smaller than -1, then you have made a mistake. Part (b) follows immediately from part (b) of the preceding theorem.

For part (c), suppose that $Y = mX + c$. Let $E(X) = \mu$ and $\mathrm{Var}(X) = \alpha$, so that $E(X^2) = \mu^2 + \alpha$. Now we just calculate everything in sight.

$$\begin{aligned}
E(Y) &= E(mX+c) = mE(X) + c = m\mu + c \\
E(Y^2) &= E(m^2X^2 + 2mcX + c^2) = m^2(\mu^2+\alpha) + 2mc\mu + c^2 \\
\mathrm{Var}(Y) &= E(Y^2) - E(Y)^2 = m^2\alpha \\
E(XY) &= E(mX^2 + cX) = m(\mu^2+\alpha) + c\mu; \\
\mathrm{Cov}(X,Y) &= E(XY) - E(X)E(Y) = m\alpha \\
\mathrm{corr}(X,Y) &= \mathrm{Cov}(X,Y)/\sqrt{\mathrm{Var}(X)\,\mathrm{Var}(Y)} = m\alpha/\sqrt{m^2\alpha^2} \\
&= \begin{cases} +1 & \text{if } m > 0, \\ -1 & \text{if } m < 0. \end{cases}
\end{aligned}$$

Thus the correlation coefficient is a measure of the extent to which the two variables are related. It is $+1$ if Y increases linearly with X; 0 if there is no relation between them; and -1 if Y decreases linearly as X increases. More generally, a positive correlation indicates a tendency for larger X values to be associated with larger Y values; a negative value, for smaller X values to be associated with larger Y values.

Example I have two red pens, one green pen, and one blue pen, and I choose two pens without replacement. Let X be the number of red pens that I choose and Y the number of green pens. Then the joint p.m.f. of X and Y is given by the following table:

		Y	
		0	1
X	0	0	$\frac{1}{6}$
	1	$\frac{1}{3}$	$\frac{1}{3}$
	2	$\frac{1}{6}$	0

From this we can calculate the marginal p.m.f. of X and of Y and hence find their expected values and variances:

$$\begin{aligned} E(X) &= 1, & \mathrm{Var}(X) &= 1/3, \\ E(Y) &= 1/2, & \mathrm{Var}(Y) &= 1/4. \end{aligned}$$

Also, $E(XY) = 1/3$, since the sum

$$E(XY) = \sum_{i,j} a_i b_j P(X = a_i, Y = b_j)$$

contains only one term where all three factors are non-zero. Hence

$$\mathrm{Cov}(X,Y) = 1/3 - 1/2 = -1/6,$$

and

$$\mathrm{corr}(X,Y) = \frac{-1/6}{\sqrt{1/12}} = -\frac{1}{\sqrt{3}}.$$

The negative correlation means that small values of X tend to be associated with larger values of Y. Indeed, if $X = 0$ then Y must be 1, and if $X = 2$ then Y must be 0, but if $X = 1$ then Y can be either 0 or 1.

Example We have seen that if X and Y are independent then $\mathrm{Cov}(X,Y) = 0$. However, it doesn't work the other way around. Consider the following joint p.m.f.

		Y		
		-1	0	1
X	-1	$\frac{1}{5}$	0	$\frac{1}{5}$
	0	0	$\frac{1}{5}$	0
	1	$\frac{1}{5}$	0	$\frac{1}{5}$

Now calculation shows that $E(X) = E(Y) = E(XY) = 0$, so $\mathrm{Cov}(X,Y) = 0$. But X and Y are not independent: for $P(X = -1) = 2/5$, $P(Y = 0) = 1/5$, but $P(X = -1, Y = 0) = 0$.

We call two random variables X and Y *uncorrelated* if $\mathrm{Cov}(X,Y) = 0$ (in other words, if $\mathrm{corr}(X,Y) = 0$). So we can say:

> Independent random variables are uncorrelated, but uncorrelated random variables need not be independent.

Here is the proof that the correlation coefficient lies between -1 and 1. Clearly this is exactly equivalent to proving that its square is at most 1, that is, that

$$\mathrm{Cov}(X,Y)^2 \leq \mathrm{Var}(X) \cdot \mathrm{Var}(Y).$$

This depends on the following fact:

> Let p, q, r be real numbers with $p > 0$. Suppose that $px^2 + 2qx + r \geq 0$ for all real numbers x. Then $q^2 \leq pr$.

For, when we plot the graph $y = px^2 + 2qx + r$, we get a parabola; the hypothesis means that this parabola never goes below the X-axis, so that either it lies entirely above the axis, or it touches it in one point. This means that the quadratic equation $px^2 + 2qx + r = 0$ either has no real roots, or has two equal real roots. From high-school algebra, we know that this means that $q^2 \leq pr$.

Now let $p = \mathrm{Var}(X)$, $q = \mathrm{Cov}(X,Y)$, and $r = \mathrm{Var}(Y)$. Equation (4.1) shows that

$$px^2 + 2qx + r = \mathrm{Var}(xX + Y).$$

(Note that x is an arbitrary real number here and has no connection with the random variable X!)

Since the variance of a random variable is never negative, we see that $px^2 + 2qx + r \geq 0$ for all choices of x. Now our argument above shows that $q^2 \leq pr$, that is, $\mathrm{Cov}(X,Y)^2 \leq \mathrm{Var}(X) \cdot \mathrm{Var}(Y)$, as required.

4.2 Conditional random variables

Remember that the *conditional probability* of event B given event A is $P(B \mid A) = P(A \cap B)/P(A)$.

Suppose that X is a discrete random variable. Then the conditional probability that X takes a certain value a_i, given A, is just

$$P(X = a_i \mid A) = \frac{P(A \text{ holds and } X = a_i)}{P(A)}.$$

This defines the probability mass function of the *conditional random variable* $X \mid A$.

So we can, for example, talk about the *conditional expectation*

$$E(X \mid A) = \sum_i a_i P(X = a_i \mid A).$$

Now the event A might itself be defined by a random variable; for example, A might be the event that Y takes the value b_j. In this case, we have

$$P(X = a_i \mid Y = b_j) = \frac{P(X = a_i, Y = b_j)}{P(Y = b_j)}.$$

In other words, we have taken the column of the joint p.m.f. table of X and Y corresponding to the value $Y = b_j$. The sum of the entries in this column is just $P(Y = b_j)$, the marginal distribution of Y. We divide the entries in the column by this value to obtain a new distribution of X (whose probabilities add up to 1).

In particular, we have

$$E(X \mid Y = b_j) = \sum_i a_i P(X = a_i \mid Y = b_j).$$

Example I have two red pens, one green pen, and one blue pen, and I choose two pens without replacement. Let X be the number of red pens that I choose and Y the number of green pens. Then the joint p.m.f. of X and Y is given by the following table:

		Y	
		0	1
X	0	0	$\frac{1}{6}$
	1	$\frac{1}{3}$	$\frac{1}{3}$
	2	$\frac{1}{6}$	0

In this case, the conditional distributions of X corresponding to the two values of Y are as follows:

a	0	1	2
$P(X = a \mid Y = 0)$	0	$\frac{2}{3}$	$\frac{1}{3}$

a	0	1	2
$P(X = a \mid Y = 1)$	$\frac{1}{3}$	$\frac{2}{3}$	0

We have

$$E(X \mid Y = 0) = \frac{4}{3}, \qquad E(X \mid Y = 1) = \frac{2}{3}.$$

If we know the *conditional expectation* of X for all values of Y, we can find the expected value of X:

Proposition 4.3 $E(X) = \sum_j E(X \mid Y = b_j) P(Y = b_j)$.

Proof: $E(X) = \sum_i a_i P(X = a_i)$

$$\begin{aligned} &= \sum_i a_i \sum_j P(X = a_i \mid Y = b_j) P(Y = b_j) \\ &= \sum_j \left(\sum_i a_i P(X = a_i \mid Y = b_j \right) P(Y = b_j) \\ &= \sum_j E(X \mid Y = b_j) P(Y = b_j). \end{aligned}$$

In the above example, we have

$$\begin{aligned} E(X) &= E(X \mid Y = 0)P(Y = 0) + E(X \mid Y = 1)P(Y = 1) \\ &= (4/3) \times (1/2) + (2/3) \times (1/2) \\ &= 1. \end{aligned}$$

Example Let us revisit the geometric random variable and calculate its expected value. Recall the situation: I have a coin with probability p of showing heads; I toss it repeatedly until heads appears for the first time; X is the number of tosses.

Let Y be the Bernoulli random variable whose value is 1 if the result of the first toss is heads, 0 if it is tails. If $Y = 1$, then we stop the experiment then and there; so if $Y = 1$, then necessarily $X = 1$, and we have $E(X \mid Y = 1) = 1$. On the other hand, if $Y = 0$, then the sequence of tosses from that point on has the same distribution as the original experiment; so $E(X \mid Y = 0) = 1 + E(X)$ (the 1 counting the first toss). So

$$\begin{aligned} E(X) &= E(X \mid Y = 0)P(Y = 0) + E(X \mid Y = 1)P(Y = 1) \\ &= (1 + E(X)) \cdot q + 1 \cdot p \\ &= E(X)(1 - p) + 1; \end{aligned}$$

rearranging this equation, we find that $E(X) = 1/p$, confirming our earlier value.

In Proposition 2.1, we saw that independence of events can be characterised in terms of conditional probabilities: A and B are independent if and only if they satisfy $P(A \mid B) = P(A)$. A similar result holds for independence of random variables:

Proposition 4.4 *Let X and Y be discrete random variables. Then X and Y are independent if and only if, for any values a_i and b_j of X and Y respectively, we have*

$$P(X = a_i \mid Y = b_j) = P(X = a_i).$$

This is obtained by applying Proposition 15 to the events $X = a_i$ and $Y = b_j$. It can be stated in the following way: X and Y are independent if the conditional p.m.f. of $X \mid (Y = b_j)$ is equal to the p.m.f. of X, for any value b_j of Y.

4.3 Joint distribution of continuous r.v.s

For continuous random variables, the covariance and correlation can be defined by the same formulae as in the discrete case; and Equation (4.1) remains valid. But we have to examine what is meant by independence for continuous random variables. The formalism here needs even more concepts from calculus than we have used before: functions of two variables, partial derivatives, double integrals. I assume that this is unfamiliar to you, so this section will be brief and can mostly be skipped.

Let X and Y be continuous random variables. The *joint cumulative distribution function* of X and Y is the function $F_{X,Y}$ of two real variables given by

$$F_{X,Y}(x,y) = P(X \leq x, Y \leq y).$$

We define X and Y to be *independent* if $P(X \leq x, Y \leq y) = P(X \leq x) \cdot P(Y \leq y)$, for any x and y, that is, $F_{X,Y}(x,y) = F_X(x) \cdot F_Y(y)$. (Note that, just as in the one-variable case, X is part of the name of the function, while x is the argument of the function.)

The *joint probability density function* of X and Y is

$$f_{X,Y}(x,y) = \frac{\partial^2}{\partial x \partial y} F_{X,Y}(x,y).$$

In other words, differentiate with respect to x keeping y constant, and then differentiate with respect to y keeping x constant (or the other way round: the answer is the same for all functions we consider.)

The probability that the pair of values of (X,Y) corresponds to a point in some region of the plane is obtained by taking the double integral of $f_{X,Y}$ over that region. For example,

$$P(a \leq X \leq b, c \leq Y \leq d) = \int_c^d \int_a^b f_{X,Y}(x,y) \mathrm{d}x \, \mathrm{d}y$$

(the right hand side means, integrate with respect to x between a and b keeping y fixed; the result is a function of y; integrate this function with respect to y from c to d.)

The marginal p.d.f. of X is given by

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x,y) \mathrm{d}y,$$

and the marginal p.d.f. of Y is similarly

$$f_Y(y) = \int_{-\infty}^{\infty} f_{X,Y}(x,y) \mathrm{d}x.$$

Then the conditional p.d.f. of $X \mid (Y = b)$ is

$$f_{X|(Y=b)}(x) = \frac{f_{X,Y}(x,b)}{f_Y(b)}.$$

The expected value of XY is, not surprisingly,

$$E(XY) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} xy f_{X,Y}(x,y)\mathrm{d}x\,\mathrm{d}y,$$

and then as in the discrete case

$$\mathrm{Cov}(X,Y) = E(XY) - E(X)E(Y), \qquad \mathrm{corr}(X,Y) = \frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{Var}(X)\,\mathrm{Var}(Y)}}.$$

Finally, and importantly,

> The continuous random variables X and Y are independent if and only if
>
> $$f_{X,Y}(x,y) = f_X(x) \cdot f_Y(y).$$

As usual this holds if and only if the *conditional* p.d.f. of $X \mid (Y = b)$ is equal to the *marginal* p.d.f. of X, for any value b. Also, if X and Y are independent, then $\mathrm{Cov}(X,Y) = \mathrm{corr}(X,Y) = 0$ (but not conversely!).

4.4 Transformation of random variables

If a continuous random variable Y is a function of another r.v. X, we can find the distribution of Y in terms of that of X.

Example Let X and Y be random variables. Suppose that $X \sim U[0,4]$ (uniform on $[0,4]$) and $Y = \sqrt{X}$. What is the support of Y? Find the cumulative distribution function and the probability density function of Y.

Solution (a) The support of X is $[0,4]$, and $Y = \sqrt{X}$, so the support of Y is $[0,2]$.

(b) We have $f_X(x) = x/4$ for $0 \le x \le 4$. Now

$$\begin{aligned} F_Y(y) &= P(Y \le y) \\ &= P(X \le y^2) \\ &= F_X(y^2) \\ &= y^2/4 \end{aligned}$$

for $0 \leq y \leq 2$; of course $F_Y(y) = 0$ for $y < 0$ and $F_Y(y) = 1$ for $y > 2$. (Note that $Y \leq y$ if and only if $X \leq y^2$, since $Y = \sqrt{X}$.)

(c) We have

$$f_Y(y) = \frac{\mathrm{d}}{\mathrm{d}y} F_Y(y) = \begin{cases} y/2 & \text{if } 0 \leq y \leq 2, \\ 0 & \text{otherwise.} \end{cases}$$

The argument in (b) is the key. If we know Y as a function of X, say $Y = g(X)$, where g is an increasing function, then the event $Y \leq y$ is the same as the event $X \leq h(Y)$, where h is the *inverse function* of g. This means that $y = g(x)$ if and only if $x = h(y)$. (In our example, $g(x) = \sqrt{x}$, and so $h(y) = y^2$.) Thus

$$F_Y(y) = F_X(h(y)),$$

and so, by the Chain Rule,

$$f_Y(y) = f_X(h(y))h'(y),$$

where h' is the derivative of h. (This is because $f_X(x)$ is the derivative of $F_X(x)$ with respect to its argument x, and the Chain Rule says that if $x = h(y)$ we must multiply by $h'(y)$ to find the derivative with respect to y.)

Applying this formula in our example we have

$$f_Y(y) = \frac{1}{4} \cdot 2y = \frac{y}{2}$$

for $0 \leq y \leq 2$, since the p.d.f. of X is $f_X(x) = 1/4$ for $0 \leq x \leq 4$.

Here is a formal statement of the result.

Theorem 4.5 *Let X be a continuous random variable. Let g be a real function which is either strictly increasing or strictly decreasing on the support of X, and which is differentiable there. Let $Y = g(X)$. Then*

(a) the support of Y is the image of the support of X under g;

(b) the p.d.f. of Y is given by $f_Y(y) = f_X(h(y))|h'(y)|$, where h is the inverse function of g.

For example, here is the proof of Proposition 3.6: *if $X \sim N(\mu, \sigma^2)$ and $Y = (X - \mu)/\sigma$, then $Y \sim N(0,1)$.* Recall that

$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} \, \mathrm{e}^{-(x-\mu)^2/2\sigma^2}.$$

We have $Y = g(X)$, where $g(x) = (x-\mu)/\sigma$; this function is everywhere strictly increasing (the graph is a straight line with slope $1/\sigma$), and the inverse function is $x = h(y) = \sigma y + \mu$. Thus, $h'(y) = \sigma$, and

$$f_Y(y) = f_X(\sigma y + \mu) \cdot \sigma = \frac{1}{\sqrt{2\pi}} \, \mathrm{e}^{-y^2/2},$$

the p.d.f. of a standard normal variable.

However, rather than remember this formula, together with the conditions for its validity, I recommend going back to the argument we used in the example.

If the transforming function g is not monotonic (that is, not either increasing or decreasing), then life is a bit more complicated. For example, if X is a random variable taking both positive and negative values, and $Y = X^2$, then a given value y of Y could arise from either of the values $\sqrt{y}$ and $-\sqrt{y}$ of X, so we must work out the two contributions and add them up.

Example $X \sim N(0,1)$ and $Y = X^2$. Find the p.d.f. of Y.

The p.d.f. of X is $(1/\sqrt{2\pi})\mathrm{e}^{-x^2/2}$. Let $\Phi(x)$ be its c.d.f., so that $P(X \le x) = \Phi(x)$, and

$$\Phi'(x) = \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-x^2/2}.$$

Now $Y = X^2$, so $Y \le y$ if and only if $-\sqrt{y} \le X \le \sqrt{y}$. Thus

$$\begin{aligned} F_Y(y) &= P(Y \le y) \\ &= P(-\sqrt{y} \le X \le \sqrt{y}) \\ &= \Phi(\sqrt{y}) - \Phi(-\sqrt{y}) \\ &= \Phi(\sqrt{y} - (1 - \Phi(\sqrt{y})) \qquad \text{(by symmetry of } N(0,1)) \\ &= 2\Phi(\sqrt{y}) - 1. \end{aligned}$$

So

$$\begin{aligned} f_Y(y) &= \frac{\mathrm{d}}{\mathrm{d}y} F_Y(y) \\ &= 2\Phi'(\sqrt{y}) \cdot \frac{1}{2\sqrt{y}} \qquad \text{(by the Chain Rule)} \\ &= \frac{1}{\sqrt{2\pi y}} \mathrm{e}^{-y/2}. \end{aligned}$$

Of course, this is valid for $y > 0$; for $y < 0$, the p.d.f. is zero.

Note the 2 in the line labelled "by the Chain Rule". If you blindly applied the formula of Theorem 4.5, using $h(y) = \sqrt{y}$, you would not get this 2; it arises from the fact that, since $Y = X^2$, each value of Y corresponds to two values of X (one positive, one negative), and each value gives the same contribution, by the symmetry of the p.d.f. of X.

4.5 Worked examples

Question Two numbers X and Y are chosen independently from the uniform distribution on the unit interval $[0,1]$. Let Z be the maximum of the two numbers. Find the p.d.f. of Z, and hence find its expected value, variance and median.

Solution The c.d.f.s of X and Y are identical, that is,

$$F_X(x) = F_Y(x) = \begin{cases} 0 & \text{if } x < 0, \\ x & \text{if } 0 < x < 1, \\ 1 & \text{if } x > 1. \end{cases}$$

(The variable can be called x in both cases; its name doesn't matter.)

The key to the argument is to notice that

$$Z = \max(X,Y) \leq x \quad \text{if and only if} \quad X \leq x \text{ and } Y \leq x.$$

(For, if both X and Y are smaller than a given value x, then so is their maximum; but if at least one of them is greater than x, then again so is their maximum.) For $0 \leq x \leq 1$, we have $P(X \leq x) = P(Y \leq x) = x$; by independence,

$$P(X \leq x \text{ and } Y \leq x) = x \cdot x = x^2.$$

Thus $P(Z \leq x) = x^2$. Of course this probability is 0 if $x < 0$ and is 1 if $x > 1$. So the c.d.f. of Z is

$$F_Z(x) = \begin{cases} 0 & \text{if } x < 0, \\ x^2 & \text{if } 0 < x < 1, \\ 1 & \text{if } x > 1. \end{cases}$$

The median of Z is the value of m such that $F_Z(m) = 1/2$, that is $m^2 = 1/2$, or $m = 1/\sqrt{2}$.

We obtain the p.d.f. of Z by differentiating:

$$f_Z(x) = \begin{cases} 2x & \text{if } 0 < x < 1, \\ 0 & \text{otherwise.} \end{cases}$$

Then we can find $E(Z)$ and $\mathrm{Var}(Z)$ in the usual way:

$$E(Z) = \int_0^1 2x^2 \mathrm{d}x = \frac{2}{3}, \qquad \mathrm{Var}(Z) = \int_0^1 2x^3 \mathrm{d}x - \left(\frac{2}{3}\right)^2 = \frac{1}{18}.$$

Question I roll a fair die bearing the numbers 1 to 6. If N is the number showing on the die, I then toss a fair coin N times. Let X be the number of heads I obtain.

(a) Write down the p.m.f. for X.

(b) Calculate $E(X)$ without using this information.

Solution (a) If we were given that $N = n$, say, then X would be a binomial $\mathrm{Bin}(n, 1/2)$ random variable. So $P(X = k \mid N = n) = {}^nC_k(1/2)^n$.

By the ToTP,

$$P(X = k) = \sum_{n=1}^{6} P(X = k \mid N = n)P(N = n).$$

Clearly $P(N = n) = 1/6$ for $n = 1, \ldots, 6$. So to find $P(X = k)$, we add up the probability that $X = k$ for a $\mathrm{Bin}(n, 1/2)$ r.v. for $n = k, \ldots, 6$ and divide by 6. (We start at k because you can't get k heads with fewer than k coin tosses!) The answer comes to

k	0	1	2	3	4	5	6
$P(X = k)$	$\frac{63}{384}$	$\frac{120}{384}$	$\frac{99}{384}$	$\frac{64}{384}$	$\frac{29}{384}$	$\frac{8}{384}$	$\frac{1}{384}$

For example,

$$P(X = 4) = \frac{{}^4C_4(1/2)^4 + {}^5C_4(1/2)^5 + {}^6C_4(1/2)^6}{6} = \frac{4 + 10 + 15}{384}.$$

(b) By Proposition 4.3,

$$E(X) = \sum_{n=1}^{6} E(X \mid (N = n))P(N = n).$$

Now if we are given that $N = n$ then, as we remarked, X has a binomial $\mathrm{Bin}(n, 1/2)$ distribution, with expected value $n/2$. So

$$E(X) = \sum_{n=1}^{6} (n/2) \cdot (1/6) = \frac{1 + 2 + 3 + 4 + 5 + 6}{2 \cdot 6} = \frac{7}{4}.$$

Try working it out from the p.m.f. to check that the answer is the same!

Appendix A

Mathematical notation

The Greek alphabet

Mathematicians use the Greek alphabet for an extra supply of symbols. Some, like π, have standard meanings. You don't need to learn this; keep it for reference. Apologies to Greek students: you may not recognise this, but it is the Greek alphabet that mathematicians use!
Pairs that are often confused are zeta and xi, or nu and upsilon, which look alike; and chi and xi, or epsilon and upsilon, which sound alike.

Name	Capital	Lowercase
alpha	A	α
beta	B	β
gamma	Γ	γ
delta	Δ	δ
epsilon	E	ε
zeta	Z	ζ
eta	H	η
theta	Θ	θ
iota	I	ι
kappa	K	κ
lambda	Λ	λ
mu	M	μ
nu	N	ν
xi	Ξ	ξ
omicron	O	o
pi	Π	π
rho	P	ρ
sigma	Σ	σ
tau	T	τ
upsilon	Υ	υ
phi	Φ	ϕ
chi	X	χ
psi	Ψ	ψ
omega	Ω	ω

Numbers

Notation	Meaning	Example
$\mathbb{N}$	Natural numbers	$1,2,3,\ldots$ (some people include 0)
$\mathbb{Z}$	Integers	$\ldots,-2,-1,0,1,2,\ldots$
$\mathbb{R}$	Real numbers	$\frac{1}{2},\sqrt{2},\pi,\ldots$
$\lvert x\rvert$	modulus	$\lvert 2\rvert = 2, \lvert -3\rvert = 3$
a/b or $\frac{a}{b}$	a over b	$12/3 = 4$, $2/4 = 0.5$
$a \mid b$	a divides b	$4 \mid 12$
${}^{m}C_n$ or $\binom{m}{n}$	m choose n	${}^{5}C_2 = 10$
$n!$	n factorial	$5! = 120$
$\sum_{i=a}^{b} x_i$	$x_a + x_{a+1} + \cdots + x_b$ (see section on Summation below)	$\sum_{i=1}^{3} i^2 = 1^2 + 2^2 + 3^2 = 14$
$x \approx y$	x is approximately equal to y	

Sets

Notation	Meaning	Example
$\{\ldots\}$	a set	$\{1,2,3\}$ NOTE: $\{1,2\} = \{2,1\}$
$x \in A$	x is an element of the set A	$2 \in \{1,2,3\}$
$\{x : \ldots\}$ or $\{x \mid \ldots\}$	the set of all x such that $\ldots$	$\{x : x^2 = 4\} = \{-2,2\}$
$\lvert A\rvert$	cardinality of A (number of elements in A)	$\lvert\{1,2,3\}\rvert = 3$
$A \cup B$	A union B (elements in either A or B)	$\{1,2,3\} \cup \{2,4\} = \{1,2,3,4\}$
$A \cap B$	A intersection B (elements in both A and B)	$\{1,2,3\} \cap \{2,4\} = \{2\}$
$A \setminus B$	set difference (elements in A but not B)	$\{1,2,3\} \setminus \{2,4\} = \{1,3\}$
$A \subseteq B$	A is a subset of B (or equal)	$\{1,3\} \subseteq \{1,2,3\}$
A'	complement of A	everything not in A
$\emptyset$	empty set (no elements)	$\{1,2\} \cap \{3,4\} = \emptyset$
(x,y)	ordered pair	NOTE: $(1,2) \neq (2,1)$
$A \times B$	Cartesian product (set of all ordered pairs)	$\{1,2\} \times \{1,3\} = \{(1,1),(2,1),(1,3),(2,3)\}$

Summation

What is it?

Let $a_1, a_2, a_3, \ldots$ be numbers. The notation

$$\sum_{i=1}^{n} a_i$$

(read "sum, from i equals 1 to n, of a_i"), means: add up the numbers $a_1, a_2, \ldots, a_n$; that is,

$$\sum_{i=1}^{n} a_i = a_1 + a_2 + \cdots + a_n.$$

The notation $\sum_{j=1}^{n} a_j$ means exactly the same thing. The variable i or j is called a "dummy variable".

The notation $\sum_{i=1}^{m} a_i$ is not the same, since (if m and n are different) it is telling us to add up a different number of terms.

The sum doesn't have to start at 1. For example,

$$\sum_{i=10}^{20} a_i = a_{10} + a_{11} + \cdots + a_{20}.$$

Sometimes I get lazy and don't bother to write out the values: I just say $\sum_i a_i$ to mean: add up all the relevant values. For example, if X is a discrete random variable, then we say that

$$E(X) = \sum_i a_i P(X = a_i)$$

where the sum is over all i such that a_i is a value of the random variable X.

Manipulation

The following three rules hold.

$$\sum_{i=1}^{n} (a_i + b_i) = \sum_{i=1}^{n} a_i + \sum_{i=1}^{n} b_i. \tag{A.1}$$

Imagine the as and bs written out with $a_1 + b_1$ on the first line, $a_2 + b_2$ on the second line, and so on. The left-hand side says: add the two terms in each line,

and then add up all the results. The right-hand side says: add the first column (all the as) and the second column (all the bs), and then add the results. The answers must be the same.

$$\left(\sum_{i=1}^{n} a_i\right) \cdot \left(\sum_{j=1}^{m} b_j\right) = \sum_{i=1}^{n}\sum_{j=1}^{m} a_i b_j. \tag{A.2}$$

The double sum says add up all these products, for all values of i and j. A simple example shows how it works: $(a_1+a_2)(b_1+b_2) = a_1b_1+a_1b_2+a_2b_1+a_2b_2$.

If in place of numbers, we have functions of x, then we can "differentiate term-by-term":

$$\frac{\mathrm{d}}{\mathrm{d}x}\sum_{i+1}^{n} f_i(x) = \sum_{i+1}^{n}\frac{\mathrm{d}}{\mathrm{d}x} f_i(x). \tag{A.3}$$

The left-hand side says: add up the functions and differentiate the sum. The right says: differentiate each function and add up the derivatives.

Another useful result is the *Binomial Theorem*:

$$(x+y)^n = \sum_{k=0}^{n} {}^nC_k x^{n-k} y^k.$$

Infinite sums

Sometimes we meet infinite sums, which we write as $\sum_{i=1}^{\infty} a_i$ for example. This doesn't just mean "add up infinitely many values", since that is not possible. We need Analysis to give us a definition in general. But sometimes we know the answer another way: for example, if $a_i = ar^{i-1}$, where $-1 < r < 1$, then

$$\sum_{i=1}^{\infty} a_i = a + ar + ar^2 + \cdots = \frac{a}{1-r},$$

using the formula for the sum of the "geometric series". You also need to know the sum of the "exponential series"

$$\sum_{i=0}^{\infty} \frac{x^i}{i!} = 1 + x + \frac{x^2}{2} + \frac{x^3}{6} + \frac{x^4}{24} + \cdots = \mathrm{e}^x.$$

Do the three rules of the preceding section hold? Sometimes yes, sometimes no. In Analysis you will see some answers to this question.

In all the examples you meet in this book, the rules will be valid.

Appendix B

Probability and random variables

Notation

In the table, A and B are events, X and Y are random variables.

Notation	Meaning	Page
$P(A)$	probability of A	3
$P(A \mid B)$	conditional probability of A given B	24
$X = Y$	the values of X and Y are equal	
$X \sim Y$	X and Y have the same distribution (that is, same p.m.f. or same p.d.f.)	41
$E(X)$	expected value of X	41
$\mathrm{Var}(X)$	variance of X	42
$\mathrm{Cov}(X,Y)$	covariance of X and Y	67
$\mathrm{corr}(X,Y)$	correlation coefficient of X and Y	68
$X \mid B$	conditional random variable	70
$X \mid (Y = b)$		71

Bernoulli random variable Bernoulli(p) (p. 48)

- Occurs when there is a single trial with a fixed probability p of success.
- Takes only the values 0 and 1.
- p.m.f. $P(X = 0) = q$, $P(X = 1) = p$, where $q = 1 - p$.
- $E(X) = p$, $\mathrm{Var}(X) = pq$.

Binomial random variable $\mathrm{Bin}(n,p)$ (p. 49)

- Occurs when we are counting the number of successes in n independent trials with fixed probability p of success in each trial, e.g. the number of heads in n coin tosses. Also, sampling with replacement from a population with a proportion p of distinguished elements.
- The sum of n independent Bernoulli(p) random variables.
- Values $0,1,2,\ldots,n$.
- p.m.f. $P(X=k) = {}^nC_k q^{n-k} p^k$ for $0 \le k \le n$, where $q = 1-p$.
- $E(X) = np$, $\mathrm{Var}(X) = npq$.

Hypergeometric random variable $\mathrm{Hg}(n,M,N)$ (p. 51)

- Occurs when we are sampling n elements without replacement from a population of N elements of which M are distinguished.
- Values $0,1,2,\ldots,n$.
- p.m.f. $P(X=k) = ({}^MC_k \cdot {}^{N-M}C_{n-k})/{}^NC_n$.
- $E(X) = n\left(\dfrac{M}{N}\right)$, $\mathrm{Var}(X) = n\left(\dfrac{M}{N}\right)\left(\dfrac{N-M}{N}\right)\left(\dfrac{N-n}{N-1}\right)$.
- Approximately $\mathrm{Bin}(n,M/N)$ if n is small compared to $N, M, N-M$.

Geometric random variable $\mathrm{Geom}(p)$ (p. 52)

- Describes the number of trials up to and including the first success in a sequence of independent Bernoulli trials, e.g. number of tosses until the first head when tossing a coin.
- Values $1,2,\ldots$ (any positive integer).
- p.m.f. $P(X=k) = q^{k-1}p$, where $q = 1-p$.
- $E(X) = 1/p$, $\mathrm{Var}(X) = q/p^2$.

Poisson random variable Poisson(λ) (p. 54)

- Describes the number of occurrences of a random event in a fixed time interval, e.g. the number of fish caught in a day.
- Values $0, 1, 2, \ldots$ (any non-negative integer)
- p.m.f. $P(X = k) = \mathrm{e}^{-\lambda}\lambda^k/k!$.
- $E(X) = \lambda$, $\mathrm{Var}(X) = \lambda$.
- If n is large, p is small, and $np = \lambda$, then $\mathrm{Bin}(n, p)$ is approximately equal to Poisson(λ) (in the sense that the p.m.f.s are approximately equal).

Uniform random variable $U[a, b]$ (p. 58)

- Occurs when a number is chosen at random from the interval $[a, b]$, with all values equally likely.
- p.d.f. $f(x) = \begin{cases} 0 & \text{if } x < a, \\ 1/(b-a) & \text{if } a \le x \le b, \\ 0 & \text{if } x > b. \end{cases}$
- c.d.f. $F(x) = \begin{cases} 0 & \text{if } x < a, \\ (x-a)/(b-a) & \text{if } a \le x \le b, \\ 1 & \text{if } x > b. \end{cases}$
- $E(X) = (a+b)/2$, $\mathrm{Var}(X) = (b-a)^2/12$.

Exponential random variable Exp(λ) (p. 59)

- Occurs in the same situations as the Poisson random variable, but measures the time from now until the first occurrence of the event.
- p.d.f. $f(x) = \begin{cases} 0 & \text{if } x < 0, \\ \lambda\,\mathrm{e}^{-\lambda x} & \text{if } x \ge 0. \end{cases}$
- c.d.f. $F(x) = \begin{cases} 0 & \text{if } x < 0, \\ 1 - \mathrm{e}^{-\lambda x} & \text{if } x \ge 0. \end{cases}$
- $E(X) = 1/\lambda$, $\mathrm{Var}(X) = 1/\lambda^2$.
- However long you wait, the time until the next occurrence has the same distribution.

Normal random variable $N(\mu, \sigma^2)$ (p. 59)

- The limit of the sum (or average) of many independent Bernoulli random variables. This also works for many other types of random variables: this statement is known as the *Central Limit Theorem*.
- p.d.f. $f(x) = \dfrac{1}{\sigma\sqrt{2\pi}}\, \mathrm{e}^{-(x-\mu)^2/2\sigma^2}$.
- No simple formula for c.d.f.; use tables.
- $E(X) = \mu$, $\mathrm{Var}(X) = \sigma^2$.
- For large n, $\mathrm{Bin}(n, p)$ is approximately $N(np, npq)$.
- *Standard normal* $N(0,1)$ is given in the table. If $X \sim N(\mu, \sigma^2)$, then $(X - \mu)/\sigma \sim N(0,1)$.

The c.d.f.s of the Binomial, Poisson, and Standard Normal random variables are tabulated in the *New Cambridge Statistical Tables*, Tables 1, 2 and 4.