

Conditional Probability and Independence

2.1 Conditional Probability

If you throw a fair die, with no clues given about the outcome, the chance of getting a six is $1/6$. But maybe you have been told that this *blue* die, and another *red* die were thrown, and their total score was four. Then you could be sure that the score on the blue die was not six. Similarly, to be told that the total score was ten makes it more likely that the blue die scored six, as all scores lower than four are eliminated. Information about the total score on the two dice can change your original opinion about the chance of a six.

“Conditional probability” addresses this sort of question. You are interested in some event A , whose probability is $P(A)$, and you learn that event B has occurred. What is your new opinion about the chance that A has occurred? The notation for what you seek is $P(A|B)$, read as “The probability of A , conditional upon B ”, or as “The probability of A , given B ”.

To see how this might relate to our previous ideas, consider a repeatable experiment with K equally likely outcomes. The probabilities of events A and B are then found by counting; how to assess the probability of A , given that B occurs? Plainly we can ignore all those experiments in which B does not occur. In the experiments where B does occur, we need to know how often A also occurs, i.e. how often $A \cap B$ occurs. If B occurs in $n(B)$ experiments, and A occurs in $n(A \cap B)$ of these, the ratio

$$n(A \cap B)/n(B)$$

should be close to the conditional probability of A given B , when $n(B)$ is large. But note that

$$\frac{n(A \cap B)}{n(B)} = \frac{n(A \cap B)}{N} \bigg/ \frac{n(B)}{N}$$

where N is the total number of experiments. Since $n(A \cap B)/N \approx P(A \cap B)$ while $n(B)/N \approx P(B)$ when N is large, we make the following definition:

Definition 2.1

$P(A|B) = P(A \cap B)/P(B)$, provided that $P(B) > 0$.

It is an immediate consequence of this definition that

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

Examples often help to settle ideas. Suppose the experiment is selecting one card at random from a well-shuffled deck, so in our model Ω has 52 equally likely outcomes. Let events A, B, C be that the selected card is, respectively, an Ace, a Black card, or a Club. Then $P(A) = 1/13$, $P(B) = 1/2$ and $P(C) = 1/4$, as you should verify. We will check that use of the definition of conditional probability is in tune with our intuitive ideas. Simply using the definition, we see that

$$P(C|B) = P(C \cap B)/P(B) = P(C)/P(B) = 1/2,$$

which makes sense – given we have a Black card, there is a 50 % chance it is a Club. The other way round leads to

$$P(B|C) = P(B \cap C)/P(C) = P(C)/P(C) = 1,$$

which is equally unsurprising – given we have a Club, we can be certain we have a Black card. And

$$P(A|C) = P(A \cap C)/P(C) = (1/52)/(1/4) = 1/13$$

i.e. given we have a Club, there is one chance in thirteen it is the Ace. The definition is working as we would anticipate.

The table in Example 1.7 shows the distribution of the total score when two fair dice are thrown. Write

- A = Total is at least seven, so that $P(A) = 21/36$
- B = Blue die scores at most four, so that $P(B) = 4/6$
- C = Blue die scores at most three, so that $P(C) = 3/6$

(it seldom helps to simplify fractions at this stage). By counting, $P(A \cap B) = 10/36$, and $P(A \cap C) = 6/36$, so the definition leads to

$$P(A|B) = (10/36)/(4/6) = 5/12, \quad P(A|C) = (6/36)/(3/6) = 4/12.$$

The information that the score on the Blue die is at most three, rather than at most four, makes it less likely that the total score is at least seven – as it should.

Theorem 2.2

Given a probability space $(\Omega, \mathcal{F}, P)$, let B be some event with $P(B) > 0$. With the same Ω and $\mathcal{F}$, define the function Q by

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

Then $(\Omega, \mathcal{F}, Q)$ is also a probability space.

Proof

The only things to check are that Q satisfies the three Laws for a probability. First, since $A \cap B \subset B$, then $P(A \cap B) \leq P(B)$, so that $0 \leq Q(A) \leq 1$. Also, it is clear from the definition that $Q(\Omega) = 1$. Finally, let $\{C_i\}$ be a collection of pairwise disjoint events in $\mathcal{F}$; then

$$Q\left(\bigcup_i C_i\right) = \frac{P((\bigcup_i C_i) \cap B)}{P(B)} = \frac{P(\bigcup_i (C_i \cap B))}{P(B)} = \frac{\sum_i P(C_i \cap B)}{P(B)} = \sum_i Q(C_i)$$

which completes the proof. $\square$

This means that any conditional probability $P(.|B)$ is also an ordinary probability. We do not have to use Definition 2.1 to compute a conditional probability, it is sometimes best to work directly.

As an illustration of this, we will compute the probability that a Bridge hand of 13 cards has at least two Aces, conditional on it containing the Ace of Spades. Using this direct approach, our model is that the hand consists of the Ace of Spades, along with 12 other cards that are selected at random from the remaining 51 cards, which comprise three Aces and 48 non-Aces. The complete hand will have at least two Aces when these 12 other cards contain at least one Ace.

This observation has removed the need to use the definition of conditional probability, we just work on the probability space for selecting 12 cards at

random from 51. There are $\binom{51}{12}$ outcomes in this space, and $\binom{48}{12}$ of them have no Aces at all. Thus the ratio $\binom{48}{12}/\binom{51}{12}$ gives the chance of no Ace. We find the chance of at least one Ace by the usual trick of subtracting this value from unity, which gives the answer of 0.5612. The chance we have at least two Aces, given we have the Ace of Spades, is about 56 %.

Suppose, on the other hand, that we want to assess the chance that our hand has at least two Aces, given it has at least one Ace. Many people will expect the answer to be the same as we have just computed. “After all”, they say, “given that the hand has at least one Ace, the Spade Ace is just as likely as any other.” An accurate observation, but not one from which the intended deduction can be drawn. Let’s do it properly.

Write A as the event that the hand has at least two Aces, and B that it has at least one Ace. To find their probabilities, it is easiest to find $P(0)$ and $P(1)$, the chances that a hand has respectively no Aces and just one Ace. Counting again gives the answers. There are $\binom{52}{13}$ possible hands, and $\binom{48}{13}$ of them have no Ace, while $\binom{48}{12}\binom{4}{1}$ have one Ace. This leads to

$$P(0) = 0.3038 \quad \text{and} \quad P(1) = 0.4388.$$

Thus

$$P(B) = 1 - P(0) = 0.6962 \quad \text{and} \quad P(A) = 1 - P(0) - P(1) = 0.2573.$$

Since $A \subset B$, then

$$P(A|B) = P(A)/P(B) = 0.3696$$

when the dust has settled. Given we have at least one Ace, the chance we have at least two Aces is about 37 %.

Intuitively, why should these two answers be so different? Think along the following lines: knowing that the hand has a particular Ace, each of the remaining 12 cards has chance $3/51$ of being an Ace so there will be, on average, $12 \times 3/51 = 36/51 \approx 0.7$ other Aces. Since this comfortably exceeds one half, there is a pretty good chance of at least one Ace. Now look at hands in general. Symmetry dictates that they have one Ace, on average, so if we eliminate the Aceless hands, just one Ace seems rather more likely than two or more. But the statement “Given there is at least one Ace” does no more nor less than eliminate the Aceless hands. Perhaps the result is no real surprise.

Explicit use of the formula $P(A \cap B) = P(A|B)P(B)$ simplifies many calculations. Take an urn that initially has six white balls and four red ones. Balls are withdrawn at random, all balls in the urn being equally likely, and after each drawing, the ball is replaced, along with another of the same colour. What is the probability that the first two balls drawn are red?

Let A mean that the *second* ball drawn is red, and B that the first one is red. We shall find $P(A \cap B)$ using this formula. Plainly $P(B) = 4/10$ and, conditional on B , the urn has six white balls and five red ones when the second ball is selected. Hence $P(A|B) = 5/11$, and the answer to our problem is $(4/10) \times (5/11) = 2/11$.

This notion extends to more than two events:

Theorem 2.3 (Multiplication Rule)

Suppose $\{A_1, A_2, \dots, A_n\}$ are events in the same probability space. Then

$$P\left(\bigcap_{i=1}^n A_i\right) = P(A_1)P(A_2|A_1)P(A_3|A_1 \cap A_2) \cdots P(A_n|A_1 \cap A_2 \cap \cdots \cap A_{n-1}),$$

provided, of course, that $P(A_1 \cap A_2 \cap \cdots \cap A_{n-1}) > 0$.

Proof

Just apply the definition of conditional probability to every term on the right side, and watch the expression collapse to that on the left side. $\square$

To illustrate this result, we will calculate the chance that each of the four players in a game of Bridge is dealt an Ace. Our model is that the 52 cards are randomly distributed, each player getting 13 of them. Let A_i denote the event that player i gets an Ace, and use the theorem to find $P(A_1 \cap A_2 \cap A_3 \cap A_4)$. The first player receives 13 cards, chosen at random from 48 non-Aces and 4 Aces, so $P(A_1) = \binom{48}{12} \binom{4}{1} / \binom{52}{13}$.

Now work with the reduced pack of 36 non-Aces and 3 Aces, which is what remains, conditional on A_1 having occurred. Then $P(A_2|A_1) = \binom{36}{12} \binom{3}{1} / \binom{39}{13}$. Continuing in this fashion, we find $P(A_3|A_1 \cap A_2)$ by working from a pack with 24 non-Aces and 2 Aces. You should write down the expression. Finally, there is no work at all to find $P(A_4|A_1 \cap A_2 \cap A_3)$, as the answer is plainly unity. (If you do not see this, you have been reading too quickly.)

Theorem 2.3 tells us to multiply together all the probabilities we have just found; there is massive cancellation among the factorial terms, and the final expression shows that the probability that all four players have exactly one Ace is $6 \times 13^3 / (51 \times 50 \times 49) \approx 0.1055$.

The next result shows another way in which using conditional probability can simplify a calculation. The essence of the bargain is to replace one fairly complex calculation by a number of easier ones. The key is finding a *partition* of the set of outcomes that is relevant to the calculation in question.

Definition 2.4

$\{B_1, B_2, \dots\}$ are said to form a *partition* of Ω if

- (i) $B_i \cap B_j = \emptyset$ if $i \neq j$
- (ii) $\bigcup B_i = \Omega$.

Theorem 2.5 (Law of Total Probability)

If A is any event, and $\{B_1, B_2, \dots\}$ form a partition of Ω , then

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

Proof

Using the definition of a partition, we have

$$P(A) = P(A \cap \Omega) = P(A \cap (\bigcup B_i)) = P(\bigcup (A \cap B_i)) = \sum P(A \cap B_i)$$

because the events $\{B_i\}$ are pairwise disjoint. The final expression arises from the definition of conditional probability. $\square$

There is the potential problem that some of the terms $P(A|B_i)$ are not yet defined when $P(B_i)$ is zero. But this is easily overcome. Whatever the expression $P(A|B_i)$ might mean, it is a probability, and so lies between zero and one; and since it gets multiplied by $P(B_i)$, the product will be zero. So we can use this result without worrying about what that conditional probability might mean.

To see this theorem in action, recall Example 1.10, and seek to decide the best strategy for Alex when his initial total is n . We know that if he sticks, his winning chance is $(n/K)^2$.

Suppose he spins; let A = Alex wins, and let B_i = Alex scores i with his second spin. Plainly these events $\{B_i\}$ form a partition, and each has probability $1/K$. For $1 \leq i \leq K - n$,

$$P(A|B_i) = P(\text{Alex wins from } n+i) = (n+i)^2/K^2.$$

For $i > K - n$, then $P(A | B_i) = 0$, as Alex's score gets reduced to zero. The theorem shows that

$$P(A) = \sum_{i=1}^{K-n} \frac{(n+i)^2}{K^2} \times \frac{1}{K} = \frac{1}{K^3} \sum_{r=n+1}^K r^2.$$

So the best tactics, assuming he wants the fun of spinning when the chances are the same, are to spin whenever $(n+1)^2 + (n+2)^2 + \cdots + K^2 \geq Kn^2$.

Given any K , there will be some unique value n_0 such that this inequality holds if, and only if, $n \leq n_0$. Theorem 2.5 also helps us evaluate each player's winning chance, if they play optimally. Again A = Alex wins, and let C_i = First spin scores i . By the above work,

1. if $i > n_0$, then Alex will stick and $P(A|C_i) = i^2/K^2$;
2. if $i \leq n_0$, Alex will spin and $P(A|C_i) = (\sum_{r=i+1}^K r^2)/K^3$.

Since $P(C_i) = 1/K$, Theorem 2.5 shows that the chance Alex wins is

$$P(A) = \frac{1}{K^3} \sum_{i=n_0+1}^K i^2 + \frac{1}{K^4} \sum_{i=1}^{n_0} \sum_{r=i+1}^K r^2.$$

Table 2.1 gives some illustrative values. Recall that Alex wins if the scores are tied, but that Bella has the advantage of going second. The table makes sense – Alex's advantage counts for less as K , the number of different scores, increases, and Bella is soon favourite.

Table 2.1 For various values of K , the optimum tactics for Alex and his winning chances

Value of K	10	20	35	50	75	100	200
Spin if at most	5	10	18	26	40	53	106
Prob. Alex wins	0.512	0.482	0.470	0.465	0.461	0.460	0.457

This law of total probability was the basis, in 1654, for Pascal's method of solving the "problem of points". This asks how a prize should be divided between two players if the series of games in which they compete has to be abandoned before the outcome is settled. When the game began, the winner would be the first to win N games, but the series stops when Clive still needs m wins, and Dave needs n .

There is no unique answer, but the approach developed by Pascal and Fermat is generally accepted as satisfactory. They suggested treating each player as having a 50 % chance of winning any of the unplayed games. Let A denote the event that Clive would win the contest, and let B mean that Clive wins the next game. Then

$$P(A) = P(A|B)P(B) + P(A|B^c)P(B^c).$$

If $p(m, n)$ is the probability that Clive would win the contest when the respective targets are m and n , this means that

$$p(m, n) = \frac{1}{2}p(m-1, n) + \frac{1}{2}p(m, n-1).$$

Since $p(m, 0) = 0$ when $m \geq 1$, and $p(0, n) = 1$ when $n \geq 1$, we can apply the above equation as often as necessary to build up to, and calculate, $p(m, n)$. The prize is then apportioned proportional to each player's chance of victory. Table 2.2 shows some sample values.

Table 2.2 Clive's winning chances $p(m, n)$ when he requires m further victories and Dave requires n

m/n	1	2	3	4
1	1/2	3/4	7/8	15/16
2	1/4	1/2	11/16	13/16
3	1/8	5/16	1/2	21/32
4	1/16	3/16	11/32	1/2

EXERCISES

2.1 An urn initially has six red balls and eight blue ones. Balls are chosen one at a time, at random, and not replaced. Find the probabilities that

- (a) the first is blue, the second is red;
- (b) the first is blue, the sixth is red;
- (c) the first three are all the same colour.

2.2 Show that

$$P(A|B) = P(A|B \cap C)P(C|B) + P(A|B \cap C^c)P(C^c|B).$$

For each of the following statements, either prove it valid, or give a counterexample, using events from the experiment of one throw of a fair die.

- (a) $P(A|B) + P(A^c|B^c) = 1$

$$(b) P(A|B) + P(A|B^c) = 1$$

$$(c) P(A|B) + P(A^c|B) = 1.$$

- 2.3 Three cards of the same size and shape are placed in a hat. One of them is red on both sides, one is black on both sides, the third is red on one side and black on the other. One card is selected at random, and one side is exposed: that side is red. What is the chance the other side is red?

Explain what is wrong with the following argument: "Seeing the red side eliminates the black-black card; since all the cards were initially equally likely, the remaining two cards are equally likely, so there is a 50 % chance the other side is red."

- 2.4 Show that, if $P(A|C) > P(B|C)$ and $P(A|C^c) > P(B|C^c)$, then $P(A) > P(B)$.
- 2.5 (Snell's problem) You have w white balls and b black balls to distribute among two urns in whatever way you like. Your friend will then select one urn at random, and withdraw one ball at random from that urn. What should you do to maximise the chance of getting a black ball?
- 2.6 In a TV quiz show, the star prize is in one of three boxes, A, B or C. When the contestant selects one of these boxes, the host (who knows which box contains the star prize) opens one of the other two boxes, showing it to be empty. He invites the contestant to stick with her original choice, or to swap to the third, unopened, box. Should she change? Or does it make no difference to her winning chances whether she sticks or swaps?

2.2 Bayes' Theorem

With the same set-up as Theorem 2.5, we can look at matters the other way round: *given* that event A has occurred, what can be said about the chances of the different events $\{B_i\}$? This is exactly the situation facing a doctor when she has examined a patient: event A is the set of symptoms and the general background information, while the events $\{B_i\}$ are the various diagnoses possible – measles, beri-beri, hypochondria or whatever. The value of i such that $P(B_i|A)$ is largest will point to the diagnosis that is most likely.

It is also the situation faced by a jury in a criminal trial. Here A represents all the evidence, while $B_1 = \text{Guilty}$ and $B_2 = \text{Not Guilty}$ are the possible ver-

dicts, forming a *partition*, since the jury must select one of them. Formally, the jury is asked to evaluate $P(B_1|A)$, and to return a verdict of Not Guilty unless this conditional probability is very close to 1 – “beyond reasonable doubt”. Each juror will have a personal threshold value (80 %? 99 %? Higher? Lower?) that they use to interpret this time-honoured phrase.

Theorem 2.6 Bayes’ Theorem

If A is any event with $P(A) > 0$ and $\{B_i\}$ form a partition, then

$$P(B_j|A) = \frac{P(A|B_j)P(B_j)}{P(A)} = \frac{P(A|B_j)P(B_j)}{\sum_i P(A|B_i)P(B_i)} \quad (2.1)$$

Proof

Because $P(A \cap B_j)$ can be written both as $P(B_j|A)P(A)$ and as $P(A|B_j)P(B_j)$, the first equality is immediate. The second comes from using Theorem 2.5 to give an alternative form of $P(A)$. $\square$

Take the special case when there are just two alternatives, B_1 and B_2 in this partition, so that $B_2 = B_1^c$. The ratio $P(B_1)/P(B_2)$ is then termed the *odds ratio* of B_1 , as it compares the chance B_1 happens to the chance it does not happen. Using Bayes’ theorem for $j = 1$ and $j = 2$, and dividing the two expressions, we have

$$\frac{P(B_1|A)}{P(B_2|A)} = \frac{P(A|B_1)}{P(A|B_2)} \times \frac{P(B_1)}{P(B_2)} \quad (2.2)$$

In a Court, if B_1 represents a Guilty verdict, this formula tells us how each piece of evidence A should influence our opinions. The left side is the odds ratio, taking account of this evidence, and it is found by multiplying the odds ratio $P(B_1)/P(B_2)$ before the evidence (the *prior odds*) by the so-called *likelihood ratio*, i.e. the ratio of the chances of the evidence arising when the accused is Guilty, or Not Guilty. As each new piece of evidence is introduced, the odds ratio is updated by multiplying by this likelihood ratio. In particular, it follows that the *order* in which the evidence is introduced makes no difference to the final answer. A major problem – and it is of crucial sensitivity – is to give a convincing initial value to $P(B_1)$ before any evidence is given.

Example 2.7 (Attributed to J.M. Keynes)

There are 2598960 different poker hands of five cards for an ordinary deck, of which just four are Royal Flushes, the highest hand possible. Thus the chance

of a Royal Flush is about 1 in 650 000. You play one hand of poker against the Archbishop of Canterbury, who promptly deals himself a Royal Flush! What is the probability he is cheating?

Let B_1 denote the event that he is cheating, and let B_2 mean he is acting honestly. Let A be the evidence, this Royal Flush on the first deal. We know from the above counting that $P(A|B_2) = 1/650\,000$, and we may as well take $P(A|B_1) = 1$, on the grounds that if he is cheating, he can certainly deal himself a Royal Flush. In Equation (2.2), the first term on the right side is 650 000, and we seek the value of the left side – the odds ratio of cheating, given the evidence.

As the equation makes clear, the answer depends entirely on the initial odds ratio. Taking the view that the Archbishop is a man of enormous integrity, our initial feeling may be that $P(B_1)$ is tiny, say 0.00000001. In that case $P(B_2) = 0.99999999$, and the initial odds ratio is effectively the tiny value of $P(B_1)$. When we apply the equation, the odds ratio becomes about 0.0065, so we retain a very strong belief in the integrity of our opponent.

But suppose he deals himself another Royal Flush next hand? After we have multiplied this last odds ratio by the factor 650 000 again, we now believe the odds are about 4000 to one that cheating is taking place.

If our opponent had been someone else, a suspected card sharp, we might have begun with the opinion $P(B_1) \approx 0.1$, and then even after the first deal, the odds ratio has shifted to over 70 000 to one in favour of cheating.

Robert Matthews has pointed out that, in certain circumstances, confession evidence may *weaken* the case against a suspect.

- G = The accused is guilty of a terrorist offence
- I = The accused is innocent of that offence
- C = Confession, extracted after long interrogation.

Equation (2.2) shows that the change in the odds of Guilt comes from the ratio $P(C|G)/P(C|I)$. Provided that $P(C|G) > P(C|I)$, then the confession does shift the odds more towards Guilt. But actual terrorists may have been trained to resist prolonged interrogation, while an innocent person may make a confession simply to end his ordeal. In these circumstances, the above inequality may be reversed, and confession would *reduce* the odds of Guilt. (A real terrorist may reach the same conclusion, and try the double bluff of confessing.)

The general point is that drawing attention to the actions of an accused only strengthens the case against him if those actions are *more likely* when he is guilty than when he is innocent. If a bus journey took the accused past both the crime scene and the shopping centre he said he was visiting, then evidence that tended to prove he took that journey would not add to the case against

him, as the two values of $P(\text{Journey}|\text{Innocent})$ and $P(\text{Journey}|\text{Guilty})$ would have the same value.

Example 2.8 (Daniel Kahneman and Amos Tversky)

In a given city, 85 % of the taxicabs are Green, 15 % are Blue. A witness to a hit-and-run accident identified the perpetrator as a Blue taxicab. Tests under similar lighting conditions showed the witness would correctly identify the colour 80 % of the time, and be wrong 20 %. What is the chance that the cab was indeed Blue?

Let B = Cab was Blue, G = Cab was Green, and W = Witness says the cab was Blue. Then

$$P(B) = 0.15, \quad P(G) = 0.85, \quad P(W|B) = 0.80 \quad \text{and} \quad P(W|G) = 0.20.$$

We require $P(B|W)$, so Equation (2.2) can be transformed into

$$\frac{P(B|W)}{P(G|W)} = \frac{P(W|B)}{P(W|G)} \times \frac{P(B)}{P(G)}.$$

Inserting the numerical values, we find that the odds ratio for a Blue cab is $(0.80/0.20) \times (0.15/0.85) = 12/17$, i.e. despite the evidence of a witness who is 80 % reliable claiming to have seen a Blue cab, it is more likely to have been Green!

For a different medical application of Bayes' theorem, suppose one person in 1000 suffers an adverse reaction to a drug, and a simple test for this reaction is on offer. The test is said to be 95 % reliable, meaning that if the person would suffer a reaction, a positive result comes up 95 % of the time, and if they would not have a reaction, a negative result occurs 95 % of the time. What can we conclude from the knowledge that Susie tests positive?

It is far too tempting to conclude that, as the test is 95 % reliable, there is a 95 % chance she would suffer a reaction. This answer is quite wrong. Let S mean that Susie tests positive, while R means she would suffer an adverse reaction. We seek $P(R|S)$, using R and R^c as the partition for Bayes' theorem. The background information can be expressed as

$$P(S|R) = 0.95 \quad \text{and} \quad P(S|R^c) = 0.05,$$

while we also know $P(R) = 1/1000$. Hence

$$P(S) = P(S|R)P(R) + P(S|R^c)P(R^c) = 0.95 \times \frac{1}{1000} + 0.05 \times \frac{999}{1000} = 0.0509.$$

Thus $P(R|S) = P(S|R)P(R)/P(S) = 0.00095/0.0509 \approx 0.0187$, using Bayes' Theorem. When Susie tests positive, the chance she would suffer the reaction is under 2 % – the test is virtually useless, even though it can claim to be 95 % reliable!

Here, and in Example 2.8, many people's intuitive answer will be wrong, because they have not properly taken into account the background information – the frequencies of Blue cabs, or the prevalence of the disease.

EXERCISES

- 2.7 Soccer player Smith has a good game two times in three, otherwise a poor game. His chance of scoring is $3/4$ in a good game, $1/4$ in a poor one. What is the chance he scores in a game? Given that he has scored, what is the chance he had a good game?
- 2.8 Of all the hats in a shop, one half come from Luton, one third from Milton Keynes, and the rest from Northampton. Two-thirds of hats from Luton are for formal wear, as are one half of those from Milton Keynes and one third of the Northampton hats. A hat is selected at random from the shop; what is the chance it is for formal wear? Given that it is for formal wear, what is the chance it originated in Northampton?
- 2.9 In the last medical example above, make one change in the parameters: it is n in 1000, not 1 in 1000, who would suffer an adverse reaction. Compute the chance that Susie would suffer a reaction, given that she tests positive, as a function of n ($1 \leq n \leq 1000$).
- 2.10 One coin amongst n in a bag is double-headed, the rest are fair. Janet selects one of these coins at random and tosses it k times, recording Heads every time. What is the chance she selected the double-headed coin?
- 2.11 Assume that if a woman carries the gene for haemophilia, any child has a 50 % chance of inheriting that gene, and that it is always clear whether or not a son has inherited the gene, but the status of a daughter is initially uncertain. Karen's maternal grandmother was a carrier, the status of her mother is unknown; but Karen's sister Penny has one son, who is healthy.
 - (a) Find the chance that Karen's first child inherits the haemophilia gene.

(b) Penny now has a second healthy son; repeat the calculation in (a).

(c) But Penny's third son is a haemophiliac; again, repeat (a).

2.12 The till in the pub contains 30 £20 notes and 20 £10 notes. There is a dispute about what denomination Derek used to pay his bill, and the initial assumption is that all 50 notes were equally likely. The barmaid, Gina, claims he used a £10 note, Derek disagrees. Both are honest, but may make mistakes. Show that, using the information that Gina correctly identifies notes 95 % of the time, the chance it was a £10 note is $38/41$.

Derek, who correctly identifies £20 notes 80 % of the time, and correctly identifies £10 notes 90 % of the time, says he used a £20 note. Update your calculation.

2.3 Independence

Suppose $P(A|B) = P(A)$, i.e. knowing that B occurs makes no difference to the chance that A occurs. Since

$$P(A) = P(A|B)P(B) + P(A|B^c)P(B^c),$$

then

$$P(A|B^c) = \frac{P(A) - P(A)P(B)}{1 - P(B)} = P(A);$$

the information that B did *not* occur has also made no difference to the chance that A occurs. Further, we also have

$$P(B|A) = P(B \cap A)/P(A) = P(A|B)P(B)/P(A) = P(B),$$

and similarly $P(B|A^c) = P(B)$. Whenever $P(A|B) = P(A)$, knowing whether or not either A or B occurs makes no difference to the chances of the other one. This is the intuitive notion that A and B are *independent* events.

Definition 2.9

Events A and B are said to be *independent* when $P(A \cap B) = P(A)P(B)$.

Making the definition in this format emphasises the symmetry, and removes any concern about whether either event has zero probability. You should check

that, by this definition, any event whose probability is zero is automatically independent of any other event. Hence the same applies to any event with probability unity.

Consider the experiment of dealing two cards, at random, from a full deck. Write

$$\begin{aligned} A &= \text{First card is an Ace}, & B &= \text{First card is a Spade}, \\ C &= \text{Second card is an Ace}, & D &= \text{Second card is a Spade}. \end{aligned}$$

We can check the working of the definition by verifying that $\{A, B\}$ are independent, as are $\{C, D\}$, $\{A, D\}$ and $\{B, C\}$, but neither $\{A, C\}$ nor $\{B, D\}$ are independent.

Suppose we have more than two events. The *idea* of independence is that knowing whether any of them occurs should make no difference to the chances of other events. For example, suppose $\Omega = \{1, 2, \dots, 8\}$, with all outcomes having probability $1/8$. Write $A = \{1, 2, 3, 4\}$, $B = \{1, 2, 5, 6\}$, $C = \{3, 4, 5, 6\}$. Each pair of these events are independent, so the information that A occurs, or the information that B occurs, makes no difference to the chance that C occurs. However, $P(A \cap B \cap C) = 0$, so if we know that *both* A and B occur, then it is impossible for C to occur.

On the same space, if also $D = \{1, 3, 4, 5\}$ and $E = \{1, 6, 7, 8\}$, we have $P(A \cap D \cap E) = 1/8 = P(A)P(D)P(E)$. A careless beginner might think that it will follow that A , D and E are independent. But simple calculations show that *none* of the pairs $\{A, D\}$, $\{A, E\}$ or $\{D, E\}$ satisfy the definition of pairwise independence. When looking at as few as three events, it is not enough to look at them just in pairs, or just as a triple, we have to look at all the pairs and the triple together. The general definition, for three or more events, is

Definition 2.10

Events $\{A_1, A_2, \dots, A_n\}$ are *independent* if

$$P(A_{1'} \cap A_{2'} \cap \dots \cap A_{r'}) = P(A_{1'})P(A_{2'}) \dots P(A_{r'})$$

whenever $\{1', 2', \dots, r'\} \subset \{1, 2, \dots, n\}$.

An infinite collection of events is independent if every finite subcollection of them is independent.

Independence among a collection of events is a very strong property. To verify all the conditions of this definition with ten events would call for over 1000 calculations! Fortunately, this terrifying prospect seldom arises: usually

either the events have been set up to be independent, or one of the conditions obviously fails.

How should we deal with “independent” repetitions of the same experiment, such as tossing a coin or rolling a die? More generally, how can $\{(\Omega_1, \mathcal{F}_1, P_1), (\Omega_2, \mathcal{F}_2, P_2), \dots\}$ be probability spaces that correspond to independent experiments? Without constructing the whole of $(\Omega, \mathcal{F}, P)$ that corresponds to all these experiments in sequence, the key step is to have

$$P(A_1 \cap A_2 \cap \dots \cap A_n) = P_1(A_1)P_2(A_2) \cdots P_n(A_n),$$

whenever each A_i belongs to $\mathcal{F}_i$, $i = 1, 2, 3, \dots$, and for all values of n . In future work, we will continue to suppress reference to the underlying probability space, unless it is absolutely necessary, and just use this fundamental property without comment. Our first application yields a result that is no surprise, but you should reflect on the steps we have taken to prepare the ground for it.

Corollary 2.11

If $P(A) > 0$ and the experiment is repeated independently, indefinitely often, then A must occur sometime.

Proof

Let B_n be the event that A does not occur in the first n experiments. Then $B_1 \supset B_2 \supset B_3 \supset \dots$, and $B = \bigcap_n B_n$ is the event that A never occurs. But Exercise 1.5 shows that $P(B) = \lim P(B_n)$; and

$$P(B_n) = (1 - P(A))^n \rightarrow 0 \quad \text{as } n \rightarrow \infty$$

since $P(A) > 0$. Hence $P(B) = 0$, so A is certain to occur sometime. □

Example 2.12

Let A and B be mutually exclusive, each with non-zero probabilities. What is the chance that A occurs before B , if the experiment is repeated indefinitely often?

Write C_n to mean that neither A nor B occur in the first n experiments, but one of them does occur in the $(n+1)$ th. These C_n form a partition, since they are plainly disjoint and exhaustive, and let D denote that A occurs before B . Then

$$P(D) = \sum_n P(D \cap C_n) = \sum_{n=0}^{\infty} (1 - P(A) - P(B))^n P(A) = \frac{P(A)}{P(A) + P(B)}$$

on summing the geometric series. To assess which of A or B is likely to appear first, look at the ratio of their probabilities.

The answer makes intuitive sense: perhaps there is a more direct way to obtain it? Try conditioning on the outcome of the first experiment. Since A , B and $(A \cup B)^c$ partition the space, we find that

$$P(D) = P(D|A)P(A) + P(D|B)P(B) + P(D|(A \cup B)^c)P((A \cup B)^c).$$

Plainly $P(D|A) = 1$ and $P(D|B) = 0$. Also, if neither A nor B occur in the first experiment, we can act as though that had never taken place, so that $P(D|(A \cup B)^c) = P(D)$. Hence

$$P(D) = P(A) + 0 + P(D)(1 - P(A \cup B)),$$

from which the previous result follows, as $P(A \cup B) = P(A) + P(B)$.

An application of this result is to the game of craps. Here the shooter throws two dice and wins immediately if the total is 7 or 11. She loses immediately with a total of 2, 3 or 12, and otherwise the game is not yet over: she throws the dice repeatedly until either she repeats her first total – she wins – or she throws a 7 – a loss. We will work out her winning chances.

The table in Example 1.7 shows the chances of the different totals on each throw. Leaving the fractions with their common denominator of 36 simplifies the steps: the chance she wins on her first throw is $6/36 + 2/36 = 8/36$. When the first throw is among the list $\{4, 5, 6, 8, 9, 10\}$ she embarks on a series of repetitions: for example, the chance she throws a 4 and subsequently wins is $3/36 \times 3/(3 + 6) = 1/36$. Making similar calculations, and using the symmetry about the midpoint of 7, her overall winning chance is

$$\frac{8}{36} + \left(\frac{1}{36} + \frac{4}{36} \cdot \frac{4}{10} + \frac{5}{36} \cdot \frac{5}{11} \right) \times 2 = \frac{244}{495}$$

about 0.493. As ever, the game favours the house.

Example 2.13

Suppose a fair coin is tossed repeatedly, and the outcomes noted. Which is more likely to appear first, HH or HT ? What about HH and TH ?

For any given pair of tosses, all four outcomes $\{HH, HT, TH, TT\}$ are equally likely, which suggests that we should use Example 2.12 to conclude that, for both problems, either sequence has a 50 % chance of arising first. But that example may not necessarily apply here, as there the events A and B are within a single experiment, but here we are looking at pairs of experiments that have some *overlap*.

We will show that, for the first problem, it is true that either sequence is equally likely to appear first, but that TH arises before HH 75 % of the time. For HH versus HT , let p be the probability that HH comes up first, and let A be that the first toss is a Head. Then

$$p = P(HH \text{ before } HT|A)P(A) + P(HH \text{ before } HT|A^c)P(A^c).$$

Now

$$P(HH \text{ before } HT|A) = P(H \text{ at second toss}) = 1/2,$$

and also

$$P(HH \text{ before } HT|A^c) = P(HH \text{ before } HT)$$

because, if the first toss is a Tail, it can be ignored. Thus

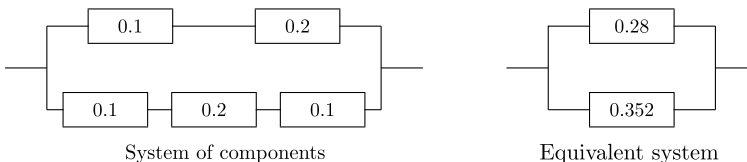
$$p = (1/2) \times (1/2) + p \times (1/2),$$

so $p = 1/2$ as we claimed.

But for HH versus TH , unless the first two tosses are both Heads, it is clear (yes?) that TH must arise first. So the chance HH is before TH is the chance of beginning HH , which is $1/4$, and TH does have a 75 % chance of arising before HH .

Of course for sequences of three successive tosses, there are eight outcomes to consider. It turns out that, whichever one of them is chosen, there is one among the remaining seven whose chance of arising before it is at least $2/3$! You might imagine playing a game against an opponent, allowing her to select whichever of the eight sequences she wants; you will (carefully) pick from the seven that are left – the winner is the one whose sequence turns up first. This contest goes under the name of Penney-ante.

Suppose a complex piece of machinery is made up from linked components, each of which may fail independently of the rest. To give extra protection, some components may be duplicated and placed in parallel, so that only one of them need work; but for components in series, all must work or the system will fail. Consider the set-up shown in the figure on the left, where the numbers in the boxes are the probabilities that component will fail.



The top row will *work* only if both components work, so the chance is $0.9 \times 0.8 = 0.72$, hence the probability of failure is 0.28. Similarly, the chance of failure in the bottom row is 0.352, so the original figure is equivalent to the new system on the right. That system has two components in parallel, so fails only when both fail, which has probability $0.28 \times 0.352 = 0.09856$, or about 10 %. The whole system works with probability 0.90144, call it 90 %.

Example 2.14 (Simpson's Paradox)

Suppose that 200 of 1000 males, and 150 of 1000 females, in University A read Economics. In University B, the figures are 30 of 100 males, and 1000 of 4000 females. In each institution, a higher proportion of males read Economics. But if the universities combine to form a super-university, it has $230/1100 \sim 21\%$ of males in Economics, and $1150/5000 = 23\%$ of females. Proportionally more females than males read Economics!

The possibility of this phenomenon was noted by Simpson (1951). To see how it might arise, write E = Student reads Economics, M = Student is male. Then, in each university separately, we have $P(E|M) > P(E|M^c)$. From the definition of conditional probability, this is equivalent to

$$P(E \cap M)/P(M) > P(E \cap M^c)/P(M^c),$$

which can be rearranged, using $P(E) = P(E \cap M) + P(E \cap M^c)$, to deduce that $P(E \cap M) > P(E)P(M)$. E and M are not independent, they are “positively associated” in each university.

But as the example shows, positive association in each institution on its own need not carry over to a combination. We have two probability spaces, Ω_1 with N_1 points and Ω_2 with N_2 . In each space, all points are equally likely, and $P(E_i \cap M_i) > P(E_i)P(M_i)$ for $i = 1, 2$. Simpson's paradox notes that if $\Omega_3 = \Omega_1 \cup \Omega_2$, and all $N_1 + N_2$ points are equally likely, it does not necessarily follow that $P((E_1 \cup E_2) \cap (M_1 \cup M_2)) > P(E_1 \cup E_2)P(M_1 \cup M_2)$ in Ω_3 .

In the first innings of a cricket match, bowler Eric takes 2 wickets for 36 runs, while Alec takes 4 for 80. In the second innings, Eric takes 1 for 9, Alec takes 6 for 60. In each innings, Eric's runs per wicket were lower, but overall, Eric took 3 for 45 at an average of 15, Alec took 10 for 140 at average of 14. Eric's average was better in each innings, Alec's match average was better. Amalgamating the two innings is appropriate.

60 of 900 people given drug A suffered a reaction (6.7 %), while 32 of 400 given drug B (8 %) suffered a reaction. Drug A appears safer. But there were 1000 healthy people, and 300 who were ill. 800 healthy people were given A, of whom 40 (5 %) reacted, only 2 of 200 healthy people (1 %) reacted to B.

Among 100 ill people given A, 20 reacted (20 %), while 30 of 200 ill people given B reacted (only 15 %). Drug B is safer for both ill people and healthy ones! Amalgamating the data would give an unsound conclusion.

Within an organisation, it is perfectly possible that division A shows better than division B for each of ten subgroups of clients, but B shows better than A when the groups are combined. Management must be very careful when deciding which division deserves a bonus.

EXERCISES

- 2.13 Confirm the remark after Definition 2.9 that if $P(A) = 0$ or 1, then A is independent of any other event. (Such events A are called *trivial* events.)
- 2.14 Suppose events A and B are mutually exclusive. Show they can be independent only if one of them has probability zero. (Thus “mutually exclusive” and “independent” are normally far apart in meaning. But many students, unaccountably, confuse the two.)
- 2.15 Suppose Ω has n equally likely outcomes. If n is prime, show that non-trivial events A and B cannot be independent.
- 2.16 A spinner is a device for selecting one of the integers $\{1, 2, \dots, n\}$ with equal probability. For what values of n are the events

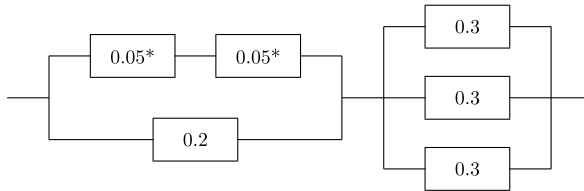
$$A = \text{Even number,} \quad \text{and} \quad B = \text{Multiple of three}$$

independent?

- 2.17 Show that, if
- (a) a fair die is thrown four times independently, it is more likely than not that at least one six appears;
 - (b) a pair of fair dice are thrown 24 times independently, it is more likely than not that a double six does *not* appear.

(This pair of calculations has an honoured place in the history of the development of the formal study of probability. Some seventeenth century gamblers are said to have believed that, since (a) holds, then having six times as many throws ($4 \rightarrow 24$) “ought” to give the same chance of getting an event that was one sixth as likely (six $\rightarrow$ double six). It is very satisfying to see a loose argument give the wrong answer.)

- 2.18 In the diagram, the numbers are the (independent) chances the components will fail within ten years. Find the chance the system fails within ten years.



Given that the system has not failed in ten years, find the chance neither component marked * has failed.

- 2.19 Poker dice have six equally likely faces labelled $\{9, 10, J, Q, K, A\}$. When five such dice are thrown independently, what are the probabilities of the different types of “hand” which, in rank order are: Five of a Kind (aaaaa); Four of a Kind (aaaab); Full House (aaabb); Threes (aaabc); Two Pairs (aabbcc); One Pair (aabcd); and No Pair (abcde)? (“Runs” are not normally considered in this game.)

Given any of these hands, you are allowed to roll again any dice that are unmatched. Find your respective chances of improving your hand for the different initial holdings (ignoring “improvements” such as turning $(KKKQJ)$ to $(KKKA9)$).

2.4 The Borel–Cantelli Lemmas

Suppose $\{A_n\}$ is an arbitrary sequence of events. Write $B_N = \bigcup_{n=N}^{\infty} A_n$, and $C = \bigcap_{N=1}^{\infty} B_N$, so that $C = \bigcap_{N=1}^{\infty} \bigcup_{n=N}^{\infty} A_n$. We claim that event C is the event that infinitely many of the events $\{A_n\}$ occur.

For, suppose infinitely many of the $\{A_n\}$ occur. Then, no matter how large we take N , some A_n with $n \geq N$ occurs, i.e. B_N occurs for every N . Thus C occurs. The argument holds in reverse: if C occurs, so does B_N for all N , and hence no matter how large N is, some A_n with $n \geq N$ occurs. Thus infinitely many A_n occur.

What can we say about $P(C)$, knowing the probabilities of the events $\{A_n\}$?

Theorem 2.15 (Borel–Cantelli Lemmas)

1. If $\sum P(A_n)$ converges, then $P(C) = 0$.

2. If $\sum P(A_n)$ diverges, and the events $\{A_n\}$ are independent, then $P(C) = 1$.

Proof

Suppose first that the series is convergent. Since $P(C) = P(\bigcap_{N=1}^{\infty} B_N)$, we find that $P(C) \leq P(B_N)$ for every choice of N . But

$$P(B_N) = P\left(\bigcup_{n=N}^{\infty} A_n\right) \leq \sum_{n=N}^{\infty} P(A_n),$$

using Boole's inequality (Equation (1.1)). Whatever $\epsilon > 0$ is given, the convergence of the series means that we can find N such that $\sum_{n=N}^{\infty} P(A_n) < \epsilon$. For such values of N , we see that $P(B_N) < \epsilon$. But we established that $P(C) \leq P(B_N)$, which means that $P(C) < \epsilon$. This holds whatever positive value we take for ϵ , no matter how small, so the inevitable conclusion is that $P(C) = 0$.

For the other result, note that, for any N ,

$$P(B_N^c) = P\left(\left(\bigcup_{n=N}^{\infty} A_n\right)^c\right) = P\left(\bigcap_{n=N}^{\infty} A_n^c\right) \leq P\left(\bigcap_{n=N}^{N+M} A_n^c\right)$$

for any $M > 0$. But when the events $\{A_n\}$ are independent, this becomes

$$P(B_N^c) \leq \prod_{n=N}^{N+M} (1 - P(A_n)) \leq \exp\left(-\sum_{n=N}^{N+M} P(A_n)\right),$$

using the simple inequality $1 - x \leq \exp(-x)$, valid when $0 \leq x \leq 1$.

If the series is divergent, whatever the value of N , we can then choose M so that the sum $\sum_{n=N}^{N+M} P(A_n)$ is as large as we like, and hence the value of $\exp(-\sum_{n=N}^{N+M} P(A_n))$ is as small as we like. Hence $P(B_N^c) = 0$, and so $P(B_N) = 1$. This holds for all values of N and so, using Exercise 1.4, $P(\bigcap_{N=1}^{\infty} B_N) = 1$, i.e. $P(C) = 1$. $\square$

The Borel–Cantelli lemmas are an example of a so-called Zero–One Law: we have an event C that is either impossible, or it is certain. The second part of the theorem has the extra condition that the $\{A_n\}$ are independent, and it is plain that some such condition will be required. For otherwise we could take $A_n = D$ for some fixed event D , with $0 < P(D) < 1$; and then $B_N = D$ for all N , so $C = D$ and $P(C)$ can be whatever we choose.

For a potential application of the theorem, let A_n be the event that, in year n , there is a catastrophic accident at some nuclear power station. Plainly,

we shall require that $P(A_n)$ is small, but the Borel–Cantelli lemmas are much more demanding. It seems plausible to take the events $\{A_n\}$ as independent, and then we have:

1. Provided $\sum_{n=1}^{\infty} P(A_n)$ converges, it is impossible that there will be infinitely many catastrophic accidents (small mercies indeed). More alarmingly
2. If $\sum_{n=1}^{\infty} P(A_n)$ diverges (even if $P(A_n) \rightarrow 0$), it is certain there will be infinitely many catastrophes.

So be warned. It is not enough that $P(A_n)$ be tiny, the successive values really ought to be getting smaller year by year, *and* get smaller fast enough for the series to converge. Safety standards in dangerous industries must be perennially made stricter, or we are doomed.



<http://www.springer.com/978-1-4471-5342-9>

Probability Models

Haigh, J.

2013, XII, 287 p. 17 illus., Softcover

ISBN: 978-1-4471-5342-9