Chapter 3

Conditional Probability

3.1 Definition of conditional probability

In spite of our misgivings, let us persist with the frequency definition of probability. Consider an experiment conducted N times under identical conditions. The number of times events A and B occur are denoted by $N(A)$ and $N(B)$, respectively. Consider the ratio:

$$\frac{N(A \cap B)}{N(B)} = \frac{N(A \cap B)}{N} \times \frac{N}{N(B)} = \frac{N(A \cap B)}{N} \div \frac{N(B)}{N}. \tag{3.1}$$

As usual, taking the limit $N \to \infty$, this ratio can be assigned a value of probability (measure). It is called the *conditional probability* and denoted as $P(A|B)$ and defined as:

$$P(A|B) \equiv \lim_{N \to \infty} \frac{N(A \cap B)}{N(B)} \quad , \tag{3.2}$$

that is,

$$\boxed{P(A|B) \equiv \frac{P(A \cap B)}{P(B)}} \quad . \tag{3.3}$$

It is said to be the probability that A occurs *given* that B occurred.

For example, a die is rolled. Given that an even number was observed, what is the probability that the number was 2 ? Clearly we know the outcome was *even* so it was one of 2,4 or 6. Since each of these outcomes is equally likely - then the chance it was 2 is:1/3. An equivalent but more convoluted route to the same result uses the ratio of probabilities. We have events $A = \{2\}$ and $B = \{2, 4, 6\}$. Then, $A \cap B = \{2\}$, so that:

$$P(A) = \frac{1}{6}, \qquad P(B) = \frac{3}{6}, \qquad P(A \cap B) = \frac{1}{6} \quad . \tag{3.4}$$

So, given that B occurred, the probability that A also occurred is:

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{1/6}{3/6} = \frac{1}{3} \quad . \tag{3.5}$$

In conditional probability, the *additional information* provided for by knowing B has occurred must be carefully weighed. Consider the following well-known example indicating where and when confusion can arise.

Example We are told a family has 2 children. Calculate the probability that both are boys, given that we know that at least one is a boy.

First, I'll give you the *wrong* answer."The probability of a boy or girl being born is $\frac{1}{2}$. The gender of any child is an independent random variable. Knowing that one child is a boy has no influence on the

gender of the other child. The probability that the other child is a boy (and thus that both are boys) is $\frac{1}{2}$."

Now the correct answer. Once we are told about the first child, this provides us with extra information before the question is asked about the second child. That is we are asking about a *conditional event*, and the additional information influences our answer. Firstly let's agree with the fact that the gender of any child is an independent random variable. So we can consider the sample space for a 2 child family as follows:

$$\Omega = \{GG, GB, BG, BB\}$$

where the first letter in each pair denotes the gender (B =boy,G =girl) of the first (older) child. We can agree that each of these 4 events is equally likely, with a probability of 0.25.

The information given in the question, that we know one is a boy, immediately eliminates the possibility of two girls (GG) from the possible outcomes. So, in answer to the question:

$$P(BB|\{BG, GB, BB\}) = \frac{P(BB \cap \{BG, GB, BB\})}{P(\{BG, GB, BB\})} = \frac{P(BB)}{P(GG^c)} = \frac{\frac{1}{4}}{1 - \frac{1}{4}} = \frac{1}{3}.$$

The correct answer is $\frac{1}{3}$, NOT $\frac{1}{2}$.

A related question that does have the *intuitive* answer would be the following. What is the probability that, given the *younger* child is a boy, the elder child is also a boy ? This can be calculated in the same manner:

$$P(BB|\{GB, BB\}) = \frac{P(BB \cap \{GB, BB\})}{P(\{GB, BB\})} = \frac{P(BB)}{P(\{GB, BB\})} = \frac{\frac{1}{4}}{\frac{1}{2}} = \frac{1}{2}.$$

3.2 Multiplication Rule

The definition of conditional probability can be rearranged from a quotient to a product, and this is called the *multiplication rule*:

$$\boxed{P(A \cap B) = P(A|B)P(B)} \tag{3.6}$$

Conditional probability obeys the axioms, and associated lemmas, of unconditional probability (see Assignment 1).

Example: In analogy to $P(A^c) = 1 - P(A)$, show that: $P(A^c|B) = 1 - P(A|B)$.

The frequency definition gives us:

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

Alternatively we can use an algebraic approach:

$$P(A^c|B) = \frac{P(A^c \cap B)}{P(B)} \quad ,$$

but since, $B = (A \cap B) \cup (B \cap A^c)$, then

$$P(B) = P(A \cap B) + P(B \cap A^c) \Rightarrow P(A^c \cap B) = P(B) - P(A \cap B) \quad ,$$

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

3.3 Partition Rule

A family of sets (events) $B_1, B_2, B_3, \ldots, B_n$ is a *partition* of Ω if

$$B_i \cap B_j = \emptyset \text{ for all } \; i \neq j \text{ and } \; \bigcup_{i=1}^{n} B_i = \Omega.$$

That is all members are *mutually exclusive* but *exhaustive* of the sample space. That is, the subsets do not overlap, but also they (as a whole) cover the entire sample space.

For *any* events A, B, such that $P(A), P(B) > 0$, the *partition rule* states that:

$$\boxed{P(A) = P(A|B)P(B) + P(A|B^c)P(B^c)} \quad . \tag{3.7}$$

Proof We have a partition $\{B, B^c\}$ (disjoint and exhaustive subsets of the event space). Then the intersections of the set A with the two elements of the partition are disjoint, that is:

$$A = (A \cap B) \cup (A \cap B^c) \qquad , \tag{3.8}$$

in which:

$$(A \cap B) \cap (A \cap B^c) = A \cap (B \cap B^c) = A \cap \emptyset = \emptyset \qquad . \tag{3.9}$$

And since this is the union of disjoint sets (figure 3.3) then

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

This can be illustrated by the Venn diagram (figure 3.3).

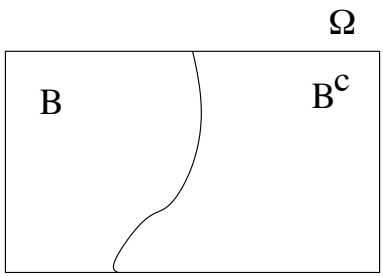

Figure 3.1: A partition of the event space: B and B^c. The subsets B and B^c are, by definition, disjoint subsets of Ω.

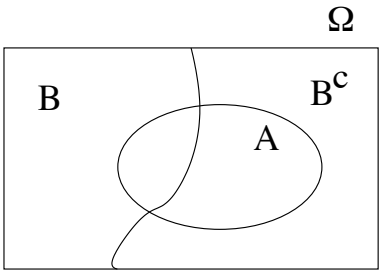

Figure 3.2: This illustrates the division of A across the partition. The set A is thus divided into two disjoint subsets: $A \cap B$ and $A \cap B^c$ according to equation (3.8).

More generally for any *partition* $\{B_1, B_2, \ldots, B_n\}$ then we have the union of disjoint subsets:

$$A = \bigcup_{i=1}^{n} (A \cap B_i) \quad .$$

Then it follows that:

$$P(A) = \sum_{i}^{n} P(A \cap B_i)$$

and using the multiplication rule (3.6), we arrive at the *partition rule*:

$$\boxed{P(A) = \sum_{i=1}^{n} P(A|B_i)P(B_i).} \tag{3.10}$$

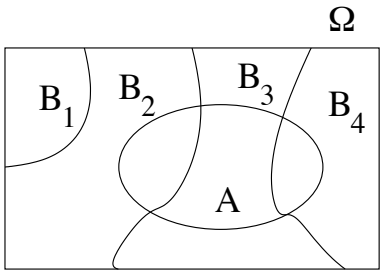

Figure 3.3: The visual representation of the *partition rule* expressed by equation (3.10). This illustration considers the set A divided between the partition sets: $\{B_1, B_2, B_3, B_4\}$.

An illustration of the partition rule is given in figure 3.3. Note that this illustration is entirely equivalent (in mathematical terms) to the algebraic expression (3.10), and moreover is much clearer in conceptual terms as to the meaning of the partition rule. Mathematics is a visual language not just algebraic!

3.4 Bayes' theorem

In general, we have,

$$P(A \cap B) = P(A|B)P(B) \quad , \quad P(A \cap B) = P(B|A)P(A) \quad , \tag{3.11}$$

that is,

$$P(A|B)P(B) = P(B|A)P(A) \quad , \tag{3.12}$$

or equivalently,

$$\boxed{P(A|B) = \frac{P(B|A)P(A)}{P(B)}} \quad . \tag{3.13}$$

which is *Bayes' theorem*. It is fair to say that this is the single most important equation in these lectures. The entire course is based on this relation. In fact, in a different context, it is a slight exaggeration, but only slight, to say that Bayes' theorem is the foundation of all *statistical inference*.

Bayes' theorem is highly-prized by Statisticians and rightly so. It is not without beauty and can, in its purest mathematical form, be developed into the theory of *measures* and *martingales* as well as *information theory*. It is accorded the status equivalent to, in classical mechanics, Newton's laws of motion.

Consider one simple application. Suppose we have a collection of events, a partition: $\{A_1, A_2, \ldots, A_n\}$ and an event B, such that $P(B) \neq 0$. Then

$$P(B) = \sum_{i=1}^{n} P(B|A_i)P(A_i) \tag{3.14}$$

by the partition rule. Then, we have a common expression of Bayes' theorem;

$$P(A_j|B) = \frac{P(A_j \cap B)}{P(B)} = \frac{P(B|A_j)P(A_j)}{\sum_{i=1}^{n} P(B|A_i)P(A_i)} \quad . \tag{3.15}$$

and in particular:

$$\boxed{P(A|B) = \frac{P(B|A)P(A)}{P(B|A)P(A) + P(B|A^c)P(A^c)}} \quad . \tag{3.16}$$

Example Consider two urns containing a mixture of white and blue balls. Urn 1 contains 2W and 7B balls. Urn 2 has 5W and 6B balls. We flip a fair coin and draw a ball from urn 1 if we get H, and draw from urn 2 if T occurs.

What is the probability that the outcome of the toss was H *given* that a W occurred (that is a white ball was selected)?

We can solve this problem by breaking it down into its elements. We seek $P(H|W$, and this can be expressed as:

$$P(H|W) = \frac{P(W|H)P(H)}{P(W)} = \frac{P(W|H)P(H)}{P(W|H)P(H) + P(W|T)P(T)}$$

$$P(H|W) = \frac{(\frac{2}{9}) \times (\frac{1}{2})}{(\frac{2}{9})(\frac{1}{2}) + (\frac{5}{11})(\frac{1}{2})} = \frac{22}{67} \quad .$$

3.5 Independent Events

We call two events, A and B, *independent* if the occurrence of one does not affect the probability that the other occurs.

Thus if $P(A), P(B) > 0$ then A and B are independent if

$$P(A|B) = P(A) \qquad \text{and} \qquad P(B|A) = P(B) \quad .$$

Then since,

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

it follows that:

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

Thus an equivalent definition, and the more commonly used expression for *independent events* is,

$$P(A \cap B) = P(A)P(B) \qquad \Leftrightarrow \qquad A \text{ and } B \text{ independent} \quad . \tag{3.17}$$

3.6 Bayes' theorem in practice

Recall that the product rule for conditional probability can be stated as:

$$P(A|B)P(B) = P(B|A)P(A) \quad ,$$

which is simply an expression of the commutativity of the intersection of the sets A and B.

Of course, this reciprocity between the conditional probabilities can be used to our advantage if we know $P(A|B)$ while, $P(B|A)$, is of interest but not known. That is we seek the (unknown) conditional probability:

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

Let us consider each of the terms on the right-hand-side in turn. $P(B)$, which we have assumed is 'known', is called the *prior* probability. That is, it is known *a priori* (beforehand). Next , $P(A|B)$ is termed the *likelihood* of event A (given that B has occurred), so that the numerator can be expressed as:

$$P(A|B) \times P(B) = \text{likelihood} \times \text{prior} \quad .$$

Finally, $P(A)$ is called the *normalization constant*, and ensures that the conditional probability on the left is a well-defined measure, in the sense that the total conditional probabilities add to 1: $P(B|A) + P(B^c|A) = 1$.

The left-hand-side, $P(B|A)$, determined at the end of the process (afterwards or *a posteriori*) is named the *posterior* term. In summary, the theorem can be expressed as:

$$\text{posterior} = \frac{\text{likelihood} \times \text{prior}}{\text{normalization constant}}$$

This apparently trivial relation, and it is nothing more than the definition of conditional probability, is the basis of all Bayesian modelling and inference.

Suppose we are interested in a proposition or hypothesis being true or false. Given a proposition B, which has probability $P(B)$ of being true, then the probability of its negation $\neg B$ (or B^c) is clearly given by: $P(B^c) = 1 - P(B)$. Now since,

$$P(A|B^c)P(B^c) = P(B^c|A)P(A) \quad ,$$

then we can eliminate the normalization constant, $P(A)$. This gives:

$$\frac{P(B|A)}{P(B^c|A)} = \frac{P(A|B)}{P(A|B^c)} \times \frac{P(B)}{P(B^c)} \quad , \tag{3.18}$$

which can be translated as follows:

Relative chance that B is true given information A = Bayes factor × Relative chance that B is true NOT given information A

Examples

1. Bayes factor

   Problem

   Suppose you run an insurance company and 40% of your customers are under-25 years old. You are aware that recent accident statistics show that a driver under 25 is three times more likely to have an accident than a driver aged 25 and over.

   You receive an insurance claim from one of your customers. What is the probability that the claim is from a driver under 25 ?

   Solution

   The mathematical formulation of the solution is as follows. Let U denote the event of the customer being under 25, and let A be the event of the insurance claim following an accident. Then the quantity we seek is: $P(U|A)$.

   This can be determined from the Bayes formula (3.18):

   $$\frac{P(U|A)}{P(U^c|A)} = \frac{P(A|U)}{P(A|U^c)}\frac{P(U)}{P(U^c)} \quad , \tag{3.19}$$

   First we calculate the *prior* probability ratio, based on the relative chance that a customer, picked at random, is under 25.

   $$\frac{P(U)}{P(U^c)} = \frac{0.4}{0.6}$$

   The additional information that makes up the Bayes factor is as follows. We are informed that the relative probability of a member of the two age-groups having an accident is:

   $$\frac{P(A|U)}{P(A|U^c)} = 3 \quad .$$

   That is, the relative probability increases by a factor of 3 following the information that an accident has occurred.

   Then the *posterior* relative probability is:

   $$\frac{P(U|A)}{P(U^c|A)} = 3 \times \frac{0.4}{0.6} = 2 \quad .$$

Thus, the claimant is twice as likely to be U (under 25) as U^c (25 or over). Then simplifying we get:

$$P(U|A) = 2(1 - P(U|A)) \quad \Rightarrow P(U|A) = \tfrac{2}{3} \quad .$$

So, while a random customer would be under-25 with a 40% probability, a customer who has had an accident would be under 25 with a 67% probability.

2. False positives

Problem

Suppose that a test exists for a performance-enhancing drug, although it is not 100% accurate. If an athlete who is taking the drug is tested, the procedure is 95% accurate in returning a positive result, and hence identifying the person as cheating. For a 'clean' athlete, someone not taking the drug, the procedure is 92% accurate in returning a (correct) negative result for doping. Suppose that, in a certain race, 5% of the athletes are on this drug.

If a random competitor is tested and the result is positive, what is the probability that the person is taking the drug ?

Solution

Let A be the event that the test is positive (indicating doping) and B the event that the tested person is taking the drug. Then we seek the value of, $P(B|A)$.

Let's tackle the calculation head on, using the partition rule to expand the denominator:

$$P(B|A) = \frac{P(A|B)P(B)}{P(A|B)P(B) + P(A|B^c)P(B^c)} \quad . \tag{3.20}$$

But we are told that $P(B) = 0.05$, and therefore $P(B^c) = 0.95$. The value of $P(A|B) = 0.95$ is also specified. Now since $P(A^c|B^c) = 0.92$, we can deduce that $P(A|B^c) = 0.08$. This leads to the result:

$$P(B|A) = \frac{0.95 \times 0.05}{0.95 \times 0.05 + 0.08 \times 0.95} \approx 0.3846 \quad .$$

Since we want a test that effectively identifies drug cheats, this result is not so impressive.

While it is important to have a test that works, and $P(A|B) = 0.95$ is fairly good, it is essential that those accused of doping based on a positive test, are guilty and not being falsely accused. The calculation shows that we can only be 38% certain that a positive test identifies the drug cheat. This is very far from our ultimate aim which is to provide a result that is nearly 100% reliable. That is:

$$P(B|A) \to 1 \qquad . \tag{3.21}$$

In order to achieve this goal we require that the denominator and numerator are roughly the same. This will arise under the conditions

$$P(A|B)P(B) \gg P(A|B^c)P(B^c) \tag{3.22}$$

in which case:

$$P(B|A) = \frac{P(A|B)P(B)}{P(A|B)P(B) + P(A|B^c)P(B^c)} \approx \frac{P(A|B)P(B)}{P(A|B)P(B)} \approx 1 \quad . \tag{3.23}$$

When a very small number are on the drug, as is often the case in practice, it follows that $P(B) \ll 1$. Then the inequality (3.22) can only be satisfied if , $P(A|B^c) \ll 1$.

Consider that an improved test is available. Under the new test, if the procedure for returning a negative result for a 'clean' athlete were 99% accurate (that is $P(A^c|B^c) = 0.99$) , then it follows that, $P(A|B^c) = 0.01$. This greatly improves the success rate to the value:

$$P(B|A) \approx 0.8333 \qquad ,$$

that is, the new test offers an 83% certainty that the positive test correctly identifies a doped athlete. Then the key to a successful detection test is firstly, to make the test very good at detection by eliminating *false negatives* , that is making $P(A^c|B) \ll 1$. Secondly, and equally importantly, in order to avoid false accusations (*false positives*) one needs to ensure that $P(A|B^c)$ is as small as possible.