Dr Isabella Gollini

Probability and Statistics

3 Conditional Probability and Bayes' Theorem

3.1 Conditional probability

Example

Consider again drawing a card at random from a standard pack of 52 cards, so that $|S| = 52$. If we are told that the card drawn is of a black suit, greater than or equal to 10 (ace high), given that information, what is the probability that the card drawn is an ace or a king?

Let B be the event that the card drawn is of a black suit, greater than or equal to 10,

$$B = \{\spadesuit A, \spadesuit K, \spadesuit Q, \spadesuit J, \spadesuit 10, \clubsuit A, \clubsuit K, \clubsuit Q, \clubsuit J, \clubsuit 10\}.$$

$|B| = 10$. Given that B has occurred, we restrict attention to the 10 outcomes in B. Of these 10 outcomes, exactly 4 are an ace or a king. Hence, using the equally likely outcomes approach, the conditional probability that the card drawn is an ace or a king is $4/10 = 2/5$.

To present the argument in a more formal way that will show us how to formulate a general definition of conditional probability, let A be the event that the card drawn is an ace or a king,

$$A = \{\spadesuit A, \spadesuit K, \heartsuit A, \heartsuit K, \diamondsuit A, \diamondsuit K, \clubsuit A, \clubsuit K\}.$$

The outcomes in B that are also in A are given by $A \cap B$,

$$A \cap B = \{\spadesuit A, \spadesuit K, \clubsuit A, \clubsuit K\}.$$

Given that B has occurred, using the equally likely outcomes approach, the conditional probability that A has occurred is given by

$$\frac{|A \cap B|}{|B|} = \frac{4}{10} = \frac{2}{5}.$$

We may also write

$$\frac{|A \cap B|}{|B|} = \frac{|A \cap B|/|S|}{|B|/|S|} = \frac{\Pr(A \cap B)}{\Pr(B)} = \frac{4/52}{10/52} = \frac{2}{5}.$$

This leads to a definition of conditional probability that does not depend upon the use of an equally likely outcomes approach.

Definition

Given some sample space S, let A and B be two events with $\Pr(B) > 0$. The *conditional probability* of A given B, $\Pr(A|B)$, is defined by

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

From the definition of conditional probability we get the *multiplication rule* for calculating $\Pr(A \cap B)$:

$$\Pr(A \cap B) = \Pr(B) \Pr(A|B). \tag{1}$$

3.2 Independence

Two events, A and B, are said to be *independent* or sometimes, to avoid confusion with other types of independence, to be *statistically independent*, if the conditional probability of one of them given the other is the same as its unconditional probability. That is,

$$\Pr(A|B) = \Pr(A).$$

From the definition of conditional probability, the above relationship is equivalent to

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

i.e.,

$$\Pr(A \cap B) = \Pr(A)\Pr(B).$$

Interchanging the roles of A and B, so that we require $\Pr(B|A) = \Pr(B)$, we arrive at the same result, $\Pr(A \cap B) = \Pr(A)\Pr(B)$. This leads to the following formal mathematical definition.

Definition

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

$$\Pr(A \cap B) = \Pr(A)\Pr(B).$$

- We do not require $\Pr(A) > 0$ or $\Pr(B) > 0$ in the definition of independence. If $\Pr(A) = 0$ or $\Pr(B) = 0$ then $\Pr(A \cap B) = 0$ and, trivially, the events A and B are independent.

- The concepts of a pair of events being independent and of them being mutually exclusive are often confused. For a pair of independent events $\Pr(A \cap B) = \Pr(A)\Pr(B)$, but for a pair of mutually exclusive events $\Pr(A \cap B) = 0$. So a pair of mutually exclusive events are <u>not</u> independent, unless $\Pr(A) = 0$ or $\Pr(B) = 0$.

Example

Consider the simple experiment of tossing a fair coin twice. We may take it that there are four equally likely outcomes. In an obvious notation,

$$S = \{hh, ht, th, tt\}.$$

Let A be the event that a head is obtained on the first toss and let B be the event that a head is obtained on the second toss, so that

$$A = \{hh, ht\}$$

and

$$B = \{hh, th\}.$$

The event $A \cap B$ is the event that a head is obtained on both tosses,

$$A \cap B = \{hh\}.$$

Hence

$$\Pr(A \cap B) = \frac{1}{4} = \frac{1}{2} \times \frac{1}{2} = \Pr(A)\Pr(B).$$

The events A and B are independent, as we might suppose (and as is implicit in the assumption of four equally likely outcomes).

Let C be the event that a tail is obtained on the first toss,

$$C = \{th, tt\}.$$

The events A and C are mutually exclusive, i.e., $A \cap C = \emptyset$, but they are <u>not</u> independent, since

$$0 = \Pr(A \cap C) \neq \Pr(A)\Pr(C) = 1/4 \ .$$

Example

Consider again drawing a card at random from a standard pack of 52 cards. Let A be the event that a queen is drawn and let B be the event that a diamond is drawn. $|A| = 4$ and $|B| = 13$ so that $\Pr(A) = 4/52 = 1/13$ and $\Pr(B) = 13/52 = 1/4$. Now $A \cap B = \{\diamondsuit Q\}$, the event that the queen of diamonds is drawn.

$$\Pr(A \cap B) = \frac{1}{52} = \frac{1}{13} \times \frac{1}{4} = \Pr(A)\Pr(B).$$

The events A and B are independent.

3.3 Bayes' Theorem

Lemma 1 *If A and B are events with $\Pr(A) > 0$ and $\Pr(B) > 0$ then*

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

Proof. From the definition of conditional probability or, equivalently, from Equation (1) applied to both $\Pr(A|B)$ and $\Pr(B|A)$,

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

Rearranging this equation we obtain the result of the lemma.

Lemma 2 (The Law of Total Probability) *Given some sample space S, let A be any event and let $B_1, B_2, \ldots, B_i, \ldots$ be a finite or infinite sequence of pairwise mutually exclusive and exhaustive events, none of which has probability zero, that is,*

1.

$$B_i \cap B_j = \emptyset \qquad (i \neq j),$$

2.

$$\bigcup_{i=1}^{\infty} B_i = S$$

3.

$$\Pr(B_i) > 0 \qquad (i \geq 1).$$

Then

$$\Pr(A) = \sum_{i=1}^{\infty} \Pr(B_i) \Pr(A|B_i).$$

Proof. Using a more general version of the distributive law than the one quoted in Section 2, together with the assumption that $\cup B_i = S$,

$$\bigcup_{i=1}^{\infty} (A \cap B_i) = A \cap \left(\bigcup_{i=1}^{\infty} B_i \right) = A \cap S = A.$$

Furthermore, for $i \neq j$,

$$(A \cap B_i) \cap (A \cap B_j) \subseteq B_i \cap B_j = \emptyset,$$

so that the $A \cap B_i$ $(i \geq 1)$ are pairwise mutually exclusive.

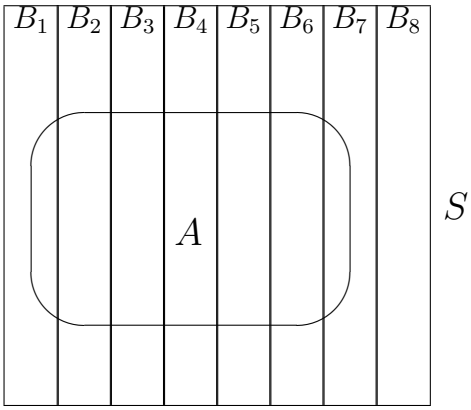

Figure 1: Mutually exclusive and exhaustive events B_i and their intersections with A

Hence, using Axiom 3,

$$\Pr(A) = \sum_{i=1}^{\infty} \Pr\left(A \cap B_i\right).$$

Using the multiplication rule of Equation (1) applied to $\Pr\left(A \cap B_i\right)$, we obtain

$$\Pr(A) = \sum_{i=1}^{\infty} \Pr(B_i) \Pr(A|B_i).$$

Theorem 3 (Bayes' Theorem) *Under the assumptions of Lemma 2 together with the assumption that* $\Pr(A) > 0$*, for all* $j \geq 1$*,*

$$\Pr(B_j|A) = \frac{\Pr(B_j) \Pr(A|B_j)}{\sum_{i=1}^{\infty} \Pr(B_i) \Pr(A|B_i)}.$$

Proof. By Lemma 1,

$$\Pr(B_j|A) = \frac{\Pr(B_j)\Pr(A|B_j)}{\Pr(A)}.$$

Substituting for $\Pr(A)$ from Lemma 2, we obtain Bayes' Theorem.

- Note that under the assumptions of Lemma 2, applying Axiom 3 to the sequence $B_1, B_2, \ldots, B_i, \ldots$,

$$\sum_{i=1}^{\infty} \Pr(B_i) = 1.$$

- Assuming in addition that $\Pr(A) > 0$ and summing the result of Theorem 3, Bayes' Theorem, over j, we obtain

$$\sum_{i=1}^{\infty} \Pr(B_i|A) = 1.$$

3.4 The application of Bayes' Theorem

Bayes' Theorem is commonly applied in situations where we are using the subjective interpretation of probability.

We shall alter the notation of the previous section to reflect this application. Instead of the infinite sequence $B_1, B_2, \ldots, B_i, \ldots$, we shall consider a finite number of mutually exclusive and exhaustive hypotheses $H_0, H_1, \ldots, H_k$, to which we assign a set of probabilities, $\Pr(H_i)$ $(0 \leq i \leq k)$ known as *prior probabilities*, such that

$$\sum_{i=0}^{k} \Pr(H_i) = 1,$$

which are a measure of how likely we think each of the hypotheses is to be true.

For example, there may be just two hypotheses, H_0 and H_1 – in a court of law H_0 that the accused is innocent of the crime of which he is accused and H_1 that the accused is guilty, or, in medical diagnosis, H_0 that the patient is free of some particular disease and H_1 that the patient has the disease.

We then observe an event E or find a piece of evidence E for which we know or can evaluate the probabilities $\Pr(E|H_i)$ $(0 \leq i \leq k)$. We can then use Bayes' Theorem to compute what are known as the *posterior probabilities*, $\Pr(H_i|E)$ $(0 \leq i \leq k)$, which satisfy

$$\sum_{i=0}^{k} \Pr(H_i|E) = 1$$

and are our probabilities for each of the hypotheses in the light of the observed E.

Rewriting Bayes' Theorem in the present notation, we have

$$\Pr(H_j|E) = \frac{\Pr(H_j)\Pr(E|H_j)}{\sum_{i=0}^{k} \Pr(H_i)\Pr(E|H_i)} \qquad (0 \leq j \leq k). \tag{2}$$

- Bayes' Theorem provides provides a mechanism for computing posterior probabilities $\Pr(H_i|E)$ from the prior probabilities $\Pr(H_i)$, a coherent method of updating our beliefs.

- If a further piece of evidence E^* is observed then, assuming that all the relevant conditional probabilities are known, we may use Bayes' Theorem again to update our beliefs, now using as our prior probabilities the posterior probabilities $\Pr(H_i|E)$ that were found on the basis of the earlier evidence E.

3.5 Example: medical screening

We consider the special case $k = 1$, where we have just two hypotheses, H_0 and H_1.

In the setting of medical screening, let H_0 represent the hypothesis that an individual does not have the particular disease under consideration and H_1 the hypothesis that the individual does have the disease. Let E be the occurrence of a positive result from the test procedure that is used in the screening.

In the case $k = 1$, the Law of Total Probability reduces to

$$\Pr(E) = \Pr(H_0)\Pr(E|H_0) + \Pr(H_1)\Pr(E|H_1) \tag{3}$$

and Bayes' Theorem to

$$\Pr(H_0|E) = \frac{\Pr(H_0)\Pr(E|H_0)}{\Pr(H_0)\Pr(E|H_0) + \Pr(H_1)\Pr(E|H_1)}\,, \tag{4}$$

$$\Pr(H_1|E) = \frac{\Pr(H_1)\Pr(E|H_1)}{\Pr(H_0)\Pr(E|H_0) + \Pr(H_1)\Pr(E|H_1)}\,. \tag{5}$$

Suppose that in the population being screened 2% of individuals have the disease. (This figure for the *prevalence* of the disease might have been obtained from knowledge of similar populations that have been investigated previously.) For an individual drawn at random from the population the probability that he/she has the disease is 0.02. So, for a randomly selected individual arriving to be screened, as our prior probabilities we may take $\Pr(H_0) = 0.98$ and $\Pr(H_1) = 0.02$.

Based on previous experience, it is known that for an individual who is free of the disease the probability that the test gives a positive result (a "false positive") is 0.01. For an individual who has the disease the probability that the test gives a positive result is 0.96. Thus $\Pr(E|H_0) = 0.01$ and $\Pr(E|H_1) = 0.96$. We now have in place all the information that we need to evaluate the posterior probabilities.

Firstly, using Equation (3), we may evaluate the probability that for a randomly selected individual the test yields a positive result:

$$\Pr(E) = (0.98)(0.01) + (0.02)(0.96) = 0.029.$$

So in the long run we expect 2.9% of tests to yield positive results.

Given a positive test result, what is the probability that the individual has the disease? Using Equation (5),

$$\Pr(H_1|E) = \frac{(0.02)(0.96)}{0.029} = 0.662,$$

to three decimal places. For an individual who receives a positive test result, there is approximately a 2/3 probability that he has the disease. Given a positive test result, the probability that the individual does not have the disease is $\Pr(H_0|E) = 0.338$. There is approximately a 1/3 probability that he/she is free of the disease, i.e., that we have a false positive.

- This probability $\Pr(H_0|E)$ of a false positive may seem to be surprisingly large, given that $\Pr(E|H_0)$ is much smaller, but it arises because the prior probability $\Pr(H_0)$ is so large.

- Note the importance of distinguishing between $\Pr(H_0|E)$ and the inverse conditional probability $\Pr(E|H_0)$, both probabilities of a false positive, but under different conditions.

3.6 Example: sudden infant death syndrome (SIDS)

The Sally Clark case and her conviction in 1999 for the murder of her two infant children — a conviction eventually overturned by the Court of Appeal in 2003 — brought publicity not only to the issue of cot deaths (deaths classified as due to SIDS) but also to the use of statistical evidence in courts of law. At the original trial, an expert witness, the paediatrician Sir Roy Meadow, gave the figure of 1 in 73 million as the chance of two infants dying of SIDS in one family. It was generally felt that the use of this figure must have had a substantial impact upon the jury in reaching a verdict of guilty.

When in 2005 Professor Meadow was charged by the General Medical Council with serious professional misconduct, a top statistician, Professor Sir David Cox, was called to give evidence. "Cot death numbers don't add up, says eminent statistician" was the headline in *The Times*.

Again taking $k = 1$ in Bayes' Theorem, let H_0 represent the hypothesis that the accused is innocent and H_1 the hypothesis that the accused is guilty of infanticide. Let E represent a piece of evidence, in this case the crucial observation that two children in the family have died from what appears to be SIDS. The accused will be found guilty if there is evidence of guilt "beyond all reasonable doubt". In terms of our notation, the accused will be found guilty if $\Pr(H_0|E)$ is very small.

There are many issues that can be debated here, including the evaluation of the 1 in 73 million figure. But even if we accept the validity of the figure quoted, there is still a major problem with its use in a criminal trial. It is being asserted that

$$\Pr(E|H_0) = \frac{1}{73000000} = 1.37 \times 10^{-8} \ .$$

But it is easy for this probability to be confused with the probability that the accused is innocent, i.e., to suppose that, given the evidence of the two deaths from what appears to be SIDS, the probability that the accused is innocent is 1 in 73 million. Indeed, unless they are given a specific warning, this is what the jury may suppose.

- This is what is sometimes known as *the prosecutors's fallacy*: to identify $\Pr(E|H_0)$ with $\Pr(H_0|E)$. It is also sometimes referred to as "the fallacy of the transposed conditional" or "confusion of the inverse".

We shall carry out a calculation of $\Pr(H_0|E)$ using Bayes' Theorem. We take it that, at least approximately,

$$\Pr(E|H_1) = 1 \ .$$

(If we take $\Pr(E|H_1) < 1$ then this only strengthens the argument below.) To make further progress, we have to specify our prior probabilities for H_0 and H_1. Given a couple from "an affluent, non-smoking family, with the mother over 26", what is the prior probability that the mother kills her two infant children? This is highly problematical to specify, but, given the statistics on infanticide, a figure of

$$\Pr(H_1) = 10^{-5}$$

does not seem unreasonable. In that case

$$\Pr(H_0) = 1 - 10^{-5}$$

and hence, from Equation (3),

$$\begin{aligned} \Pr(E) &= \Pr(H_0)\Pr(E|H_0) + \Pr(H_1)\Pr(E|H_1) \\ &= (1 - 10^{-5})(1.37 \times 10^{-8}) + (10^{-5})(1) \ \approx \ 10^{-5}. \end{aligned}$$

It follows from Equation (4) that

$$\Pr(H_0|E) = \frac{(1 - 10^{-5})(1.37 \times 10^{-8})}{10^{-5}} \approx \frac{1.37 \times 10^{-8}}{10^{-5}} = 1.37 \times 10^{-3}.$$

The posterior probability of innocence, $\Pr(H_0|E) \equiv 1.37 \times 10^{-3}$, is small, but not nearly as small as $\Pr(E|H_0) \equiv 1.37 \times 10^{-8}$. Is the accused guilty beyond all reasonable doubt?

Reminder: from next week onwards you are advised to bring a copy of the *New Cambridge Statistical Tables* with you to lectures.