

Use **Bayes' Theorem** to find the conditional probability of an event $P(A | B)$, say, when the "reverse" conditional probability $P(B | A)$ is the probability that is known.

Objectives

- Learn how to find the probability of an event by using a partition of the sample space S .
- Learn how to apply Bayes Theorem to find the conditional probability of an event when the "reverse" conditional probability is the probability that is known.

Example:

A desk lamp produced by The Luminar Company was found to be defective (D). There are three factories (A, B, C) where such desk lamps are manufactured. A Quality Control Manager (QCM) is responsible for investigating the source of found defects. This is what the QCM knows about the company's desk lamp production and the possible source of defects:

Factory	% of total production	Probability of defective lamps
A	$0.35 = P(A)$	$0.015 = P(D A)$
B	$0.35 = P(B)$	$0.010 = P(D B)$
C	$0.30 = P(C)$	$0.020 = P(D C)$

The QCM (Quality Control Unit) would like to answer the following question: If a randomly selected lamp is defective, what is the probability that the lamp was manufactured in factory C?

$$\begin{aligned}
 P(C|D) &\stackrel{CP}{=} \frac{P(C \cap D)}{P(D)} \stackrel{MR}{=} \frac{P(D|C) \cdot P(C)}{P(D)} \\
 &= \frac{P(D|C) \cdot P(C)}{P[(D \cap A) \cup (D \cap B) \cup (D \cap C)]} \\
 &\stackrel{ME}{\stackrel{MR}{=}} \frac{P(D|C) \cdot P(C)}{P(D|A) \cdot P(A) + P(D|B) \cdot P(B) + P(D|C) \cdot P(C)} \\
 &= \frac{(0.02)(0.30)}{(0.015)(0.35) + (0.01)(0.35) + (0.02)(0.30)} \\
 &= \frac{0.006}{0.01475} = 0.407
 \end{aligned}$$

Now, if a randomly selected lamp is defective, what is the probability that the lamp was manufactured in factory A? And, if a randomly selected lamp is defective, what is the probability that the lamp was manufactured in factory B?

Solution. In our previous work, we determined that $P(D)$, the probability that a lamp manufactured by The Luminar Company is defective, is 0.01475. In order to find $P(A | D)$ and $P(B | D)$ as we are asked to find here, we need to perform a similar calculation to the one we used in finding $P(C | D)$. Our work here will be simpler, though, since we've already done the hard work of finding $P(D)$. The probability that a lamp was manufactured in factory A given that it is defective is:

$$P(A|D) = \frac{P(A \cap D)}{P(D)} = \frac{P(D|A) \times P(A)}{P(D)} = \frac{(0.015)(0.35)}{0.01475} = 0.356$$

And, the probability that a lamp was manufactured in factory B given that it is defective is:

$$P(B|D) = \frac{P(B \cap D)}{P(D)} = \frac{P(D|B) \times P(B)}{P(D)} = \frac{(0.01)(0.35)}{0.01475} = 0.237$$

Note that in each case we effectively turned what we knew upside down on its head to find out what we really wanted to know! We wanted to find $P(A | D)$, but we knew $P(D | A)$. We wanted to find $P(B | D)$, but we knew $P(D | B)$. We wanted to find $P(C | D)$, but we knew $P(D | C)$. It is for this reason that I like to say that we are interested in finding "reverse conditional probabilities" when we solve such problems.

The probabilities $P(A)$, $P(B)$ and $P(C)$ are often referred to as **prior probabilities**, because they are the probabilities of events A , B , and C that we know *prior* to obtaining any additional information. The conditional probabilities $P(A | D)$, $P(B | D)$, and $P(C | D)$ are often referred to as **posterior probabilities**, because they are the probabilities of the events *after* we have obtained additional information.

As a result of our work, we determined:

- $P(C | D) = 0.407$
- $P(B | D) = 0.237$
- $P(A | D) = 0.356$
-

Calculated posterior probabilities should make intuitive sense, as they do here.

For example, the probability that the randomly selected desk lamp was manufactured in Factory C has increased, that is, $P(C | D) > P(C)$, because Factory C generates the greatest proportion of defective lamps ($P(D | C) = 0.02$).

And, the probability that the randomly selected desk lamp was manufactured in Factory B has decreased, that is, $P(B | D) < P(B)$, because Factory B generates the smallest proportion of defective lamps ($P(D | B) = 0.01$). It is, of course, always a good practice to make sure that your calculated answers make sense.

Let's now go and generalize the kind of calculation we made here in this defective lamp example in doing so, we summarize what is called **Bayes' Theorem**.

Generalization of Bayes theorem

Bayes' Theorem. Let the m events B_1, B_2, \dots, B_m constitute a **partition** of the sample space S . That is, the B_i are mutually exclusive:

$$B_i \cap B_j = \emptyset \text{ for } i \neq j$$

and exhaustive:

$$S = B_1 \cup B_2 \cup \dots \cup B_m$$

Also, suppose the prior probability of the event B_i is positive, that is, $P(B_i) > 0$ for $i = 1, \dots, m$. Now, if A is an event, then A can be written as the union of m mutually exclusive events, namely:

$$A = (A \cap B_1) \cup (A \cap B_2) \cup \dots \cup (A \cap B_m)$$

Therefore:

$$\begin{aligned} P(A) &= P(A \cap B_1) + P(A \cap B_2) + \dots + P(A \cap B_m) \\ &= \sum_{i=1}^m P(A \cap B_i) \\ &= \sum_{i=1}^m P(B_i) \times P(A|B_i) \end{aligned}$$

Now, even though I've presented the formal Bayes' Theorem to you, as I should have, the reality is that I still find "reverse conditional probabilities" using the brute force method I presented in the example on the last page. That is, I effectively re-create Bayes' Theorem every time I solve such a problem.

Example

A common blood test indicates the presence of a disease 95% of the time when the disease is actually present in an individual. Joe's doctor draws some of Joe's blood, and performs the test on his drawn blood. The results indicate that the disease is present in Joe.

Here's the information that Joe's doctor knows about the disease and the diagnostic blood test:

- One-percent (that is, 1 in 100) people have the disease. That is, if D is the event that a randomly selected individual has the disease, then $P(D) = 0.01$.
- If H is the event that a randomly selected individual is disease-free, that is, healthy, then $P(H) = 1 - P(D) = 0.99$.
- The **sensitivity** of the test is 0.95. That is, if a person has the disease, then the probability that the diagnostic blood test comes back positive is 0.95. That is, $P(T+ | D) = 0.95$.
- The **specificity** of the test is 0.95. That is, if a person is free of the disease, then the probability that the diagnostic test comes back negative is 0.95. That is, $P(T- | H) = 0.95$.
- If a person is free of the disease, then the probability that the diagnostic test comes back positive is $1 - P(T- | H) = 0.05$. That is, $P(T+ | H) = 0.05$.

What is the **positive predictive value** of the test? That is, given that the blood test is positive for the disease, what is the probability that Joe actually has the disease?

$$\begin{aligned} P(D|T+) &\stackrel{CP}{=} \frac{P(D \cap T+)}{P(T+)} \\ &\stackrel{MR}{=} \frac{P(T+|D) \cdot P(D)}{P[(T+ \cap D) \cup (T+ \cap H)]} \\ &\stackrel{ME}{=} \frac{P(T+|D) \cdot P(D)}{P(T+|D) \cdot P(D) + P(T+|H) \cdot P(H)} \\ &= \frac{(0.95)(0.01)}{(0.95)(0.01) + (0.05)(0.99)} \\ &= 0.0095 / 0.059 = \boxed{0.16} \end{aligned}$$

The test is seemingly not all that accurate! Even though Joe tested positive for the disease, our calculation indicates that he has only a 16% chance of actually having the disease. Is the test bogus? Should the test be discarded? Not at all! This kind of result is quite typical of screening tests in which the disease is fairly unusual. It is informative after all to know that, to begin with, not many people have the disease. Knowing that Joe has tested positive increases his chances of actually having the disease (from 1% to 16%), but the fact still remains that not many people have the disease. Therefore, it should still be fairly unlikely that Joe has the disease.

One strategy doctors often employ with inexpensive, not-too-invasive screening tests, such as Joe's blood test, is to perform the test again if the first test comes back positive. In that case, the population of interest is not all people, but instead those people who got a positive result on a first test. If a second blood test on Joe comes back positive for the disease, what is the probability that Joe actually has the disease now?

$$\begin{aligned}
 P(D|T^+) &\stackrel{BR}{=} \frac{P(T^+|D) \cdot P(D)}{P(T^+|D) \cdot P(D) + P(T^+|H) \cdot P(H)} \\
 &= \frac{(0.95)(0.16)}{(0.95)(0.16) + (0.05)(0.84)} \\
 &= \frac{0.152}{0.194} \\
 &= 0.78
 \end{aligned}$$

Incidentally, there is an alternative way of finding "reverse conditional probabilities," such as finding $P(D|T^+)$, when you know the "forward conditional probability" $P(T^+|D)$. Let's take a look:

		TEST		
		Diseased +	Healthy -	
TRUTH	Diseased +	950	50	1,000
	Healthy -	4950	94050	99,000
		5900	94100	100,000

$$P(D|T+) = \frac{950}{5900} = 0.16$$

Some Miscellaneous Comments

(1) It is quite common, even for people in the seeming know, to confuse forward and reverse conditional probabilities. A 1978 article in the *New England Journal of Medicine* reports how a problem similar to the one above was presented to 60 doctors at four Harvard Medical School teaching hospitals. Only eleven doctors gave the correct answer, and almost half gave the answer 95%.

(2) A person can be easily misled if he or she doesn't pay close attention to the difference between probabilities and conditional probabilities. As an example, consider that some people buy sport utility vehicles (SUV's) so that they will be safer on the road. In one way, they are actually correct. If they are in a crash, they would be safer in an SUV. (What kind of probability is this? A conditional probability!) Conditioned on an accident, the probability that a driver or passenger will be safe is better when in an SUV. But you might not necessarily care about this conditional probability. You might instead care more about the probability that you are in an accident. The probability that you are in an accident is actually higher when in an SUV! (What kind of probability is this? Just a probability, not conditioned on anything.) The moral of the story is that, when you draw conclusions, you need to make sure that you are using the right kind of probability to support your claim.

(3) The Reverend Thomas Bayes (1702-1761), a Nonconformist clergyman who rejected most of the rituals of the Church of England, did not publish his own theorem. It was only published posthumously after a friend had found it among Bayes' papers after his death. The theorem has since had enormous influence on scientific and statistical thinking.