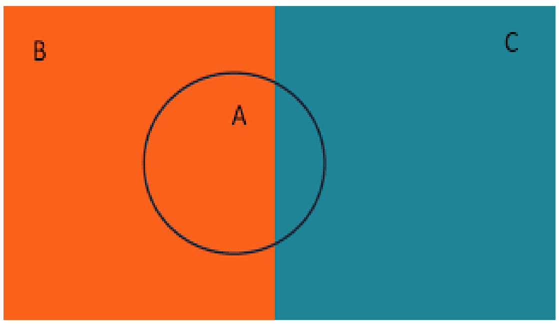
Exercise (conditional probability):

$P(A \mid B) = 0.20$. Determine (i) $P(A^c \mid B)$ (ii) $P(A \mid B^c)$ (iii) $P(A)$ (iv) $P(B)$.

Solution:

(i) $1 - 0.20 = 0.80$ (ii) cannot say (iii) cannot say (iv) cannot say.

Bayes' Theorem



Statement:

Let B and C be two partitions of S. That is, $B \cup C = S$ and $B \cap C = \emptyset$. Let A be any event. Let $P(B)$, $P(C)$, $P(A \mid B)$ and $P(A \mid C)$ be known. Then

(i)

$$P(A) = P(B)\, P(A \mid B) + P(C)\, P(A \mid C)$$

(ii)

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

Proof:

(i)

$A = (B \cap A) \cup (C \cap A)$

$\therefore P(A) = P(B \cap A) + P(C \cap A)$

$\therefore P(A) = P(B)\, P(A \mid B) + P(C)\, P(A \mid C)$

(ii)

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

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

Example (Quality of a blood test):

A blood test gives 97% positive results among the diseased people, and 2% positive results among the healthy people.

In a community, 1% of the people have that particular disease. A person is selected at random from the community and the blood test is performed on that person.

(i) What is the probability that the test result will be positive?

(ii) If the test is positive, what is the probability that the person has the disease?

Solution:

D: the selected person has the disease.

H: the selected person is healthy.

$+$: the test result of the person is positive.

(Set notation rule is ignored for comfort.)

$P(D) = 0.01$, $P(+ \mid D) = 0.97$

$P(H) = 0.99$, $P(+ \mid H) = 0.02$

(i)

$P(+) = 0.01 \times 0.97 + 0.99 \times 0.02 = 0.0295$

(ii)

$$P(D \mid +) = \frac{P(D)\, P(+ \mid D)}{P(+)} = \frac{0.01 \times 0.97}{0.0295} = 0.33$$

Note:

1. If the test result is positive, there is only 33% chance that the person has the disease, and 67% chance that the person is healthy!! Why? [Hint: Healthy people are creating the problem. 😊]

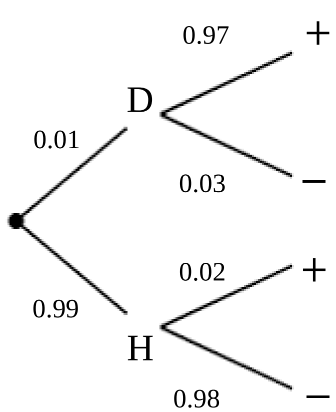
2. $P(D)$ is called the '**prior**' probability, that is, probability before getting 'data' (test result). $P(D \mid +)$ is called the '**posterior**' probability, that is, probability after getting data.

3. What is the contribution of the blood test? Before getting the test result, we could say that there is only 1% chance (prior probability 0.01) that the person has the disease. If the test result is positive, we can say that there is a 33% chance (posterior probability 0.33) that the person has the disease. From 1% to 33% - this is the contribution of the blood test. Unfortunately, this is not enough.

4. If the test result is negative, there is 0.03% (much less than 1%) chance that the person has the disease. (Where did I get this number?)

5. How to develop a useful blood test? Make sure that there are no 'false positives' (healthy people testing positive). If we cannot do that, 'false positive' cases will dominate 'true positive' cases (because, there are so many healthy people), and doctors will have nightmares.

Solution using a tree:
(Quality of a blood test)



From the above tree,

$$P(+) = 0.01 \times 0.97 + 0.99 \times 0.02 = 0.0295$$

$$P(D \mid +) = \frac{P(D)\, P(+ \mid D)}{P(+)} = \frac{0.01 \times 0.97}{0.0295} = 0.33$$

Exercise:

Suppose that 40% of emails are spam. You want to design an "email filtering system" to deal with spam. You have prepared a list of "spam words" (words and phrases that are usually found in spam). If an email is spam, it contains spam words with probability 0.90. On the other hand, a regular email contains spam words with probability 0.05.

(i) What is the probability that a randomly chosen email contains spam words?

(ii) If the email contains spam words, what is the probability that it is spam?

Solution: Do it yourself.