Examples in Conditional Probabilities

1 False Positives and False Negatives (Conditional Probabilities and Bayes' Rule)

Recall that

$$P[A|B] = \frac{P[A \cap B]}{P[B]}$$

so that, since

$$P[A \cap B] = P[A|B]\,P[B] = P[B|A]\,P[A]$$

$$P[A|B] = \frac{P[B|A]\,P[A]}{P[B]}$$

(Bayes' Rule). While the formula may seem a curiosity, it reflects a basic method of inquiry ("induction")e . The following exercise is a typical example.

Suppose we are testing for a rare illness that is known to affect 2% of the population. Suppose that we are using a test known to have a rate of 5% false positives (people who are not ill test as ill), and 1% false negatives (people who are ill do not test for the illness).

1. Suppose an individual tests positive: what is *now* the probability that he/she is actually ill?
2. Suppose an individual tests negative: what is *now* the probability that he/she is actually ill?
3. Suppose an individual who tested positive a first time is tested again with the same procedure. If the result is positive, what are the odds now? And what are they if the test is negative?

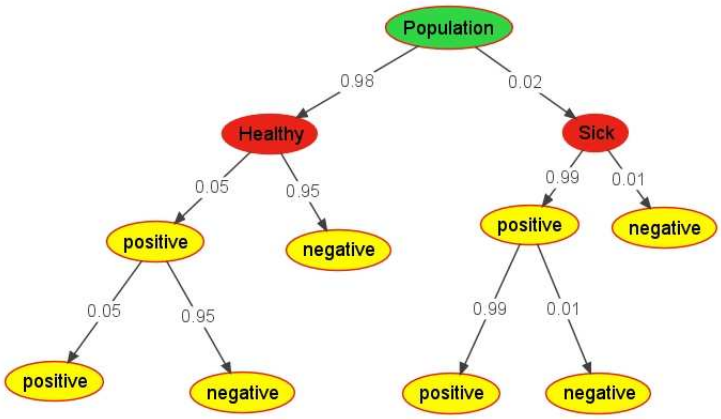
Hints: Give names to the various possibilities. For example we may define

- A =the individual is ill, and A^c =the individual is healthy
- E =the test is positive, and E^c =the test is negative

Note how we know $P[A]$. The "false positive" and "false negative" probabilities are *conditional probabilities* (for example, a "false positive" corresponds to E, given A^c). The requested probabilities in 1 and 2 can be found using Bayes' Rule, since we want the probability of being ill, knowing the result of the test. As for 3, it is the same question as 1 or 2, but we use the result from 1 in place of the original $P[A]$.

Remark: examples of this type show why it is imperative to test more than once for illnesses or other conditions, especially if the probability of being affected is relatively low.

Solutions: The various possibilities (a patient being healthy or not, a test resulting positive - signaling illness - or negative) can be represented as a tree, where each connecting line is labeled with the probability of the target (the "offspring" node), given the source (the "parent" node), that is the corresponding conditional probabilities:



The probability of each node is found by multiplying the numbers on the lines (the "branches") leading from the "root" (labeled "Population") to the node.

1. We are told that

$$P[A] = 0.02, P[E|A^c] = 0.05, P[E^c|A] = 0.01$$

from which we can also find that

$$P[A^c] = 0.98, P[E^c|A^c] = 0.95, P[E|A] = 0.99$$

This allows us to compute

$$P[E] = P[E \cap A] + P[E \cap A^c] = P[E|A]\,P[A] + P[E|A^c]\,P[A^c] =$$

$$= 0.99 \cdot 0.02 + 0.05 \cdot 0.98 = 0.0688$$

Now, we can look for what we want, that is $P[A|E]$ (the probability of a patient being ill, given that the test was positive). To this end, we use Bayes' Formula:

$$P[A|E] = \frac{P[E|A]\,P[A]}{P[E]}$$

and simply plug in the numbers:

$$P[A|E] = \frac{0.99 \cdot 0.02}{0.0688} \approx 0.288$$

This is way higher than 0.02, but is far from certainty – in fact, it is still much more likely, even after a positive test, that our patient is healthy, rather than ill!

2. This is similar: we are asking for

$$P[A|E^c] = \frac{P[E^c|A]P[A]}{P[E^c]} = \frac{0.01 \cdot 0.02}{1 - 0.0688} \approx 2 \cdot 10^{-4}$$

That's really minimal...

3. If we take a second test on a patient that tested positive already, we cannot use $P[A|E] = \frac{P[E|A]P[A]}{P[E]}$ as a formula with $P[A] = 0.02$. We now *know* that a first test was positive, hence, that the probability of illness is no longer 2%, but rather 29%. Hence, we now have a new model. Without changing notation, we will have (rounding the numbers for simplicity)

$$P[A] = 0.29, P[E|A^c] = 0.05, P[E^c|A] = 0.01$$

$$P[A^c] = 0.71, P[E^c|A^c] = 0.95, P[E|A] = 0.99$$

$$P[E] = P[E \cap A] + P[E \cap A^c] = P[E|A]P[A] + P[E|A^c]P[A^c] =$$

$$0.99 \cdot 0.29 + 0.05 \cdot 0.71 = 0.323$$

and

$$P[A|E] = \frac{P[E|A]P[A]}{P[E]} \approx \frac{0.99 \cdot 0.29}{0.323} \approx 0.88996$$

Now we are really scared!

2 Why the Naive "Law of Averages" Does Not Hold

Suppose you are repeatedly playing a "pure game of chance", like a lottery, or most casino games. Assume an "ideal" situation: the outcome of each game is independent of all others, and they are all equal as far as the chance of winning.

Assume your chance of winning one game is $p < 1$ (in a lottery, p is extremely small, in some casino games, like roulette, there are bets where $p < \frac{1}{2}$, but not by much). Your probability of *not* winning one game is then $q = 1 - p$. Define the random variable, $X = n$ if the first time you win is the nth. Since all games are independent, the probability that $X = n$ is the probability of losing the first $n - 1$ games, and winning the nth, and is the product of the probabilities of each of these events:

$$P[X = n] = pq^n$$

It also follows that the probability of never winning in n attempts is

$$P[X > n] = q^n \tag{1}$$

since it equals the probability of losing the first n games.

1. Since $q < 1$, show that the probability of never winning over n games is very small if n is sufficiently large. For this purpose, you can compute an example, e.g., $q = 0.55$ and $n = 10$. Try a few similar numbers.

2. This observation corresponds to the intuitive fact that "sooner or later you have to win". However, consider the following situation: you have lost already n attempts (the following event is given: $\{X > n\}$). You are asking now what is your probability of wining within the next m attempts (that is you are looking at $P[X \leq n + m]$, knowing that $\{X > n\}$. Show that this probability is precisely the same as that of winning within the first m attempts (it is easier to how that the probability of *not* winning in the next m attempts is the same as that of not winning in the first m attempts). In other words, that having lost n times did not earn you any brownie points: you are right where you were at the beginning...

Solutions:

1. From the formula (1), we have that

$$P[X > 10] = 0.55^{10} \approx 0.0025$$

   In general, q^n will become very small very fast (it's called "exponential decay", and it is indeed fast) if $q < 1$.

2. We just write the formula for conditional probabilities (it is faster to work with $\{X > n + m\}$):

$$P[X > n + m \,|X > n] = \frac{P[X > n + m, X > n]}{P[X > n]}$$

   Now, if you lost $n + m$ games, you certainly lost n, so the numerator is actually equal to $P[X > n + m]$. We now refer to (1), to see that

$$P[X > n + m \,|X > n] = \frac{q^{n+m}}{q^n} = q^{n+m-n} = q^m = P[X > m]$$

   thanks to the peculiarities of the exponential function. We conclude that losing n games at the start didn't make it any more likely to win in the next m games, than it was at the beginning. The folk statement of this feature (it's peculiar to the geometric and to the exponential distributions) is that "the geometric distribution has no memory".