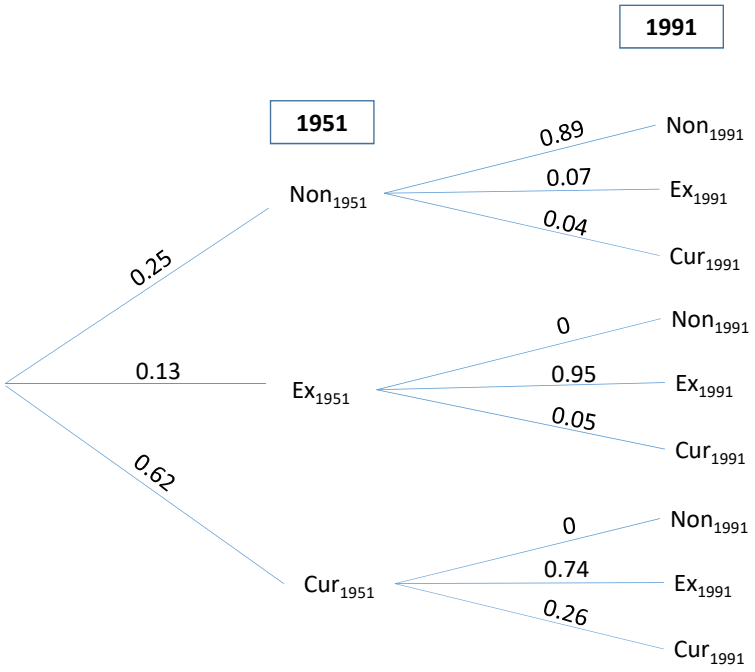
Practical 2 Solutions

Question 1

(a) For convenience, denote non-smoker, ex-smokers and current smokers in 1991 by Non_{1991}, Ex_{1991}, and Cur_{1991}, respectively, with smoking status in 1951 denoted in an analogous manner. The probability tree can be drawn as follows:



(b) Probability of being an ex-smoker in 1991:

$$Pr(\text{Ex}_{1991}) = (0.25 \times 0.07) + (0.13 \times 0.95) + (0.62 \times 0.74) = 0.60$$

It is clear that smoking habits have changed drastically between 1951 and 1991 for these doctors.

(c) We want $P(\text{Non}_{1951} \,|\, \text{Ex}_{1991}), P(\text{Ex}_{1951} \,|\, \text{Ex}_{1991})$ and $P(\text{Cur}_{1951} \,|\, \text{Ex}_{1991})$. Bayes theorem says:

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

We apply this to the current example:

$$\begin{aligned} P(\text{Non}_{1951} \,|\, \text{Ex}_{1991}) &= \frac{P(\text{Ex}_{1991} \,|\, \text{Non}_{1951}) \times P(\text{Non}_{1951})}{P(\text{Ex}_{1991})} \\ &= \frac{0.07 \times 0.25}{0.60} = 0.029 \end{aligned}$$

where the denominator comes from our calculations in part (b). In this way, we find that:

$P(\text{Non}_{1951} \,|\, \text{Ex}_{1991}) = 0.07 \times 0.25/0.60 = 0.03$
$P(\text{Ex}_{1951} \,|\, \text{Ex}_{1991}) = 0.95 \times 0.13/0.60 = 0.21$
$P(\text{Cur}_{1951} \,|\, \text{Ex}_{1991}) = 0.74 \times 0.62/0.60 = 0.76$

Note that 76% of the doctors who were ex-smokers in 1991 were smokers in 1951. Three percent of the doctors started smoking after 1951 and stopped smoking before 1991. The other 21% of the doctors were ex-smokers in 1951 and 1991. This means they stopped smoking before 1951. We do not know whether they had relapses between 1951 and 1991.

Question 2

Cystic fibrosis is an inherited condition in which the lungs and digestive system can become clogged with thick, sticky mucus. The gold standard diagnostic test is the sweat chloride test. The sensitivity of the sweat chloride test is 91.7% and the specificity is 99.9%.

Tests are only performed on patients suspected of having cystic fibrosis (i.e. those with symptoms typical of cystic fibrosis). The probability of cystic fibrosis in this group is 3.25%.

(a) The positive predictive value is calculated as

$$\begin{aligned} P(CF|+ve) &= \frac{P(+ve|CF)P(CF)}{P(+ve|CF)P(CF)+P(+ve|\bar{CF})P(\bar{CF})} \\ &= \frac{0.917\times 0.0325}{0.917\times 0.0325+(1-0.999)\times(1-0.0325)} = 0.968 \end{aligned}$$

(b) The PPV is 96.8%, meaning that a patient whose sweat test comes back positive has a 96.8% chance of having cystic fibrosis.

(c) The PPV in the whole population is

$$\begin{aligned} P(CF|+ve) &= \frac{P(+ve|CF)P(CF)}{P(+ve|CF)P(CF)+P(+ve|\bar{CF})P(\bar{CF})} \\ &= \frac{0.917\times 0.0004}{0.917\times 0.0325+(1-0.999)\times(1-0.0004)} = 0.268 \end{aligned}$$

If we screened everyone in the general population using this test, 26.8% of those who had a positive test would actually have cystic fibrosis. This is probably not, therefore, a very useful test in this setting. (It would also be very expensive and logistically difficult!).

Question 3

(a) Using the definition of expectation we have that

$$E(X) = 0\times 0.3 + 1\times 0.1 + 2\times 0.6 = 1.3$$

To find $Var(X)$ we first find $E(X^2)$:

$$E(X^2) = 0\times 0.3 + 1^2\times 0.1 + 2^2\times 0.6 = 2.5$$

and so $Var(X) = E(X^2) - E(X)^2 = 2.5 - 1.3^2 = 0.81$.

(b) There are many different ways of showing that X and Y are not independent. One is to note that that $P(X=0, Y=1) = 0.05$ is not equal to $P(X=0)P(Y=1) = 0.3\times 0.1 = 0.03$, therefore violating the independent definition.

(c) (i) $P(Y|X=0)$ is found by dividing the joint probabilities by $P(X=0)$, giving conditional probabilities $P(Y=1|X=0) = 0.05/0.3 = 0.166$, $P(Y=2|X=0) = 0.333$, $P(Y=3|X=0) = 0.5$.

(ii)

$$E(Y|X=0) \quad = \quad 1 \times 0.166 + 2 \times 0.333 + 3 \times 0.5 = 2.333$$

(iii) $P(Y|X=0)$ is different to $P(Y|X=1)$, i.e. the conditional distribution of Y differs according to the value of X, which is consistent with X and Y not being independent. Similarly, the conditional expectation of Y given X depends on X, which again reflects the dependency between X and Y.

Additional: Question 4

(a) $E(2+3X) = 2 + 3E(X) = 2 + 3 \times 1 = 5$

(b) $E((2+X)^2) = E(4 + 4X + X^2)$. Since $Var(X) = 5$ and $E(X) = 1$, $5 = E(X^2) - 1^2$, and so $E(X^2) = 6$. Therefore $E((2+X)^2) = E(4 + 4X + X^2) = 4 + 4 \times 1 + 6 = 14$.

(c) $Var(10 + 3X) = 3^2 Var(X) = 9 \times 5 = 45$.

Additional: Question 5

(a) The standard definition of expectation gives

$$\begin{aligned} E(XY) &= \sum_x \sum_y x\, y\, P(X = x, Y = y) \\ &= \sum_x \sum_y x\, y\, P(X = x)\, P(Y = y). \end{aligned}$$

where the second line follows since X and Y are *independent* discrete random variables. Then we can separate out the terms involving x from those involving y,

$$\begin{aligned} E(XY) &= \sum_x x\, P(X = x) \sum_y y\, P(Y = y) \\ &= E(X)E(Y). \end{aligned}$$

(b) The standard definition of variance gives

$$Var(X+Y) \quad = \quad E((X+Y)^2) - E(X+Y)^2$$

We expand the brackets and use some of the properties of expectation that we have met:

$$\begin{aligned} Var(X+Y) &= E(X^2 + Y^2 + 2XY) - (E(X) + E(Y))^2 \\ &= E(X^2) + E(Y^2) + 2E(XY) \\ &\qquad -E(X)^2 - E(Y)^2 - 2E(X)E(Y) \end{aligned}$$

We now use the equality $E(XY) = E(X)E(Y)$, which requires our assumption that X and Y are independent. Then

$$\begin{aligned} Var(X+Y) &= E(X^2) - E(X)^2 + E(Y^2) - E(Y)^2 \\ &= Var(X) + Var(Y). \end{aligned}$$

Optional: Brainteaser

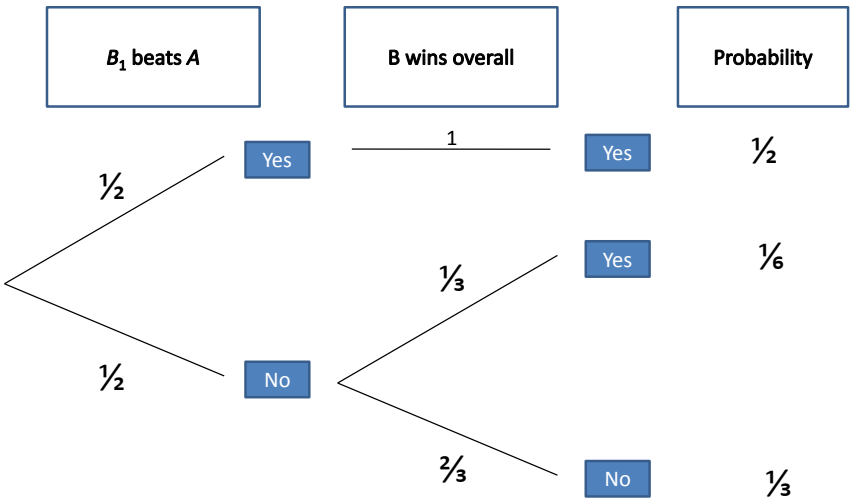
Solution 1 is correct.

The phrase *his first throw is now irrelevant* is the downfall of Solution 2. The throws themselves, A, B_1 and B_2 are indeed independent, but who wins and who loses is based on the *differences*, $B_1 - A$ and $B_2 - A$, and these two are *not* independent: $-A$ appears in both!

B's probability of winning on his second throw is not independent of the fact that he lost on his initial throw. The probabilities in the second column of the probability tree should be calculated using

$$\begin{aligned} P(B_2 > A \,|\, B_1 < A) &= \frac{P(B_2 > A \cap B_1 < A)}{P(B_1 < A)} \\ &= \frac{P(B_1 < A < B_2)}{P(B_1 < A)} \end{aligned}$$

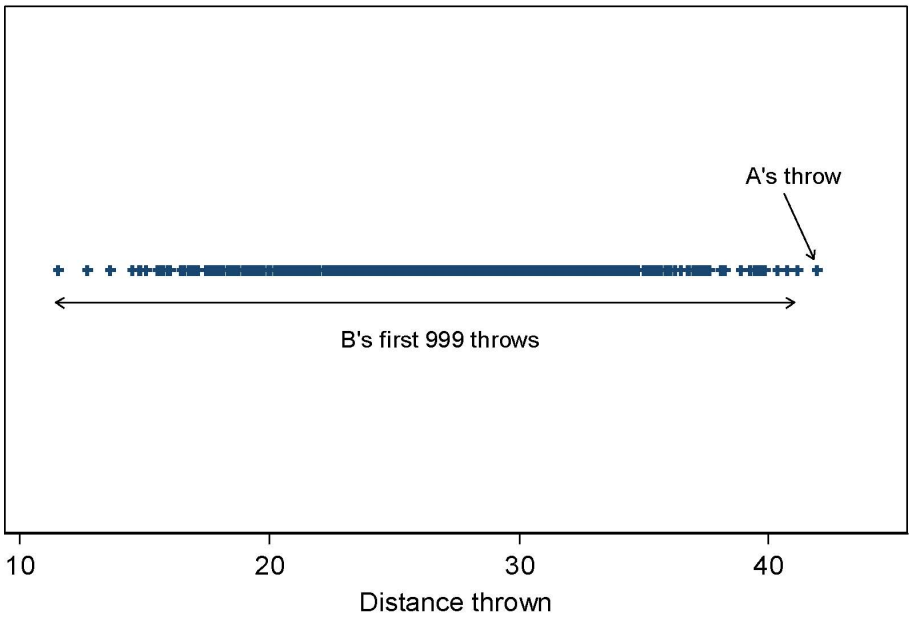
If we look at the permutations given in Solution 1, we see that $P(B_1 < A < B_2) = 1/6$ and $P(B_1 < A) = 1/2$. Thus $P(B_2 > A \,|\, B_1 < A) = 1/3$ and the tree should have looked like this:



which then gives a probability $1/2 + 1/6 = 2/3$ that B wins, in agreement with the first solution.

In other words, knowing that B lost on his first throw increases our belief that he will lose again on the second throw. The fact that A's throw won against B's first throw gives us some reason to believe that A's throw was, by chance, quite good, and that therefore, in order to beat this good throw on his second attempt, B also needs a good throw, which happens with probability $< 1/2$.

If you're still not convinced that his probability of winning on his second attempt decreases once we condition on his first loss, imagining a more extreme example (as we did with the Monty Hall problem last week) might help. Suppose that A gets one throw but that B gets 1000. Now, suppose that after B has thrown 999 of his 1000 attempts, A's one single attempt is the best of the lot (as is shown in the diagram below). This is of course unlikely, but suppose it has happened. Now, we must be pretty convinced that A's single throw was *extremely* lucky, and that it probably went much further than the average expected from both A's and B's throws. *Given this*, it is now extremely unlikely that — on his one remaining attempt — B can win the game. The incorrect tree diagram



given in Solution 2, would be correct for a different game in which A had two throws as well as B but that A could only win if she beat B on both occasions. Then, we would be looking at $B_1 - A_1$ and $B_2 - A_2$, which *are* independent. So the first 'moral' of this story is that, in every column except for the first column in a tree diagram, the probabilities are all **conditional on the previous branches** along which you have travelled. The other moral of the story is that we can sometimes be tricked into thinking that events are independent when in fact they are not, so careful thought is needed when calculating these conditional probabilities. As Dennis Lindley says in the introduction to his book *Understanding Uncertainty*:

> Although we shall meet conclusions that at first surprise, further reflection suggests that they are correct and that our common sense is faulty. Indeed, one of the merits of our approach is that it does produce results that conflict with common sense and yet, on careful consideration, are seen to be sound. In other words, it is possible to improve on common sense.