Think of assigning conditional probabilities as “re-scaling” original probabilities

BIO210 Biostatistics

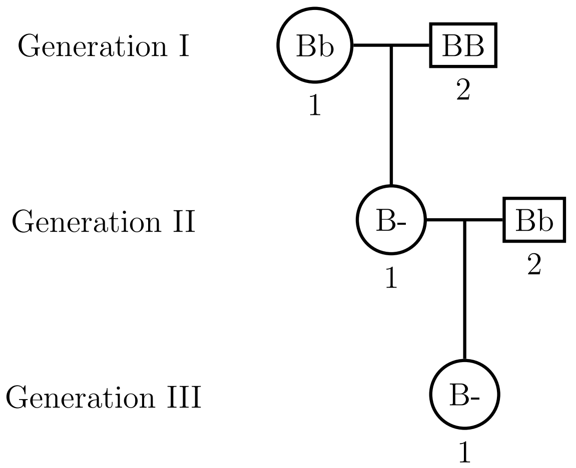
Extra Reading Material for **Lecture 6**

Xi Chen
School of Life Sciences
Southern University of Science and Technology

Fall 2024

The pedigree example

Let's revisit the pedigree example we talked about during the lecture. We do not have any problems with the first example. The second example is drawn below:



$\mathbb{P}\left(\text{III}_1 \text{ is heterozygous}\right) = ?$

Now let's define some events to help us with the visualisation:

$\mathbf{E} = \{ \text{ II}_1 \text{ is } \mathbf{BB} \text{ } \}$, $\mathbf{E}^C = \{ \text{ II}_1 \text{ is } \mathbf{Bb} \text{ } \}$

$\mathbf{F1} = \{ \text{ III}_1 \text{ is } \mathbf{BB} \text{ } \}$, $\mathbf{F2} = \{ \text{ III}_1 \text{ is } \mathbf{Bb} \text{ } \}$, $\mathbf{F3} = \{ \text{ III}_1 \text{ is } \mathbf{bb} \text{ } \}$

$\mathbf{G} = \{ \text{ III}_1 \text{ is } \mathbf{B\text{-}} \text{ } \} = \mathbf{F1} \cup \mathbf{F2}$

We should be very clear and there is no doubt that $\mathbb{P}(E) = \mathbb{P}\left(E^C\right) = 0.5$. However, once we know that event **G** has occurred, we want to use this information to update $\mathbb{P}(E)$ and $\mathbb{P}\left(E^C\right)$. Therefore, it is $\mathbb{P}(E|G)$ and $\mathbb{P}\left(E^C|G\right)$ that we do not know and want to find out.

Like discussed during the lecture, we should build the tree based on all possibilities. The fact that event **G** has occurred is a piece of information that ***we only observe afterwards***.

Therefore, we should build our tree like this:

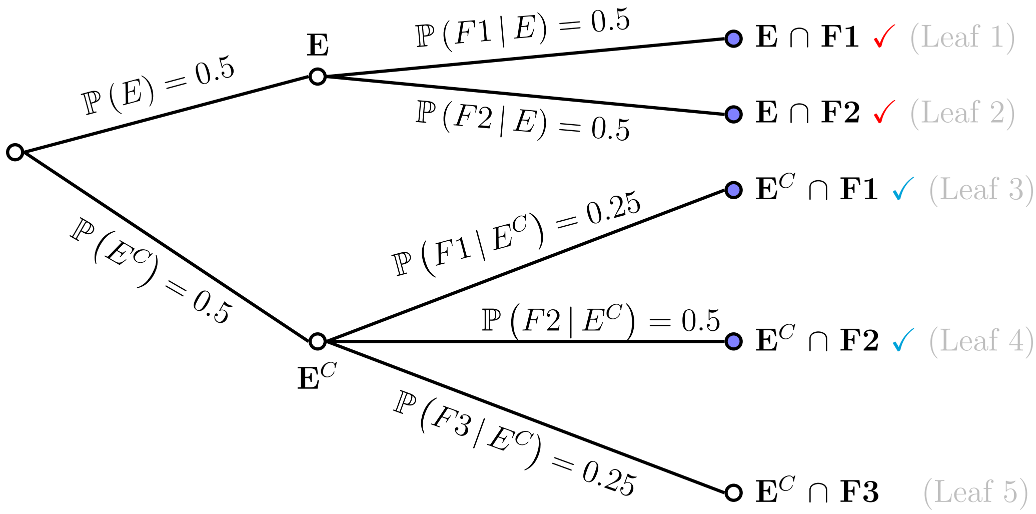


Figure 1. The full model

Now we observe that event **G** has occurred, meaning all of a sudden, we are living in a universe of the blue leaves. Therefore, we should update our beliefs about $\mathbb{P}(E)$ (red checks) and $\mathbb{P}(E^C)$ (cyan checks) accordingly:

$$\mathbb{P}(E \mid G) = \frac{\mathbb{P}(\checkmark)}{\mathbb{P}(\bullet)} = \frac{0.5 \times 0.5 + 0.5 \times 0.5}{0.5 \times 0.5 + 0.5 \times 0.5 + 0.5 \times 0.25 + 0.5 \times 0.5} = \frac{4}{7}$$

$$\mathbb{P}(E^C \mid G) = \frac{\mathbb{P}(\checkmark)}{\mathbb{P}(\bullet)} = \frac{0.5 \times 0.25 + 0.5 \times 0.5}{0.5 \times 0.5 + 0.5 \times 0.5 + 0.5 \times 0.25 + 0.5 \times 0.5} = \frac{3}{7}$$

Even though it is not recommended, some of you are still wondering how to build the tree with the information incorporated into the model. That is, when you build your tree, you want to eliminate the $E^C \cap F3$ leaf (Leaf 5) from the very beginning. In this case, you are building your tree under the universe where event **G** has occurred. It is okay. Since your tree is under the universe where event **G** has occurred, what you should do is to to put all those probabilities with the condition of **G** in them. In this case, we are not sure about the conditional probability of $\mathbb{P}(E \mid G)$ and $\mathbb{P}(E^C \mid G)$. Let's denote them as p and $1 - p$, respectively.

Your tree will look like:

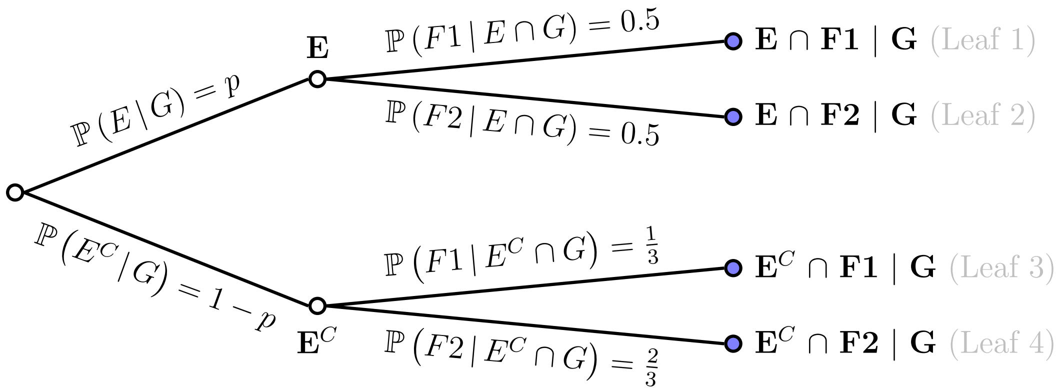
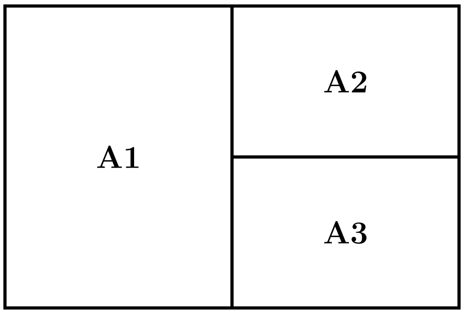


Figure 2. The conditional model (conditioned on **G**)

How do we proceed from here? We need to re-scale the probability of those four leaves. As always, let's look at the simplest case where we have three disjoint events **A1**, **A2** and **A3** that form a sample space with the ratio of the areas as 2 : 1 : 1, like this:



Therefore, the probabilities are: $\mathbb{P}(A1) = \frac{1}{2}$, $\mathbb{P}(A2) = \frac{1}{4}$, $\mathbb{P}(A3) = \frac{1}{4}$. Now we are told that event **D** = { **A3** does NOT ocurr } has occurred. Given that **D** has occurred, what are the probabilities of **A1-3**? Given that **D** has occurred, all of a sudden, we know that we are living in a universe of **A1** ∪ **A2**. Everything outside **A1** ∪ **A2** becomes irrelevant. It is relatively straightforward to see that $\mathbb{P}(A1 \mid D) = \frac{2}{3}$ and $\mathbb{P}(A2 \mid D) = \frac{1}{3}$. There are two things you should notice:

1. We have the normalisation property: $\mathbb{P}(A1 \mid D) + \mathbb{P}(A2 \mid D) = 1$
2. The ratio of the probabilities of $\mathbb{P}(A1 \mid D)$ and $\mathbb{P}(A2 \mid D)$ is still 2 : 1, which is the same as the original ratio. That is:

$$\frac{\mathbb{P}(A1 \mid D)}{\mathbb{P}(A2 \mid D)} = \frac{\mathbb{P}(A1)}{\mathbb{P}(A2)} = \frac{2}{1}$$

That is, when we assign conditional probabilities, we just need to let the sum of the partitions to be 1, and ***rescale*** them based on their original ratios.

The ratio of Leaf 1 : Leaf 2 : Leaf 3 : Leaf 4 in the original model (**Figure 1**) is:

$$(0.5 \times 0.5) : (0.5 \times 0.5) : (0.5 \times 0.25) : (0.5 \times 0.5) = 2 : 2 : 1 : 2$$

The ratio of Leaf 1 : Leaf 2 : Leaf 3 : Leaf 4 in the conditional model (**Figure 2**) is:

$$(p \times 0.5) : (p \times 0.5) : \left[(1-p) \times \frac{1}{3}\right] : \left[(1-p) \times \frac{2}{3}\right]$$

The ratio in the conditional model should also be 2 : 2 : 1 : 2. Solve for p, we have:

$$p = \frac{4}{7}$$