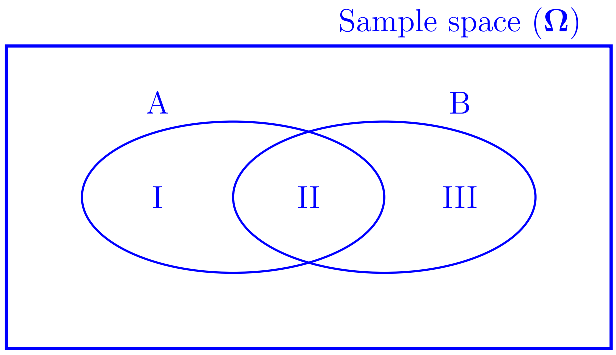
Southern University of Science And Technology
School of Life Sciences
BIO210 Biostatistics
(Spring, 2022)

Assignment 2
Due on 7th Oct, 11 p.m.

1. **Independence of event complements (5 points)**: If events A and B are independent, are A^c and B^c independent? Prove your answer.

   A^C and B^C are independent. Put the Venn diagram in your head:



Then, we could break certain events into disjoint events, and use the additivity axiom for the calculation.

Proof. We could divide the sample space into two disjoint events:

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

Then we have:

$$\begin{aligned} P(\Omega) &= P[(A^C \cap B^C) \cup (A \cup B)] \\ &= P(A^C \cap B^C) + P(A \cup B) \\ \Rightarrow P(A^C \cap B^C) &= P(\Omega) - P(A \cup B) \\ &= 1 - P(A \cup B) \end{aligned} \tag{1}$$

Now, we could divide $A \cup B$ into three disjoint events:

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

Therefore, we have:

$$P(A \cup B) = P(A \cap B^C) + P(A \cap B) + P(A^C \cap B)$$

Doing a bit algebraic manipulation by adding and subtracting $P(A \cap B)$, we have the following:

$$\begin{aligned}
P(A \cup B) &= P(A \cap B^C) + P(A \cap B) + P(A^C \cap B) + P(A \cap B) - P(A \cap B) \\
&= [P(A \cap B^C) + P(A \cap B)] + [P(A^C \cap B) + P(A \cap B)] - P(A \cap B)
\end{aligned} \tag{2}$$

Since $A \cap B^C$ and $A \cap B$ are disjoint, and $A^C \cap B$ and $A \cap B$ are disjoint. **Equation 2** becomes:

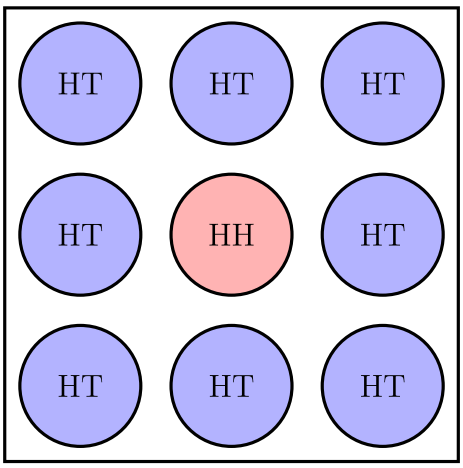
$$\begin{aligned}
P(A \cup B) &= P[(A \cap B^C) \cup (A \cap B)] + P[(A \cap B) \cup (A^C \cap B)] - P(A \cap B) \\
&= P(A) + P(B) - P(A \cap B)
\end{aligned} \tag{3}$$

Now, put **Equation 3** into **Equation 1**, and using the fact that A and B are independent, implying $P(A \cap B) = P(A) \cdot P(B)$, we have:

$$\begin{aligned}
P(A^C \cap B^C) &= 1 - [P(A) + P(B) - P(A \cap B)] \\
&= 1 - P(A) - P(B) + P(A) \cdot P(B) \\
&= [1 - P(A)] \cdot [1 - P(B)] \\
&= P(A^C) \cdot P(B^C)
\end{aligned}$$

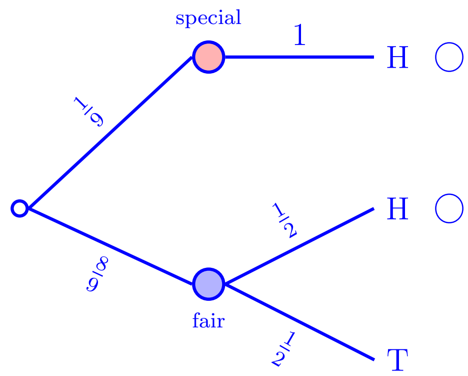
By definition, A^C and B^C are independent. $\square$

2. A coin box with strange coins: Li Lei has a coin box that contains 9 coins. 8 of them are fair, standard coins (heads and tails) and 1 coin has heads on both sides.



2.1) **(5 points)** If Li Lei randomly chooses a coin and flip it, what is the probability of getting a head?

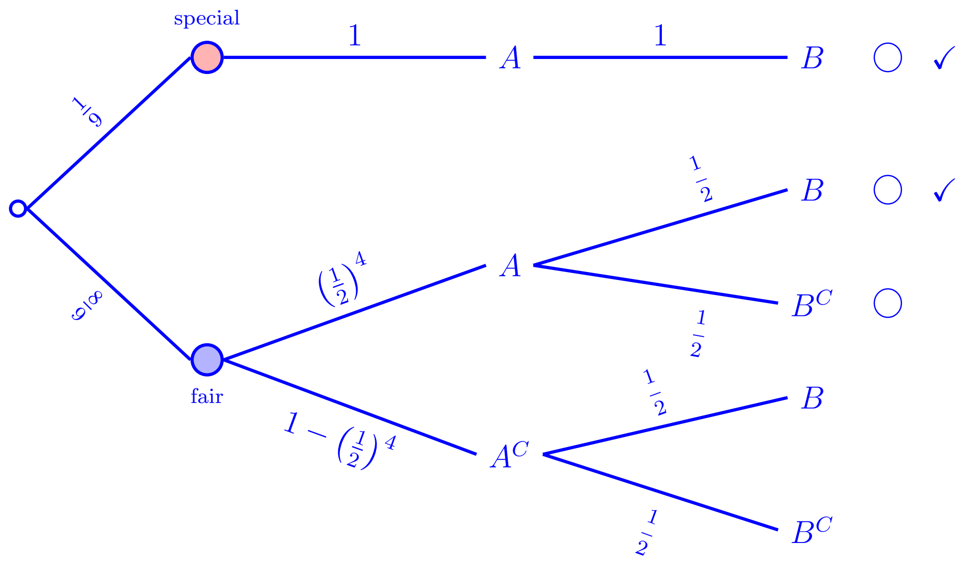
The answer is $\mathbf{\frac{5}{9}}$.
Build the following tree to help the calculation:



$$
\begin{aligned}
P(\text{getting a head}) &= \text{the sum of the two} \bigcirc \text{ branches} \\
&= \frac{1}{9} \times 1 + \frac{8}{9} \times \frac{1}{2} \\
&= \frac{5}{9}
\end{aligned}
$$

2.2) **(10 points)** Li Lei randomly chose a coin without looking at it. He flipped it four times, and his friend Han Meimei told him that he got four heads. Now, if Li Lei flip the coin again, what is the probability of getting a head?

The answer is $\mathbf{\frac{5}{6}}$:
You can approach this problem by different ways. The first way is just use brutal force to build a tree for the full model from scratch. Let events $\boldsymbol{A}$ = { The first four flips are all heads } and $\boldsymbol{B}$ = { The fifth flip is head }. Now construct the tree as follows:



Now that we know that event A has occurred, we are living in a universe in the ◯ branches, and the ✓ branches are what we are interested in. Therefor the probability is simply the ratio of the two:

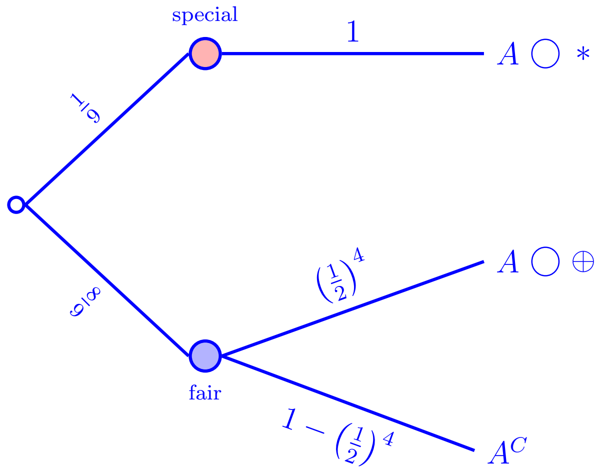
$$P(\text{the next flip is H} \mid \text{observing four heads}) = \frac{P(\checkmark)}{P(\bigcirc)}$$

$$= \frac{\frac{1}{9}\times 1\times 1+\frac{8}{9}\times\left(\frac{1}{2}\right)^4\times\frac{1}{2}}{\frac{1}{9}\times 1\times 1+\frac{8}{9}\times\left(\frac{1}{2}\right)^4\times\frac{1}{2}+\frac{8}{9}\times\left(\frac{1}{2}\right)^4\times\frac{1}{2}} = \frac{5}{6}$$

The second way of solving the problem is by our intuition and interpretation of the prior and posterior probability.

Intuitively, the probability of getting a head the next time depends on the probability of the coin at hand is fair or special. If we do not have any information, there is $\frac{1}{9}$ of the chance of having the special coin and $\frac{8}{9}$ of the chance of having a fair coin. Those are our prior probabilities. Then this problem is exactly the same as **2.1)**. However, we now know the first four flips are all heads. This fact that already happened provides new information, and we should use this to update our prior beliefs to get the probability of the coin at hand being fair or special (posterior probabilities). Then we should use those posterior probabilities to calculate the probability of getting a head for the next flip.

With that in mind, we could first calculate the probability of having fair or special coin at hand given that the first four flips are all heads. This is the tree:

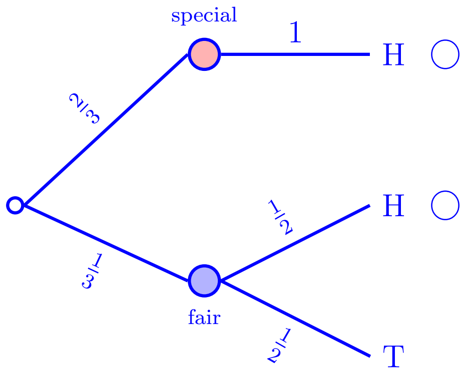


Apparently, we are living in a universe of the ◯ branches. Now we could update our beliefs about the probability of the coin at hand being fair or special:

$$P(\text{special} \mid A) = \frac{P(*)}{P(\bigcirc)} = \frac{\frac{1}{9}\times 1}{\frac{1}{9}\times 1 + \frac{8}{9}\times\left(\frac{1}{2}\right)^4} = \frac{2}{3}$$

$$P(\text{fair} \mid A) = \frac{P(\oplus)}{P(\bigcirc)} = \frac{\frac{8}{9}\times\left(\frac{1}{2}\right)^4}{\frac{1}{9}\times 1 + \frac{8}{9}\times\left(\frac{1}{2}\right)^4} = \frac{1}{3}$$

Now we could use the new calculation as our prior to calculate the probability of the next flip being H:

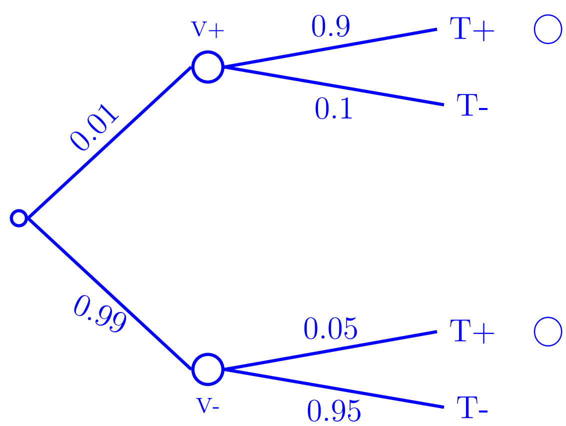


Now the probability of getting a head in the next flip after seeing four heads is basically $P(\bigcirc) = \frac{2}{3}\times 1 + \frac{1}{3}\times\frac{1}{2} = \frac{5}{6}$, which is the same as the previous brutal force answer.

3. Virus detection kits: It is known that the ***prevalence*** of a virus is 0.01. There are two detection kits for this particular virus used by many hospitals. They are produced by different manufacturers and based on completely different assays. **Kit A** has a ***sensitivity*** of 0.9 and a ***specificity*** of 0.95. **Kit B** has a ***sensitivity*** of 0.95 and a ***specificity*** of 0.9. A random person goes to a hospital to get tested by both kits **A** and **B**. Let events $\boldsymbol{E_1}$ = { Kit **A** shows a positive result }, $\boldsymbol{E_2}$ = { Kit **B** shows a positive result } and $\boldsymbol{E_3}$ = { The person carries the virus }.

3.1) **(10 points)** Calculate $P(\boldsymbol{E_1})$

As always, build the tree to help the visualisation:



With that, we could calculate the probability of the ○ branches:

$$P(E_1) = P(\text{sum of two } \bigcirc) = 0.01 \times 0.9 + 0.99 \times 0.05 = 0.0585$$

3.2) **(10 points)** Calculate $P(\boldsymbol{E_2})$

Similar to **3.1)**, simply replace the number based on the sensitivity and specificity of **Kit B**, we could get:

$$P(E_2) = 0.01 \times 0.95 + 0.99 \times 0.1 = 0.1085$$

3.3) **(10 points)** Are events $\boldsymbol{E_1}$ and $\boldsymbol{E_2}$ independent? Why or why not? You can use the mathematical formula to prove your conclusion, or you can also use your own words to describe your intuition about this question.

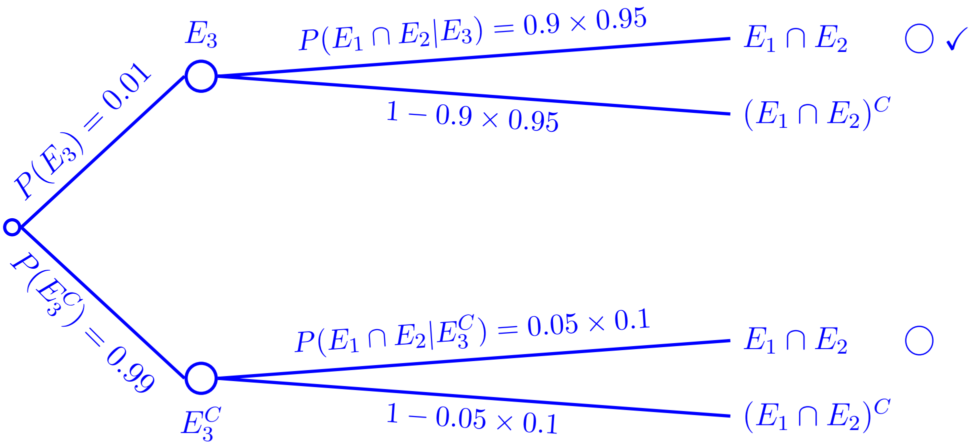
The following two answers are both correct:

Answer 1: E_1 and E_2 are NOT independent.

Answer 2: If we know the ground truth about whether the random person carry the virus or not, then E_1 and E_2 are independent. If we don't, then E_1 and E_2 are NOT independent.

Intuitive explanation: If we know the ground truth about the virus status of the person, then no matter what other information we have, it should not change our belief, because we know for a fact that the person carries the virus (or not). That is a fact that already happened. Then, since the two kits are produced by different manufacturers and based on completely different assays, they are independent. However, if we don't know the ground truth, then the event that a kit shows a positive result depends on the probability of the person carries the virus. The higher the chance that a person carries the virus, the more likely a kit will show a positive results. Before we do any tests, the best guess for a person to carry the virus is the prevalence. If a kit shows a positive result, this is a piece of new information that changes our belief about how likely the person carries the virus. It increases the chance that the person carries the virus. Now if a second test is done, the probability will not be the same as before when no tests are done. Therefore, E_1 and E_2 are NOT independent.

Use the definition of independence: We already know $P(E_1)$ and $P(E_2)$ from the previous two questions, now we only need to check if $P(E_1 \cap E_2) = P(E_1) \cdot P(E_2)$ is correct or not. Again, let's build trees:



Now it is obvious that:

$$\begin{aligned} P(E_1 \cap E_2) &= P(\bigcirc \text{ branches}) = 0.01 \times 0.9 \times 0.95 + 0.99 \times 0.05 \times 0.1 \\ &= 0.0135 \neq P(E_1) \cdot P(E_2) \end{aligned}$$

Therefore, they are NOT independent.

3.4) **(10 points)** Calculate $P(\boldsymbol{E_3} \,|\, \boldsymbol{E_1} \cap \boldsymbol{E_2})$

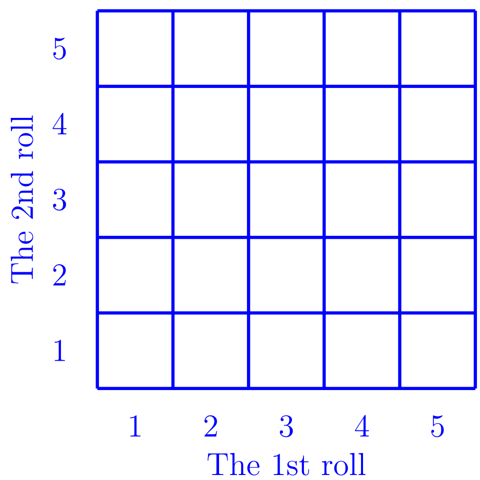
Before we actually makes the calculation, we should always start with intuition. What does this probability mean? It tells us if both kits show positive results, what is the probability of the person carries the virus. Our intuition tells us that if two different kits both show positive results, we should be kind of confident that the person carries the virus. Therefore, this probability should be much higher than the one we calculated during the lecture, which only one kit shows a positive result. Use the tree built in **3.3)**, we get:

$$P(E_3 \,|\, E_1 \cap E_2) = \frac{P(\checkmark)}{P(\bigcirc)} = \frac{0.01 \times 0.9 \times 0.95}{0.01 \times 0.9 \times 0.95 + 0.99 \times 0.05 \times 0.1} = \frac{19}{30}$$

4. Five-sided die: You have a fair five-sided die. The sides of the die are numbered from 1 to 5. Each die roll is independent of all others, and all faces are equally likely to come out on top when the die is rolled. Suppose you roll the die twice.

4.1) **(5 points)** Let event $\boldsymbol{A}$ = { the total of two rolls is 10 }, event $\boldsymbol{B}$ = { at least one roll resulted in 5 } and event $\boldsymbol{C}$ = { at least one roll resulted in 1 }. Are events $\boldsymbol{A}$ and $\boldsymbol{B}$ independent? Are events $\boldsymbol{A}$ and $\boldsymbol{C}$ independent?

Draw the following picture to have a sense of the sample space and each event:



It is relatively easy to see that: $P(A) = \frac{1}{25}$, $P(B) = \frac{9}{25}$, $P(C) = \frac{9}{25}$, $P(A \cap B) = \frac{1}{25}$ and $P(A \cap C) = 0$. Therefore, A and B are not independent. A and C are not independent, either.

4.2) **(5 points)** Let event $\boldsymbol{D}$ = { the total of two rolls is 7 }, event $\boldsymbol{E}$ = { the difference between the two roll outcomes is exactly 1 } and event $\boldsymbol{F}$ = { the second roll resulted in a higher number than the first roll }. Are events $\boldsymbol{E}$ and $\boldsymbol{F}$ independent? Are events $\boldsymbol{E}$ and $\boldsymbol{F}$ independent given that event $\boldsymbol{D}$ has occurred?

Using the figure above, it is straightforward to calculate: $P(D) = \frac{4}{25}$, $P(E) = \frac{8}{25}$, $P(F) = \frac{10}{25}$, $P(E \cap F) = \frac{4}{25}$. Now, we know that E and F are not independent. For the 2nd question, first we have:

$$P(E|D) = \frac{|E \cap D|}{|D|} = \frac{2}{4} = \frac{1}{2}$$

Then, we have:

$$P(F|D) = \frac{|F \cap D|}{|D|} = \frac{2}{4} = \frac{1}{2}$$

Finally, we have:

$$P(E \cap F|D) = \frac{|E \cap F \cap D|}{|D|} = \frac{1}{4}$$

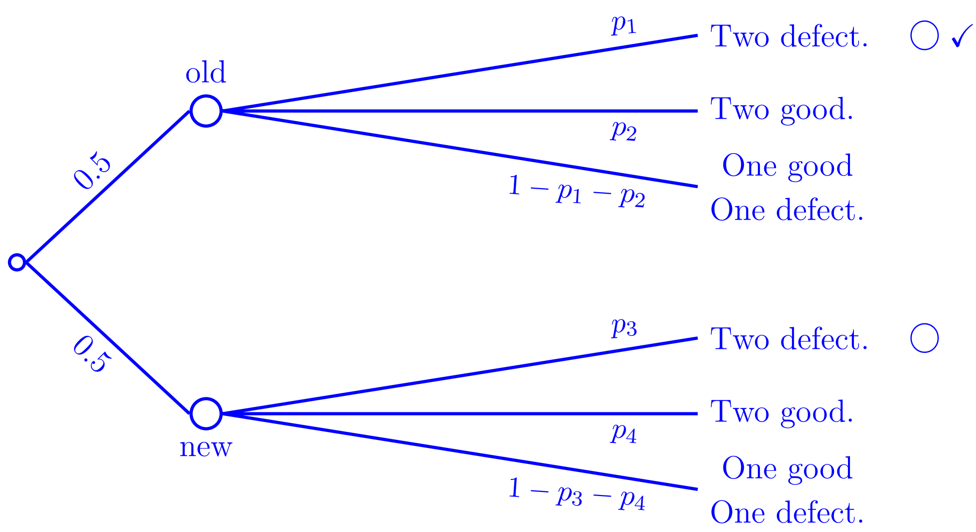
Now, we know that $P(E \cap F|D) = P(E|D) \cdot P(F|D)$. Therefore, E and F are independent given that D has occurred.

5. Buying CD-ROMs: The local computer store is having a blowout CD-ROM sale, because everything now is digital and on the cloud. It seems very few people actually need CD-ROMs. The store wants to clear out the CD-ROMs they have. The store has 500 old CD-ROMs, and 1500 new CD-ROMs in stock. The problem is that 15% of the old CD-ROMs are defective, and 5% of the new ones are defective as well. You can assume that CD-ROMs are selected at random when an order comes in. You are the first customer since the sale was announced.

5.1) **(5 points)** You flip a fair coin once to decide whether to buy old or new CD-ROMs. You order two CD-ROMs of the same type, chosen based on the outcome of the coin toss. What is the probability that they will both be defective?

5.2) **(5 points)** Given that both CD-ROMs turn out to be defective, what is the probability that they were old CD-ROMs?

We could build the tree for our problem:



Remember p_1, p_2, p_3, p_4 are all conditional probabilities based on the previous events. Using discrete uniform law, we could calculate:

$$p_1 = \frac{\binom{75}{2}}{\binom{500}{2}} \approx 0.15^2 = 0.0225$$

$$p_3 = \frac{\binom{75}{2}}{\binom{1500}{2}} \approx 0.05^2 = 0.0025$$

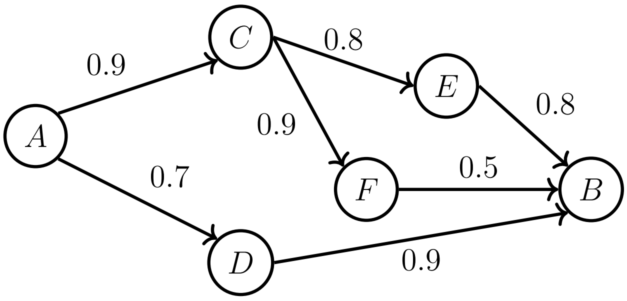
Therefore, we have:

$$P(\text{both defect.}) = P(\bigcirc \text{ branches}) \approx 0.5p_1 + 0.5p_3 = 0.0125$$

With the tree, it is also straightforward to see:

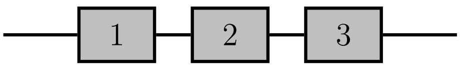
$$P(\text{old}|\text{two defect.}) = \frac{P(\checkmark)}{P(\bigcirc)} = \frac{0.5 \times p_1}{0.5p_1 + 0.5p_3} \approx \frac{0.01125}{0.0125} = 0.9$$

6. Network reliability and signalling pathway (17 points): Let's start with a simple computer network connecting two nodes from A to B through intermediate nodes C, D, E, F as shown in the picture below:

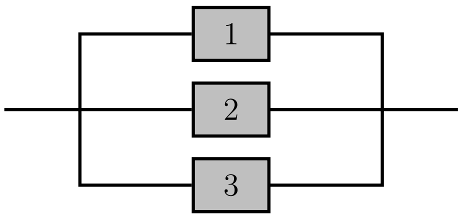


For every pair of directly connected nodes, say i and j, there is a given probability p_{ij} that the link from i to j is up. It is generally of interest to calculate the probability that there is a path connecting A and B in which all links are up.

This is a typical problem of assessing the reliability of a system consisting of components that can fail ***independently***. You will encounter this type of system very often in real life. Such a system can often be divided into subsystems, where each subsystem consists in turn of several components that are connected either in series or in parallel, like shown below:



Series connection



Parallel connection

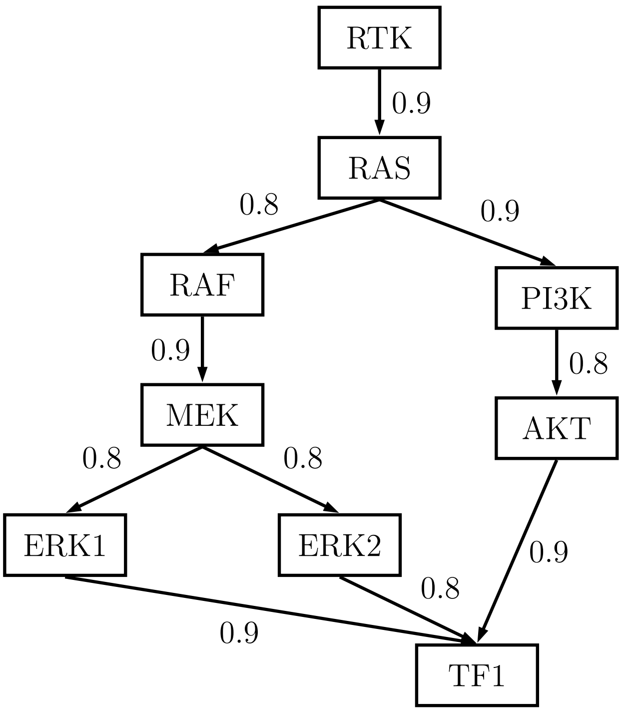
Let a subsystem consist of components $1, 2, 3, ..., m$, and let p_i be the probability that component i is up ("succeeds"). Then a series subsystem succeeds if ***all*** of its components are up, so its probability of success is the product of the probabilities of success of the corresponding components, *i.e.*

$$P(\text{series subsystem succeeds}) = \prod_{i=1}^{m} p_i = p_1\, p_2\, p_3 \cdots p_m$$

A parallel subsystem succeeds if ***any one*** of its components succeeds, so its probability of failure is the product of the probabilities of failure of the corresponding components, *i.e.*

$$\begin{aligned}P(\text{parallel subsystem succeeds}) &= 1 - P(\text{parallel subsystem fails})\\ &= 1 - \prod_{i=1}^{m}(1-p_i)\\ &= 1-(1-p_1)(1-p_2)(1-p_3)\cdots(1-p_m)\end{aligned}$$

Now let's look at a signalling pathway inside a cell, which is very important for the functions of a cell. The signalling pathway consists of a cascade of protein phosphorylation events. Many proteins are only activated when they are phosphorylated. In the following diagram depicting a simplified version of the MAP and PI3 kinase pathways, "$A \to B$" means "when A is activated, it has a probability of phosphorylating B", and the probability of the phosphorylation is indicated on the edge:



Assume all phosphorylation events are independent and all proteins shown in the above picture must be activated first before they can phosphorylate their downstream proteins to activate them. Now, the protein RTK is activated, what is the probability that the protein TF1 is activated ?

The answer is 0.7883039232, some degree of rounding errors is allowed.

First, we could divide the entire pathway into a series connection with two components:

$$P(RTK \to TF1) = P(RTK \to RAS) \cdot P(RAS \to TF1) \\ = 0.9 \times P(RAS \to TF1) \tag{4}$$

The route $RAS \to TF1$ has a parallel configuration, simply denoted as "left" and "right" here. Therefore:

$$P(RAS \to TF1) = 1 - [1 - P(\text{left})][1 - P(\text{right})] \tag{5}$$

$P(\text{right})$ is easy to calculate:

$$P(\text{right}) = P(RAS \to PI3K) \cdot P(PI3K \to AKT) \cdot P(AKT \to TF1) \\ = 0.9 \times 0.8 \times 0.9 = 0.648 \tag{6}$$

$P(\text{left})$ requires some extra calculations, but it is relatively straghtforward:

$$P(\text{left}) = P(RAS \to RAF) \cdot P(RAF \to MEK) \cdot P(MEK \to TF1) \\ = 0.8 \times 0.9 \times P(MEK \to TF1) = 0.72 \times P(MEK \to TF1) \tag{7}$$

Now we only need to get $P(MEK \to TF1)$ which consists of a parallel configuration:

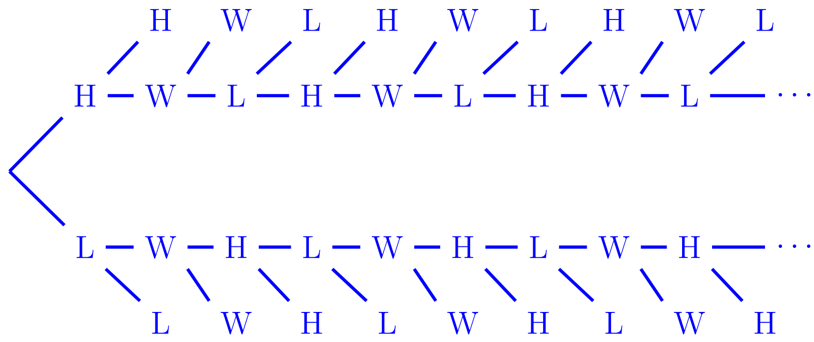
$$P(MEK \to TF1) = 1 - [1 - P(MEK \to ERK1 \to TF1)] \cdot \\ [1 - P(MEK \to ERK2 \to TF1)] \\ = 1 - (1 - 0.8 \times 0.9) \times (1 - 0.8 \times 0.8) \\ = 0.8992 \tag{8}$$

Take all those values back, and we get:

$$P(\text{TF1 is activated}) = 0.7883039232$$

7. * **Chess tournament (3 points)**: Han Meimei, Li Lei and Wang Gang play a chess tournament. The first game is played between Han Meimei and Li Lei. The player who sits out a given game plays next the winner of that game. The tournament ends when some player wins two successive games. Let a tournament history be the list of game winners, so for example $HWLHH$ corresponds to the tournament where Han Meimei won games 1, 4 and 5, Wang Gang won game 2, and Li Lei won game 3.

7.1) Provide a tree-based sequential description of a sample space where the outcomes are the possible tournament histories.



7.2) We are told that every possible tournament history that consists of k games has probability $1/2^k$, and that a tournament history consisting of an infinite number of games has zero probability. Check if this assignment of probability follows the three probability axioms we talked about during the lecture.

It is easy to see that $\frac{1}{2^k} \geqslant 0$, which satisfy the non-negativity axiom. The additivity axiom is also obvious. Now let's check the normalisation axiom, that is, we need to check if the sum of the probability of each possible outcomes in the sample space is 1 or not.

Note for very number that k can take (apparently $k \geqslant 2$), we always have a symmetric scenario: one outcome comes from the upper tree and the other outcome comes from the lower tree. They have the same probability. *e.g.*

$$k = 2 : P(HH) = \frac{1}{2^2} \text{ and } P(LL) = \frac{1}{2^2}$$

$$k = 3 : P(HWW) = \frac{1}{2^3} \text{ and } P(LWW) = \frac{1}{2^3}$$

$$k = 4 : P(HWLL) = \frac{1}{2^4} \text{ and } P(LWHH) = \frac{1}{2^4}$$

$$\vdots$$

Therefore, the sum of the probability of all outcomes is:

$$P(\Omega) = \cdot \sum_{k=2}^{\infty} 2 \cdot \frac{1}{2^k} = \lim_{n\to\infty} \left[\frac{1}{2} \cdot \frac{1-(\frac{1}{2})^n}{1-\frac{1}{2}} \right] = 1$$

Note: the sum of a geometric sequence is $a_1 \cdot \dfrac{1-q^n}{1-q}$, where a_1 is the first number of the sequence, and q is the common ratio.

7.3) Assuming the probability law from part 7.2) to be correct, find the probability that the tournament lasts no more than 5 games, and the probability for each of Han Meimei, Li Lei, and Wang Gang winning the tournament.

If you figured out the previous two questions, it will be easy to see:

$$\begin{aligned} P(\text{no more than } 5) &= P(k=2) + P(k=3) + P(k=4) + P(k=5) \\ &= 2 \cdot \frac{1}{2^2} + 2 \cdot \frac{1}{2^3} + 2 \cdot \frac{1}{2^4} + 2 \cdot \frac{1}{2^5} \\ &= \frac{15}{16} \end{aligned}$$

By looking at the tree, you can see the number of games is fixed for each winner:

$$\begin{aligned} P(\text{Han Meimei wins}) &= P(\text{win from upper tree}) + P(\text{win from lower tree}) \\ &= [\frac{1}{2^2} + \frac{1}{2^5} + \frac{1}{2^8} + \cdots] + [\frac{1}{2^4} + \frac{1}{2^7} + \frac{1}{2^{10}} + \cdots] \\ &= \frac{5}{14} \end{aligned}$$

$$\begin{aligned} P(\text{Li Lei wins}) &= P(\text{win from upper tree}) + P(\text{win from lower tree}) \\ &= [\frac{1}{2^4} + \frac{1}{2^7} + \frac{1}{2^{10}} + \cdots] + [\frac{1}{2^2} + \frac{1}{2^5} + \frac{1}{2^8} + \cdots] \\ &= \frac{5}{14} \end{aligned}$$

$$\begin{aligned} P(\text{Wang Gang wins}) &= P(\text{win from upper tree}) + P(\text{win from lower tree}) \\ &= [\frac{1}{2^3} + \frac{1}{2^6} + \frac{1}{2^9} + \cdots] + [\frac{1}{2^3} + \frac{1}{2^6} + \frac{1}{2^9} + \cdots] \\ &= \frac{2}{7} \end{aligned}$$

Sanity check: $\frac{5}{14} + \frac{5}{14} + \frac{2}{7} = 1$, and note:

$$P(\text{Wang Gang wins}) < P(\text{Han Meimei wins}) = P(\text{Li Lei wins})$$

Wang Gang has some slight disadvantages by not playing the first game.