This is a draft version of the lecture notes. We aim to keep improving it but at the current stage it is most likely far from perfect. Please contact us if you notice any typos, errors, subtle points, or if you have any questions or suggestions for improvements.

maciej.lisicki[at]fuw.edu.pl
milosz.panfil[at]fuw.edu.pl

Maciej Lisicki & Miłosz Panfil
October 13, 2022

1 Introduction and discrete probability distributions

1.1 Propagation of rare diseases

We inherit a copy of each gene from each parent. Those copies are called *alleles*. In the simplest case we can imagine that alleles can be in two variants. We could say 1 and 0 or 1 and -1. In biology we usually say dominant and recessive. As an example let's denote A the dominant allele and a the recessive one. For such gene, in human population we can distinguish 3 *genotypes*: those with both dominant alleles AA, those with one copy of a recessive gene Aa, and those with two copies of a recessive gene aa.

Imagine now, that this gene is related to some rare genetic disease, in the sense that only people with both recessive genes are sick. The question is: how abundant such disease could be in the population?

We assume that each parent gives one of the alleles with equal probability. Let us see how the genes and the disease propagate in one generation. We can make a table

M / F	AA	Aa	aa
AA	$AA(1)$	$AA(1/2)$, $Aa(1/2)$	$Aa(1)$
Aa	$AA(1/2)$, $Aa(1/2)$	$AA(1/4)$, $Aa(1/2)$, $aa(1/4)$	$Aa(1/2)$, $aa(1/2)$
aa	$Aa(1)$	$Aa(1/2)$, $aa(1/2)$	$aa(1)$

Table 1. Possible combinations of genes in an offspring

With the help of this table we could compute the probability of a rare disease in a child knowing the genes of parents. But we are interested in the whole population. We have a population of N people with N_p, N_q and N_r of them having exactly AA, Aa and aa genes. Obviously $N_p + N_q + N_r = N$. It's convenient to introduce relative abundances $p = N_p/N$, $2q = N_q/N$ and $r = N_r/N$ such that $p + 2q + r = 1$.

The relative abundances have a *probabilistic* meaning. If we draw a random person from the population it would have the AA genotype with probability p and similarly for other genotypes.

With the help of the table let us write the relative abundances in the next generation assuming that initially we have p_0, q_0 and r_0. We have

$$p_1 = p_0^2 + 2p_0q_0 + q_0^2 = (p_0 + q_0)^2, \tag{1.1}$$

$$2q_1 = 2p_0q_0 + 2p_0r_0 + 2q_0^2 + 2q_0p_0 = 2(p_0 + q_0)(r_0 + q_0), \tag{1.2}$$

$$r_1 = q_0^2 + 2q_0r_0 + r_0^2 = (r_0 + q_0)^2. \tag{1.3}$$

This set of equations determines relative abundances in the first generation starting from the zeroth generation. What about further generations? We get a recursive relation in which generation $(n + 1)$ is obtained from the knowledge of the nth-generation

$$p_{n+1} = (p_n + q_n)^2, \tag{1.4}$$

$$q_{n+1} = (p_n + q_n)(r_n + q_n), \tag{1.5}$$

$$r_{n+1} = (r_n + q_n)^2, \tag{1.6}$$

given the initial data (p_0, q_0, r_0). We observe that this process does not depend on its past only on its current state. Such processes are called *Markovian* or *Markov process* after a Russian mathematician Andrey Markov (1856–1922).

In principle we could now try to solve this recursive problem. Problems of such type can be difficult or even impossible to solve exactly. However we might not need the full solution to answer our question. Recall, that the question is: how abundant are such rare diseases? Essentially we would like to know if we start with non-zero r_0 will the evolution (in the sense of time evolution, not biological) lead to $r_0 = 0$ or not? One way of addressing this problem is by looking at *stationary solutions*.

A stationary solution is a choice of (p_s, q_s, r_s) such that they do not change under the evolution. Therefore, they obey

$$p_s = (p_s + q_s)^2, \tag{1.7}$$

$$q_s = (p_s + q_s)(r_s + q_s), \tag{1.8}$$

$$r_s = (r_s + q_s)^2. \tag{1.9}$$

The solution is $p_sr_s = q_s^2$. From the stationary solution we learn that $r_s \neq 0$ as long as $q_s \neq 0$. We see that there are stationary solutions with the disease present, despite its "rareness".

An interesting question is how quickly the system approaches the stationary state. By looking at equations for p_1, q_1 and r_1 we see that $q_1^2 = p_1r_1$. Therefore it takes just one step. This simple solution allows us to express the stationary state through the initial data. We find

$$p_s = (p_0 + q_0)^2, \tag{1.10}$$

$$q_s = 2(p_0 + q_0)(r_0 + q_0), \tag{1.11}$$

$$r_s = (r_0 + q_0)^2. \tag{1.12}$$

Therefore it's enough to have a small presence of allele a in the population for the disease to persist.

1.2 Difficulties with probabilities

To illustrate some of the challenges with using probabilistic thinking we consider two problems.

1) Imagine that you have to make a yes/no decision and you would like to make it random and fair. You have a coin, but you don't know if the coin is fair. What's the procedure (algorithm) to make a fair decision with such coin?

2) Imagine that there are two boxes with numbered balls. One contains balls numbered from 1 to 10, the other contains balls numbered from 1 to 1000. Someone gives you a ball with number 7. What's the probability that it came from the second box?

This second example is used as an illustration of a doomsday argument. The argument goes like this (according to https://what-if.xkcd.com/65/):

> Humans will go extinct someday. Suppose that, after this happens, aliens somehow revive all humans who have ever lived. They line us up in order of birth and number us from 1 to N. Then they divide us divide them into three groups—the first 5%, the middle 90%, and the last 5%:
>
> Now imagine the aliens ask each human (who doesn't know how many people lived after their time), "Which group do you think you're in?"
>
> Most of them probably wouldn't speak English, and those who did would probably have an awful lot of questions of their own. But if for some reason every human answered "I'm in the middle group", 90% of them will (obviously) be right. This is true no matter how big N is.
>
> Therefore, the argument goes, we should assume we are in the middle 90% of humans. Given that there have been a little over 100 billion humans so far, we should be able to assume with 95% probability that N is less than 2.2 trillion humans. If it's not, it means we're assuming we're in 5% of humans—and if all humans made that assumption, most of them would be wrong.
>
> To put it more simply: Out of all people who will ever live, we should probably assume we're somewhere in the middle; after all, most people are.
>
> If our population levels out around 9 billion, this suggests humans will probably go extinct in about 800 years, and not more than 16,000.

1.3 Probabilites

Let's now formalize the concepts that we used in the previous section. To talk about probability we need two ingredients and one condition

1. We need a *sample space* Ω that is a set of all possible elementary outcomes. In the previous example the sample space were possible genotypes and $\Omega = \{"AA", "Aa", "aa"\}$.

2. With each element of the sample space we associate a number between $[0, 1]$. In the previous example these numbers where $p, 2q$ and r. This number we call *probability* that some random event happens. For example, p was the probability that a randomly drawn person from the population had a genotype AA.

3. Finally, since the sample space covers all possible outcomes the respective probabilties must sum up to 1. Hence $p+2q+r=1$, which we should read that randomly drawn person will have (with probability 1 a genotype AA or Aa or aa.

We often denote probabilities $p(X)$ where $X \in \Omega$. We then write

$$\sum_{X\in\Omega} p(X) = 1. \tag{1.13}$$

We can also talk about probabilities of events composed of elementary events. For example probability that someone has genotype "AA" or "Aa" is $p+2q$. This leads to a definition of set S which is a set of all possible subsets of Ω. To an element A of S we associate probability

$$P(A) = \sum_{X\in A} P(X). \tag{1.14}$$

The collection of probabilities associated with all elementary events will be called the *probability distribution*, $p_i = P(X_i)$. The knowledge of the entire probability distribution is often too detailed and instead we characterise the distribution, e.g. by its expected (mean) value μ – the most probable result of an experiment – and variance σ, the expected probable deviations from the mean. They are defined as

$$\mu = \langle X\rangle = \bar{X} = \sum_i x_i p_i, \tag{1.15}$$

$$\sigma^2 = \langle (X - \langle X\rangle)^2\rangle = \sum_i (x_i - \langle x\rangle)^2 p_i. \tag{1.16}$$

One can further characterise the distribution by calculating the so-called higher *moments* of the distribution

$$\langle x^k\rangle = \sum_i x_i^k p_i. \tag{1.17}$$

A distribution is fully characterised by its moments.

1.4 Conditional probability

Conditional probability, denoted $P(X|Y)$ is a probability of an event X knowing that event Y is true. For example, you throw a dice an the result is even. What is the probability that the result is 4?

The conditional probability obeys an important relation, *the law of total probability*

$$P(X) = \sum_{Y\in Part(\Omega)} P(X|Y)P(Y), \tag{1.18}$$

where $Part(\Omega)$ is some partition of the sample space Ω. In the dice example it was could be a partition into odd and even results.

Let's interpret now $P(X|Y)P(Y)$, it's a probability that X happens given that Y happened multiplied by probability that Y happens. So this is a probability that X and Y happen, also denoted $P(X\cup Y)$. We can write

$$P(X\cup Y) = P(X|Y)P(Y). \tag{1.19}$$

The *Bayes' theorem* theorem is an observation that the left hand side is symmetric in X and Y. Therefore

$$P(X|Y)P(Y) = P(Y|X)P(X), \tag{1.20}$$

usually stated as

$$P(Y) = \frac{P(Y|X)P(X)}{P(X|Y)}. \tag{1.21}$$

Let us see how the Bayes' theorem works in practice. We stick to genetic diseases and we assume take that 1% of people have a certain genetic disease and that there is a test that detects it with 90% of accuracy (true positive). There is also 5% chance that the test is wrong (false positive). Given that the test is positive what are the chances of the disease?

Let's denote by A chance of having a gene defect and by B a chance of positive test. We want to know $P(A|B)$. Using the Bayes theorem we write

$$\begin{aligned} P(A|B) &= \frac{P(B|A)P(A)}{P(B)} = \frac{P(B|A)P(A)}{P(B|A)P(A) + P(B|\bar{A})P(\bar{A})} \\ &= \frac{0.9 \cdot 0.01}{0.9 \cdot 0.01 + 0.05 \cdot 0.99} \approx 15\%. \end{aligned} \tag{1.22}$$

Monty Hall problem

Following Wikipedia (https://en.wikipedia.org/wiki/Monty_Hall_problem): *Suppose you're on a game show, and you're given the choice of three doors: Behind one door is a car; behind the others, goats. You pick a door, say No. 1, and the host, who knows what's behind the doors, opens another door, say No. 3, which has a goat. He then says to you, "Do you want to pick door No. 2?" Is it to your advantage to switch your choice?*