

The cutting problem

Jasper Dekoninck

Summary

In this very short paper a probability problem is introduced and solved. The problem seeks to get the probability that k numbers on an edge of a given length aren't separated if you were to cut the edge n times.

1 Introduction

The first time I came across this problem was in a biology class a few years back. More specifically we were learning about the chromosomal crossover: one of the final steps in reproduction were the chromosomes of the mother and father recombine with each other. When our teacher said that traits could be caused by multiple genes spread over the chromosome, one of my classmates wondered about the probability that the offspring (including grandchild and further) had the trait caused by these genes based on their position on the chromosome. In this paper, we will try to answer that question. More specifically, we'll answer the following question:

Let $a_1, \dots, a_k \in [0, 1]$ (the genes) and let $b_0 = 0 < b_1 < \dots < b_n < 1 = b_{n+1}$ (the cuts). What is the probability there exists an $i \in \{0, 1, 2, \dots, n\}$ so that $b_i \leq a_1, a_2, \dots, a_n \leq b_{i+1}$?

You may note that the problem we're going to study is not exactly the same as the original gene-based problem: in the new problem we don't take into account the fact that even when all numbers lie on the same piece of the chromosome, that piece still doesn't have to be the piece chosen for reproduction. However, just multiplying our result by $\frac{1}{2^n}$ results in the answer we're looking for.

2 A simplified version

First, we'll take a look at a simplified version of the problem. We'll assume $k = 2$ and $a_1 = 0$. In fact, this is the only version where we're going to find a simple algebraic expression for the probability. In the general case, we'll only be able to construct a recursive probability function. Calculating this recursive expression will prove to be quite difficult for larger values of n .

We'll call the probability of this problem $P(a_2, n)$ (as a function of a_2 and n , the number of cuts). We'll prove the following formula using induction:

$$P(a_2, n) = 1 + \sum_{k=0}^{n-1} \frac{(-1)^{k+1}}{k!} \cdot a_2 \cdot \ln^k(a_2)$$

For $n = 1$, the formula checks out. Indeed, for $n = 1$ it says that

$$P(a_2, 1) = 1 - a_2$$

and this is the probability that 0 and a_2 are not separated.

Now suppose the formula checks out for $n - 1$ cuts. Then we'll prove that it is correct for n cuts. Suppose that the first cut happens at $b_1 \geq a_2$. By scaling the axis so that b_1 is mapped to 1, we can determine the probability that the next $n - 1$ cuts don't separate a_2 from 0. This probability is $P(\frac{a_2}{b_1}, n - 1)$. The probability that $b_1 \geq a_2$ is $1 - a_2$. Now we can use an integral to solve the problem:

$$P(a_2, n) = (1 - a_2) \cdot \frac{1}{1 - a_2} \int_{a_2}^1 P\left(\frac{a_2}{s_1}, n - 1\right) db_1$$

Calculating this formula only takes a few steps:

$$\begin{aligned}
P(a_2, n) &= (1 - a_2) \cdot \frac{1}{1 - a_2} \int_{a_2}^1 P\left(\frac{a_2}{b_1}, n - 1\right) db_1 \\
&= \int_{a_2}^1 1 + \sum_{k=0}^{n-2} \frac{(-1)^{k+1}}{k!} \cdot \frac{a_2}{b_1} \cdot \ln^k\left(\frac{a_2}{b_1}\right) db_1 && \text{given} \\
&= \left[b_1 - \sum_{k=0}^{n-2} \frac{(-1)^{k+1}}{k!} \cdot a_2 \cdot \frac{\ln^{k+1}\left(\frac{a_2}{b_1}\right)}{k+1} \right]_{a_2}^1 && \text{calculate integral} \\
&= 1 - a_2 + \sum_{k=0}^{n-2} \frac{(-1)^{k+2}}{(k+1)!} \cdot a_2 \cdot \ln^{k+1}(a_2) \\
&= 1 - a_2 + \sum_{k=1}^{n-1} \frac{(-1)^{k+1}}{k!} \cdot a_2 \cdot \ln^k(a_2) \\
&= 1 + \sum_{k=0}^{n-1} \frac{(-1)^{k+1}}{k!} a_2 \cdot \ln^k(a_2) && \text{rewrite}
\end{aligned}$$

3 General version

We can make a simplification of the problem by only looking at the two ends of a_1, \dots, a_k . Say $a_1 < a_2 < \dots < a_k$. Then the probability that they are separated is just the probability that a_1 and a_k are separated. Note that our previous calculation can be extended so that k doesn't have to be 2. The only assumption in that calculation is thus that $a_1 = 0$. Now we won't make that assumption. Using the exact same principle as a minute ago (first cutting once, then scaling the number line) but now on both sides (not only for $b_1 \geq a_k$ but also for $b_1 \leq a_1$), we can find the following recursive formula:

$$P(a_1, a_k, n) = a_1 \cdot \frac{1}{a_1} \int_0^{a_1} P\left(\frac{a_1 - b_1}{1 - b_1}, \frac{a_k - b_1}{1 - b_1}, n - 1\right) db_1 + (1 - a_k) \cdot \frac{1}{1 - a_k} \int_{a_k}^1 P\left(\frac{a_1}{b_1}, \frac{a_k}{b_1}, n - 1\right) db_1$$

Or simplifying just a bit:

$$P(a_1, a_k, n) = \int_0^{a_1} P\left(\frac{a_1 - b_1}{1 - b_1}, \frac{a_k - b_1}{1 - b_1}, n - 1\right) db_1 + \int_{a_k}^1 P\left(\frac{a_1}{b_1}, \frac{a_k}{b_1}, n - 1\right) db_1$$

Note that if we define $P(a_1, a_k, 0) = 1$, the formula is also correct for $n = 1$ (namely $P(a_1, a_k, 1) = 1 - (a_k - a_1)$).

We can calculate $P(a_1, a_k, n)$ for the first values of n . The case $n = 3$ was calculated using [1].

$$P(a_1, a_k, 0) = 1$$

$$P(a_1, a_k, 1) = P(a_1, a_k, 0) - (a_k - a_1)$$

$$P(a_1, a_k, 2) = P(a_1, a_k, 1) - (a_k - a_1) + (a_k - a_1) \log(1 - a_1) + (a_k - a_1) \log(a_k)$$

$$P(a_1, a_k, 3) = P(a_1, a_k, 2) - \frac{a_k - a_1}{2} (\log^2(1 - a_1) + \log^2(a_k)) + (a_k - a_1) \left(\text{Li}_2(1 - a_k) + \text{Li}_2(a_1) - \text{Li}_2\left(\frac{a_k - 1}{a_1 - 1}\right) - \text{Li}_2\left(\frac{a_1}{a_k}\right) \right)$$

Li_2 is the dilogarithm function defined as $\text{Li}_2(x) = \sum_{k=0}^{\infty} \frac{x^k}{k^2}$.

4 Conclusion

In this very short paper, a problem about genetic probability was introduced. It was solved in closed form for a simplified case. However, for the general case only a recursive formula could be found.

References

- [1] Scherfgen, D. (2020), *Integral Calculator*. Accessed at 10 September 2020 on integral-calculator.com.
- [2] Chromosomal crossover. (2020, August 14). In *Wikipedia*. Accessed at 10 September 2020 on https://en.wikipedia.org/wiki/Chromosomal_crossover
- [3] Weisstein, Eric W. "Dilogarithm." From MathWorld—A Wolfram Web Resource. <https://mathworld.wolfram.com/Dilogarithm.html>