On the Probabilistic Method and Permutations

Maturitätsarbeit MNG Rämibühl

Valentin Imbach

Under supervision of Eric Fitze and Valentin Féray

Zürich, January 2019

Contents

1 Introduction

There is no doubt that probability theory has many direct applications in physics, economics and everyday life. However, during the mid-twentieth century mathematicians discovered a further and highly interesting potential of the theory: It can be used as a tool for proving theorems, which may have nothing to do at all with probabilities. This so-called "Probabilistic Method" was pioneered by the Hungarian Paul Erdős, one of the most influential mathematicians of the era. The method has developed into one of the most powerful and widely used strategies in the field of combinatorics. Nowadays, we know of many elegant results from all areas of mathematics, that can be proven using the probabilistic method.

1.1 Motivation

When we first deal with probability theory, we use the structure of mathematical objects and various counting arguments from combinatorics in order to find certain probabilities. However, what fascinates me about it, is when this process is switched around. Probability theory is then used as part of the chain of reasoning when proving results from combinatorics and other areas of mathematics. The mere idea of introducing randomness to deduce a non-random statement may sound paradoxical at first. The aim of this paper is to get across the thought process necessary for understanding, why the probabilistic method works and what makes it so powerful.

I had the ambition to find a proof for every single theorem used in my paper, and I succeeded to do so. For every new concept introduced, I have chosen an illustrating example, in order to help the reader visualise the individual ideas. Moreover, my intention was to keep the mathematical knowledge required to understand my argumentation on a level accessible for high school students.

In the first part, I talk about the basic idea of the probabilistic method, and introduce the reader to a range of useful concepts and strategies. To illustrate how they can be applied to more complex structures, I introduce the notion of permutations. Using the ideas established in the first part, I analyse three properties of permutations, investigating how they are distributed and how often they occur. Having found many probabilistic properties of permutations, the next chapter is about how we can use the specific concept of higher moments. I prove a number of well-known inequalities and, using higher moments, I deduce more concrete results. In the last chapter I then give some famous examples of theorems from different areas of mathematics, which can be approached using the probabilistic method. In these proofs, I use the ideas discussed throughout the paper.

1.2 Notation

It is assumed that the reader is familiar with the basic concepts of probability and set theory. I will use the following conventions:

$$\begin{aligned}
\mathbb{N} &:= \{1, 2, 3, \dots\} \\
\mathbb{N}_0 &:= \{0, 1, 2, 3, \dots\} \\
\mathbb{Z} &:= \{\dots, -2, -1, 0, 1, 2, \dots\} \\
\mathbb{R} &:= \text{The set of real numbers} \\
\mathbb{R}^+ &:= \{x \in \mathbb{R} \text{ where } x \geq 0\} \\
|A| &:= \text{The size of set } A \\
\lfloor x \rfloor &:= \text{The greatest number } n \in \mathbb{Z} \text{ with } n \leq x \\
\lceil x \rceil &:= \text{The least number } n \in \mathbb{Z} \text{ with } n \geq x
\end{aligned}$$

Additionally, I write $f(x) \sim g(x)$ if and only if

$$\lim_{x \to \infty} \frac{f(x)}{g(x)} = 1.$$

Throughout the paper, there are many occasions where I will talk about randomly picking an element from a set. Those random choices are always thought of as being uniform, which means that every element is equally likely to get chosen.

2 The Probabilistic Method

2.1 Outline

The probabilistic method is used for proving the existence of some mathematical object with certain properties. This is achieved by selecting some object at random and proving that the probability of it having the properties we want is strictly greater than 0. Although the method can get arbitrarily more complex, it always boils down to this basic principle.

First of all, I will introduce some crucial concepts from probability theory that many of you will already be familiar with.

2.2 Probability

Definition 1 (Probability). *We denote the probability that some event E is true by*

$$\mathbb{P}(E), \text{ where } 0 \leq \mathbb{P}(E) \leq 1.$$

An event in the mathematical context can be anything that either occurs or does not occur. As the concept of probability is rather intuitive, there is no need for a technical definition. The following example will illustrate the concept.

Example 1 (Probability). Consider the following events and their corresponding probabilities:

E	$\mathbb{P}(E)$
A dice rolls a six	$\frac{1}{6}$
$2 > 1$	1
A coin lands on heads	$\frac{1}{2}$
$0.5 \in \mathbb{N}$	0

Statements which are always true have a probability of 1, others that are impossible have a probability of 0.

2.3 Random Variables

To further quantify random behaviour and shorten the notation, we introduce the idea of random variables.

Definition 2 (Random Variables). *A random variable X is defined as a variable, whose value is determined by the outcome of a random phenomenon. For every possible value x, there is a corresponding probability that $X = x$. In particular, we know that*

$$\sum_x \mathbb{P}(X = x) = 1,$$

where the sum is over all possible values of X.

Some values for a random variable can be more likely than others, and since the variable has to take a certain value, it is clear why all the probabilities have to add up to 1. We call the set of probabilities which determine, how likely each outcome is to occur, the probability distribution of the random variable.

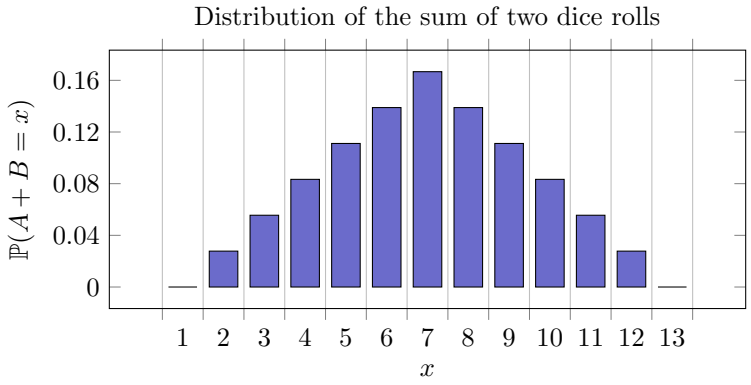
I will always denote random variables by capital letters. Furthermore, the corresponding lower case letter under the summation sign denotes the sum over all possible values of the variable. Note that while in this paper I will only focus on random variables in the real and integer numbers, the concept is much broader and the outcome of a random variable could be any kind of mathematical object.

Example 2 (Random Variables). Let us roll two dice and let the random variables A and B denote the number of pips showing on the first and second dice respectively. We can easily calculate the following:

E	$\mathbb{P}(E)$
$A = 3$	$\frac{1}{6}$
$B = 7$	0
$A = B$	$\frac{1}{6}$
$A + B = 8$	$\frac{5}{36}$
$A > B$	$\frac{5}{12}$

This is done by counting how many of the 36 possible combinations of A and B satisfy the desired properties.

In case of a single roll, all possible outcomes are equally likely and the probability distribution is constant. Yet, once we start adding the outcomes of the two dice, we find that some results are more likely than others:



Here, $\mathbb{P}(A+B=7)$ is 6 times the value of $\mathbb{P}(A+B=2)$.

2.4 Dependency

Another important idea to get familiar with is the concept of dependency. This is where probability theory gets more subtle and less intuitive.

Definition 3 (Dependency). *We say that two random variables are dependent, if and only if the probability distribution of one variable is influenced by the outcome of the other. We say that two variables are independent, if and only if they are not dependent.*

Understanding the shape of a certain probability distribution often turns out to be the key for deducing useful facts about the random system. One very intuitive property of a distribution is its "average". The next section will give a more rigorous definition of this concept.

2.5 Expected Value

As the name already gives away, the expected value of a random variable is a measure for what value should be expected.

Definition 4 (Expected Value). *Let $X \in \mathbb{R}$ be a random variable. The expected value of X is defined as*

$$\mathbb{E}(X) = \sum_x \Big(x \cdot \mathbb{P}(X = x) \Big).$$

The expected value is therefore an average over all possible values of the random variable, where each outcome is weighted by its corresponding probability. In this way, values which are more likely to be the outcome of the random variable, are treated as "more important" than other, less likely values. The expected value is also referred to as the mean of the probability distribution.

Example 3 (Rolling Dice). Let us roll a dice and let X again denote the number of pips showing. Using the definition of the expected value we find that

$$\mathbb{E}(X) = 1 \cdot \frac{1}{6} + 2 \cdot \frac{1}{6} + 3 \cdot \frac{1}{6} + 4 \cdot \frac{1}{6} + 5 \cdot \frac{1}{6} + 6 \cdot \frac{1}{6} = \frac{21}{6} = 3.5.$$

It is very important to note that the expected value does not have to be a possible outcome of the random variable itself. Therefore it is wrong to say that the expected value is the outcome with highest probability. It should rather be thought of as the average value of the outcomes over many tries. A more technical description of why this is true, is given by the so-called "Law of Large Numbers", a fundamental result in probability theory first stated by the Italian Gerolamo Cardano in the sixteenth century. Now for a more advanced example.

Application 1 (Coin Game). Consider the game where you flip a coin several times. For every time you flip heads, you receive 1 token. As soon as you flip tails for the first time, the game stops. What is the fair price for playing this game?

Due to its properties, the expected value is exactly what we need in order to answer this question. It is easy to see that the fair price for any game is given by the expected winnings of the game.

Imagine now that you agreed to play this game. Denote by T the number of tokens you win. Note that for all $t \in \mathbb{N}_0$, we have

$$\mathbb{P}(T = t) = \frac{1}{2^{t+1}},$$

since for you to win exactly t tokens, the coin has to land on heads t times in a row followed by one tail. It follows by definition that

$$\mathbb{E}(T) = \sum_t \Big(t \cdot \mathbb{P}(T = t) \Big) = \sum_{t=0}^{\infty} \frac{t}{2^{t+1}} = \frac{1}{4} + \frac{2}{8} + \frac{3}{16} + \frac{4}{32} + \cdots$$

All that remains to do is to evaluate this sum. Note that we can write

$$\begin{array}{rcccccccccc}
\mathbb{E}(T) & = & \frac{1}{4} & + & \frac{1}{8} & + & \frac{1}{16} & + & \frac{1}{32} & + & \frac{1}{64} & + & \cdots \\
 & & & + & \frac{1}{8} & + & \frac{1}{16} & + & \frac{1}{32} & + & \frac{1}{64} & + & \cdots \\
 & & & & & + & \frac{1}{16} & + & \frac{1}{32} & + & \frac{1}{64} & + & \cdots \\
 & & & & & & & + & \frac{1}{32} & + & \frac{1}{64} & + & \cdots \\
 & & & & & & & & & + & \frac{1}{64} & + & \cdots
\end{array}$$

Instead of summing everything at once, we will sum by columns first. In this way we merely need to evaluate geometric sums. Using the well-known formula, it directly follows that

$$\mathbb{E}(T) = \frac{1}{2} + \frac{1}{4} + \frac{1}{8} + \frac{1}{16} + \frac{1}{32} + \cdots = 1.$$

We conclude that the fair price for playing this game is exactly 1 token.

Theorem 1 (Linearity of Expectation). *Let $X, Y \in \mathbb{R}$ be two random variables. We know that*

$$\mathbb{E}(X + Y) = \mathbb{E}(X) + \mathbb{E}(Y)$$

and

$$\mathbb{E}(a \cdot X) = a \cdot \mathbb{E}(X)$$

for all $a \in \mathbb{R}$.

This first part of the theorem is, especially in the case where X and Y happen to be dependent, not intuitive at all and turns out to be a very strong tool.

Proof. Let $X, Y \in \mathbb{R}$ be two random variables. Using the definition of the expected value we deduce that

$$\begin{aligned}
\mathbb{E}(X+Y) &= \sum_x \sum_y \Big((x+y) \cdot \mathbb{P}(X=x, Y=y) \Big) \\
&= \sum_x \sum_y \Big(x \cdot \mathbb{P}(X=x, Y=y) \Big) + \sum_y \sum_x \Big(y \cdot \mathbb{P}(X=x, Y=y) \Big) \\
&= \sum_x \Big(x \cdot \sum_y \mathbb{P}(X=x, Y=y) \Big) + \sum_y \Big(y \cdot \sum_x \mathbb{P}(X=x, Y=y) \Big) \\
&= \sum_x \Big(x \cdot \mathbb{P}(X=x) \Big) + \sum_y \Big(y \cdot \mathbb{P}(Y=y) \Big) \\
&= \mathbb{E}(X) + \mathbb{E}(Y)
\end{aligned}$$

and reach the desired equality. The second part of the theorem follows immediately from the definition of the expected value.

□

Note that the first step of the proof makes use of the fact that we can split the summation over all values of $x+y$ into a sum over x and y separately where certain values of $x+y$ appear multiple times. Moreover, the theorem can easily be extended to sums of more than two terms by just applying it multiple times.

2.6 Indicator Variables

Linearity of expectation is often used in conjunction with the concept of indicator variables.

Definition 5 (Indicator Variables). *Let E be some event. The indicator variable of the event E is the random variable defined as*

$$\begin{cases} 1, & \text{if the event } E \text{ occurs.} \\ 0, & \text{otherwise.} \end{cases}$$

With this definition it is easy to see that the expected value of the indicator variable is equal to the probability of the corresponding event occurring: Let I be some indicator variable. It follows that

$$\mathbb{E}(I) = 0 \cdot \mathbb{P}(I=0) + 1 \cdot \mathbb{P}(I=1) = \mathbb{P}(I=1),$$

since 0 and 1 are the only possible states of I.

Given a random variable X which counts how many of n events occurred, we can write

$$X = \sum_{k=1}^{n} I_k,$$

where I_k is the indicator variable of the k-th event. Applying the definition and linearity of the expected value, we find that

$$\mathbb{E}(X) = \mathbb{E}\left(\sum_{k=1}^{n} I_k\right) = \sum_{k=1}^{n} \mathbb{E}(I_k) = \sum_{k=1}^{n} \mathbb{P}(I_k = 1).$$

This is just the sum of probabilities, that each event occurs. When all the events are pairwise independent, this result seems intuitive. More surprising is the fact, that this still works when the n events are dependent in any conceivable way.

Application 2 (Cowboy Duel). There are $n \geq 2$ cowboys on a field duelling each other. Simultaneously, every cowboy picks one of the other $n-1$ at random and shoots at him. We assume that every shot is lethal. What is the expected number of cowboys surviving the duel?

Let C be the number of surviving cowboys. We number the cowboys from 1 to n and define indicator variables as

$$I_k = \begin{cases} 1, & \text{if the } k\text{-th cowboy survives} \\ 0, & \text{otherwise} \end{cases}$$

for all $1 \leq k \leq n$. It follows that

$$\mathbb{E}(C) = \mathbb{E}\left(\sum_{k=1}^{n} I_k\right) = \sum_{k=1}^{n} \mathbb{E}(I_k) = \sum_{k=1}^{n} \mathbb{P}(I_k = 1).$$

The probability to survive is the same for every cowboy. It is given by

$$\mathbb{P}(I_k = 1) = \left(1 - \frac{1}{n-1}\right)^{n-1}$$

for all $1 \leq k \leq n$. We find that

$$\mathbb{E}(C) = n \cdot \left(1 - \frac{1}{n-1}\right)^{n-1}.$$

Note that as n gets very large, this expression tends towards $\frac{n}{e}$, where e denotes Euler's number. This is due to the well-known fact that

$$\lim_{x \to \infty} \left(1 - \frac{1}{x}\right)^{x} = \frac{1}{e}.$$

We conclude: As the number of cowboys increases, the fraction of survivors, and therefore also the probability for any particular cowboy to survive, gets closer and closer to $1/e \approx 0.368$.

Given the expected value of a random variable, we can start making conclusions about the structure of its probability distribution. In particular, the following result is very useful.

Lemma 1. *Let $X \in \mathbb{R}$ be a random variable.*

$$\mathbb{P}\Big(X \geq \mathbb{E}(X)\Big) > 0 \text{ and } \mathbb{P}\Big(X \leq \mathbb{E}(X)\Big) > 0.$$

This theorem states that not all possible outcomes of a random variable can be above or below its expected value.

Proof. Let $X \in \mathbb{R}$ be a random variable and let $\mathbb{E}(X) = \mu$. Assume that all possible values of X are below its expected value. In other words, let

$$\mathbb{P}\Big(X \geq \mu\Big) = 0.$$

By the definition of the expected value it follows that

$$\mu = \sum_{x} \Big(x \cdot \mathbb{P}(X = x)\Big) = \sum_{x<\mu} \Big(x \cdot \mathbb{P}(X = x)\Big) + \sum_{x \geq \mu} \Big(x \cdot \mathbb{P}(X = x)\Big).$$

Because the probability of $X \geq \mu$ is equal to zero, the second sum vanishes. We deduce that

$$\mu = \sum_{x<\mu} \Big(x \cdot \mathbb{P}(X = x)\Big) < \sum_{x<\mu} \Big(\mu \cdot \mathbb{P}(X = x)\Big) = \mu,$$

which is a clear contradiction. We therefore must conclude that our initial assumption was wrong and that

$$\mathbb{P}\Big(X \geq \mathbb{E}(X)\Big) > 0.$$

The opposite inequality follows similarly.

□

This theorem can be used to give basic proofs of existence, since it implies that there exists at least one possible outcome above and below the expected value. We will have a look at how to apply this strategy next.

Application 3 (Post Offices). Consider n houses and n post offices. Every house is usually connected to every post office by a road. Unfortunately, today $n-1$ of those roads are under construction and can not be used. We want to prove that it is still possible to assign every house to a different post office, connecting them directly by a functional road.

In the language of graph theory, we want to show that given a bipartite graph on n nodes with $n^2 - n + 1$ edges, there exists a perfect matching.

We pair the houses with the post offices randomly and let C denote the number of pairs which are connected by a functional road. After numbering the pairs from 1 to n, we define indicator variables as

$$I_k = \begin{cases} 1, & \text{if the } k\text{-th pair is connected} \\ 0, & \text{otherwise} \end{cases}$$

for $1 \leq k \leq n$. Again we write

$$\mathbb{E}(C) = \mathbb{E}\left(\sum_{k=1}^{n} I_k\right) = \sum_{k=1}^{n} \mathbb{E}(I_k) = \sum_{k=1}^{n} \mathbb{P}(I_k = 1).$$

For a random pair, the probability of it being connected by a road is given by the total number of functional roads divided by the number of possible pairs.

$$\mathbb{P}(I_k = 1) = \frac{n^2 - n + 1}{n^2}$$

for all $1 \leq k \leq n$. The expected value is therefore given by

$$\mathbb{E}(C) = n \cdot \frac{n^2 - n + 1}{n^2} = n - 1 + \frac{1}{n}.$$

Lemma 1 tells us that

$$\mathbb{P}\left(C \geq n - 1 + \frac{1}{n}\right) > 0.$$

Using the fact that C only takes values in $\mathbb{N}_0$ and $C \leq n$, we conclude

$$\mathbb{P}(C = n) > 0.$$

This implies that there is a nonzero chance of picking a pairing where all the pairs are connected. Therefore such a pairing exists.

In the previous example we were given a set of mathematical objects and we wished to show that there exists a subset of size s, where all elements satisfy some condition. We achieved this by selecting a random subset of size s and showing that on average almost all the elements of this subset satisfy our condition. Unfortunately, this strategy only works on rather basic configurations. If we want to prove more advanced results, we need a more sophisticated approach.

2.7 Alteration

Again assume that we are given a set of mathematical objects and we want to prove the existence of a subset of size s, where all elements satisfy some condition. Instead of selecting a random subset of size s, we will construct a subset by picking every element from the initial set with some probability p. From the resulting subset we will then delete all elements which interfere with our condition. If we can show that the expected number of remaining elements in our subset is greater or equal to s, we are done. Naturally, we will choose the value of p such that this expectancy is maximised.

Application 4 (Lonely Unicorns). Somewhere over the rainbow there live n unicorns. Assuming the average unicorn knows $d > 0$ others, we want to show that there is a group of at least $n/2d$ unicorns, of which no two know each other.

In the language of graph theory, we want to prove that given a graph on n nodes with average degree $d > 0$, we can find an independent set of size at least $n/2d$. If we approach this problem with the previous strategy, we will ultimately fail. When we try to select a group of $\lceil n/2d \rceil$ unicorns at random and to show that the expected number of acquainted pairs is smaller than 1, we find that

$$\mathbb{E}(\text{acquainted pairs}) = \binom{\lceil \frac{n}{2d} \rceil}{2} \cdot \frac{A}{\binom{n}{2}} = \frac{\lceil \frac{n}{2d} \rceil \cdot \left(\lceil \frac{n}{2d} \rceil - 1 \right)}{2} \cdot \frac{2 \cdot A}{n \cdot (n-1)}.$$

Here, A denotes the total number of acquainted pairs given by

$$A = \frac{n \cdot d}{2},$$

since every pair is counted twice. Putting everything together we get

$$\mathbb{E}(\text{acquainted pairs}) = \frac{n \cdot (\lceil \frac{n}{2d} \rceil - 1)}{4 \cdot (n-1)}.$$

To our disappointment we realise now that this leads nowhere, as choosing $n = 10$ and $d = 1$ results in a value of $10/9 > 1$, and our attempt of showing that the term is always less than 1 has failed.

To prove the statement, we need to make use of alteration. Instead of selecting a group of size $\lceil n/2d \rceil$, we will construct a group by selecting every unicorn with a fixed probability p. This group will contain some acquainted pairs that we do not want. For every such pair we will then just select one of the two unicorns and discard it.

In this way, we will end up with a group of unicorns, which contains no acquainted pair, of some expected size U dependent on p:

$$\mathbb{E}(U) = \mathbb{E}(\text{selected}) - \mathbb{E}(\text{discarded})$$

The individual parts can be calculated very easily. Let us number the unicorns from 1 to n and define indicator variables as

$$I_k = \begin{cases} 1, & \text{if the } k\text{-th unicorn was picked} \\ 0, & \text{otherwise} \end{cases}$$

for all $1 \leq k \leq n$. It now follows that

$$\mathbb{E}(\text{selected}) = \mathbb{E}\left(\sum_{k=1}^{n} I_k\right) = \sum_{k=1}^{n} \mathbb{E}(I_k) = \sum_{k=1}^{n} \mathbb{P}(I_k = 1) = n \cdot p.$$

To count the unicorns we have discarded, we must count the acquainted pairs where we picked both unicorns. As seen before, the number of acquainted pairs is given by

$$A = \frac{n \cdot d}{2}.$$

Let us again number all acquainted pairs from 1 to A and define new indicator variables as

$$I'_k = \begin{cases} 1, & \text{if the } k\text{-th acquainted pair was picked} \\ 0, & \text{otherwise} \end{cases}$$

for all $1 \leq k \leq A$. The expected number of unicorns we have to discard is now given by

$$\mathbb{E}(\text{discarded}) = \mathbb{E}\left(\sum_{k=1}^{A} I'_k\right) = \sum_{k=1}^{A} \mathbb{E}(I'_k) = \sum_{k=1}^{A} \mathbb{P}(I'_k = 1) = A \cdot p^2.$$

Putting everything together, we find that

$$\mathbb{E}(U) = n \cdot p - F \cdot p^2 = n \cdot p \cdot \left(1 - \frac{d \cdot p}{2}\right).$$

It is now time to choose the appropriate value of p. The expression is a quadratic in p and takes its maximum at $p = 1/d$. Since we want to maximise the expectation, we will choose this value and conclude that

$$\mathbb{E}(U) = \frac{n}{2d}.$$

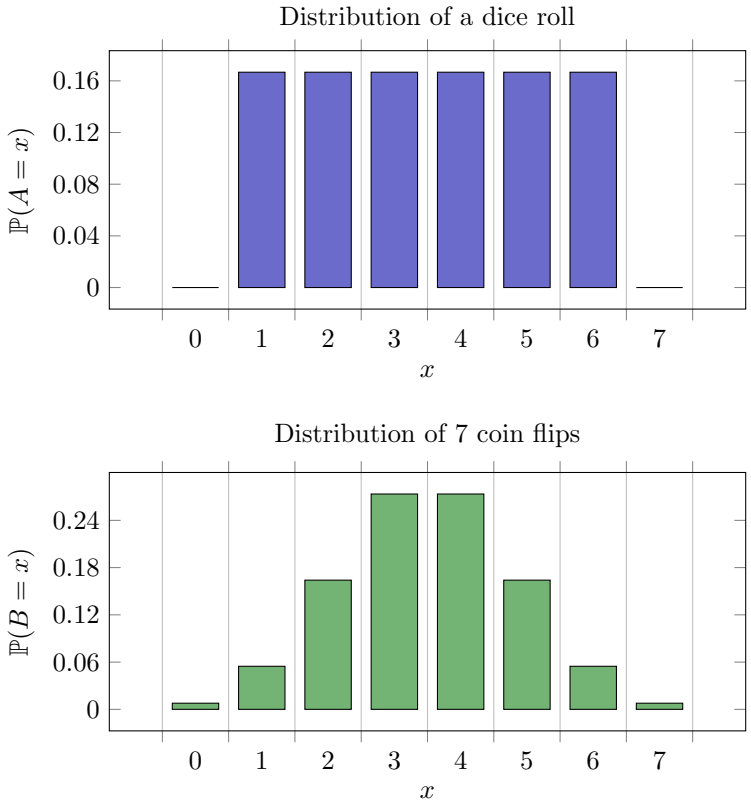
Therefore there must be some way of choosing a group of at least $n/2d$ unicorns with the property of having no acquainted pairs. In the last section of this paper (see p. 41) we will prove a theorem that gives a stronger version of this result.

2.8 Moments

Even if we know the expected value of a random variable, there is still a lot of ambiguity as to how the probability is distributed. To illustrate this, let us roll a dice and flip 7 coins. Let A be the number of pips showing on the dice and B the number of coins that landed on heads. It is easy to verify that the expected values of the two variables are in fact equal:

$$\mathbb{E}(A) = 3.5 = \mathbb{E}(B)$$

But when comparing the individual probabilities of all the outcomes, we see that the two random variables have very different distributions.



We now wish to be able to further characterise the shape of probability distributions. This can be achieved by using higher moments.

Definition 6 (Moments). *Let X be a random variable. The m-th moment of X is defined as*

$$\mathbb{E}\big(X^m\big)$$

for any $m \in \mathbb{N}_0$.

For every possible outcome of the random variable X, instead of just weighting it accordingly with its corresponding probability, we also raise X to the m-th power. Doing so, we get different information about the distribution for every $m \in \mathbb{N}_0$. Note that the 0-th moment is always equal to 1 and the first moment is just the expected value of the random variable. The more moments of a distribution we know, the more we can say about its structure as will be demonstrated later on.

Example 4 (Moments). Let us again roll a dice and flip seven coins and let A be the number of pips showing. With the definition of moments we can now calculate the following:

m	$\mathbb{E}\big(A^m\big)$	
1	$\frac{1}{6}\big(1+2+3+4+5+6\big)$	$= 3.5$
2	$\frac{1}{6}\big(1^2+2^2+3^2+4^2+5^2+6^2\big)$	≈ 16.17
3	$\frac{1}{6}\big(1^3+2^3+3^3+4^3+5^3+6^3\big)$	≈ 73.83

Similarly, let B be the number of heads we flipped and compute the following:

m	$\mathbb{E}\big(B^m\big)$	
1	$\frac{1}{128}\big(7\cdot 1+21\cdot 2+35\cdot 3+35\cdot 4+21\cdot 5+7\cdot 6+1\cdot 7\big)$	$= 3.5$
2	$\frac{1}{128}\big(7\cdot 1^2+21\cdot 2^2+35\cdot 3^2+35\cdot 4^2+21\cdot 5^2+7\cdot 6^2+1\cdot 7^2\big)$	$= 14$
3	$\frac{1}{128}\big(7\cdot 1^3+21\cdot 2^3+35\cdot 3^3+35\cdot 4^3+21\cdot 5^3+7\cdot 6^3+1\cdot 7^3\big)$	$= 61.25$

Whereas the first moment is equal for both random variables, higher moments allow us to distinguish between the two probability distributions.

A more intuitive tool to capture information about the shape of a probability distribution is the so-called variance of a random variable. This value tells us, how "spread-out" the distribution is by taking the expected value of the squared distance to the mean.

Definition 7 (Variance). *Let X be a random variable. The variance of X is defined as*

$$\mathrm{Var}(X) = \mathbb{E}\left(\Big(X - \mathbb{E}(X)\Big)^2\right).$$

Similar to what we did with the expected value, this concept can be extended by using higher exponents than 2. While the variance and its extensions describe the distance of the distribution to the mean, moments look at the distance to 0. This is why the variance and its extensions are sometimes called "Moments around the mean". Despite this idea being more intuitive, moments around 0 allow for cleaner calculations and are easier to work with. Luckily, there is a nice connection between the two.

Lemma 2 (Connecting Variance and Moments). *Let $X \in \mathbb{R}$ be a random variable. We know that*

$$\mathrm{Var}(X) = \mathbb{E}\big(X^2\big) - \mathbb{E}(X)^2.$$

The proof follows from the linearity of the expected value.

Proof. Let X be a random variable. By definition we deduce that

$$\mathrm{Var}(X) = \mathbb{E}\left(\Big(X - \mathbb{E}(X)\Big)^2\right) = \mathbb{E}\Big(X^2 - 2 \cdot X \cdot \mathbb{E}(X) + \mathbb{E}(X)^2\Big)$$

$$= \mathbb{E}\big(X^2\big) - 2 \cdot \mathbb{E}(X)^2 + \mathbb{E}(X)^2 = \mathbb{E}\big(X^2\big) - \mathbb{E}(X)^2,$$

keeping in mind that $E(X)$ is just a real constant.

□

It is easy to see, that there is a similar connection also for higher exponents. We conclude that higher moments around 0 carry all the information necessary for finding higher moments around the mean. As the two concepts are now shown to be equivalent, we will focus our attention on finding moments around 0.

Example 5 (Variance). Let A and B be defined as before. We find:

$$\mathrm{Var}(A) = \mathbb{E}\big(A^2\big) - \mathbb{E}(A)^2 \approx 16.17 - 12.25 = 3.92$$

$$\mathrm{Var}(B) = \mathbb{E}\big(B^2\big) - \mathbb{E}(B)^2 = 14 - 12.25 = 1.75$$

Just as expected, the dice roll has a greater variance because its distribution is more "spread out", whereas the distribution of the coin tosses is more concentrated around the expected value.

3 Permutations

To illustrate how moments are calculated and how we can apply the other results already discussed, we will have a look at permutations, a very interesting concept which is used in all branches of mathematics.

Definition 8 (Permutations). *A permutation π of length $n \in \mathbb{N}$ is defined as a function from set $\{1, 2, ..., n\}$ onto itself where no element is hit twice. We write*

$$\pi = \pi(1)\pi(2)\cdots\pi(n).$$

Furthermore, we define $\mathcal{S}_n$ to be the set of all permutations of length n.

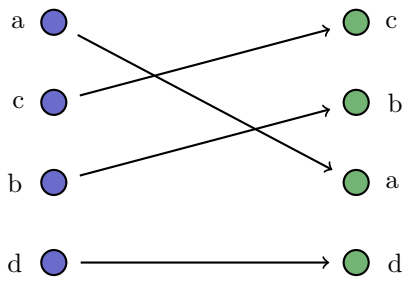
This definition is rather technical, and most of the time it is easier to think of a permutation just as a rearrangement of the numbers from 1 up to n.

Example 6 (Permutations). Looking at the first few S_k, we find the following:

$$\begin{aligned}
\mathcal{S}_1 &= \{1\} \\
\mathcal{S}_2 &= \{12, 21\} \\
\mathcal{S}_3 &= \{123, 132, 213, 231, 312, 321\} \\
&\vdots
\end{aligned}$$

It is important to note that we only use numbers to represent the structure of the permutation. In some sense a permutation is just a function which we can apply to a list of elements to rearrange them accordingly: Let $\pi = 3214$.

When Appying π to (a, c, b, d), we end up with (c, b, a, d):



The first element is moved to position $\pi(1)$, the second element to position $\pi(2)$, and so on.

The following lemma tells us how many permutations there are of a given size:

Lemma 3. *We know that*

$$|\mathcal{S}_n| = n!$$

for all $n \in \mathbb{N}$.

We will give a combinatorial proof of this fact.

Proof. When shuffling n objects we can think of picking them in a random order. There are n possibilities to pick the first object, $(n-1)$ possibilities for the second one and so on. This continues, until there remains only one choice for the last object. Since those decisions are independent, we conclude that

$$|\mathcal{S}_n| = n \cdot (n-1) \cdots 2 \cdot 1 = n!$$

for all $n \in \mathbb{N}$.

□

In the following sections we will have a look at various interesting properties of permutations and at how they are distributed over the set of all permutations of a given size.

3.1 Fixpoints

Definition 9 (Fixpoint). *Let* $n \in \mathbb{N}$. *A permutation* $\pi \in \mathcal{S}_n$ *is said to have a fixpoint at position* k *for* $1 \leq k \leq n$ *if and only if*

$$\pi(k) = k.$$

We further define $\mathcal{F}(\pi)$ *to be the number of fixpoints of* π.

The fixpoints therefore are just the positions that stay fixed, when the permutation is applied to a list of objects.

Example 7 (Fixpoints). Consider the following permutations of length 7:

$\pi \in \mathcal{S}_7$	$\mathcal{F}(\pi)$
15**3**645**7**	3
4167352	0
1234567	7

All the fixpoints are highlighted.

It is clear that the minimal number of fixpoints is 0 and the maximal number is given by the length of the permutation. Let us now start to investigate how the fixpoints are distributed by picking a random permutation.

Theorem 2 (Expected Number of Fixpoints). *Let $\pi \in \mathcal{S}_n$ be chosen at random. It holds that*

$$\mathbb{E}\Big(F(\pi)\Big) = 1$$

for all $n \in \mathbb{N}$.

This result is rather surprising, since the expected value of fixpoints does not depend on the length of the permutations at all.

Proof. Using the methods already discussed, this result is evident: Let $n \in \mathbb{N}$ and let $\pi \in \mathcal{S}_n$ be chosen uniformly at random. We now introduce indicator variables defined as

$$I_k = \begin{cases} 1, & \text{if } \pi(k) = k \\ 0, & \text{otherwise} \end{cases}$$

for $1 \leq k \leq n$. Given linearity of expectancy, we deduce that

$$\mathbb{E}\Big(F(\pi)\Big) = \mathbb{E}\left(\sum_{k=1}^{n} I_k\right) = \sum_{k=1}^{n} \mathbb{E}(I_k) = \sum_{k=1}^{n} \mathbb{P}(I_k = 1).$$

However, the probability of some index k to be a fixpoint is just given by the ratio of all permutations for which this is true, divided by the total number of permutations.

$$\mathbb{P}(I_k = 1) = \frac{(n-1)!}{n!} = \frac{1}{n}$$

for all $1 \leq k \leq n$. It follows that

$$\mathbb{E}\Big(F(\pi)\Big) = n \cdot \frac{1}{n} = 1,$$

which is exactly what we wanted to show.

□

We will now prove a connection between higher moments of fixpoints and a sequence of numbers frequently used in combinatorics. Let us first introduce those numbers.

Definition 10 (Stirling Numbers of the Second Kind). *The Stirling numbers of the second kind, denoted by $\left\{ {n \atop k} \right\}$ for $n, k \in \mathbb{N}_0$, are defined as the numbers of possible partitions of n elements into exactly k nonempty sets.*

From this it is nothing but natural to extend to the following:

Definition 11 (Bell Numbers). *Let $n \in \mathbb{N}_0$. The n-th Bell number, denoted by B_n, is defined to be the number of possible partitions of n elements into any number of nonempty sets.*

It is very hard to find an explicit formula for the Stirling numbers and therefore also for the Bell numbers. To get familiar with those new definitions, let us look at an example.

Example 8 (Bell Numbers). Let us compute B_3. By definition its value is equal to the number of ways to split up 3 elements into groups of arbitrary sizes. There are a total of 5 ways to do so:

Number of sets	1	2	3
Partitions	{a,b,c}	{a},{b,c} {c},{a,b} {b},{c,a}	{a},{b},{c}

Therefore $B_3 = 5$. The first few Bell numbers, starting with B_0, are as follows:

$$1, 1, 2, 5, 15, 52, 203, 877, 4140, \ldots$$

Intuitively we have already made use of the lemma following below.

Lemma 4. *We know that*

$$B_n = \sum_{k=0}^{n} \left\{ {n \atop k} \right\}$$

for all $n \in \mathbb{N}_0$.

Proof. This follows directly from the definition of the Bell numbers, as we sum over all possible numbers of sets.

□

The next surprising result will show how Bell numbers are connected to higher moments of fixpoints.

Theorem 3 (Moments of Fixpoints). *Let $\pi \in \mathcal{S}_n$ be chosen uniformly at random and let $m \in \mathbb{N}_0$ such that $m \leq n$. We know that*

$$\mathbb{E}\Big(F(\pi)^m\Big) = B_m$$

for all $n \in \mathbb{N}$.

Proof. Let $\pi \in \mathcal{S}_n$ be chosen uniformly at random and let $m \in \mathbb{N}$ with $m \leq n$. We discard the case $m = 0$ since it is trivial. As always we will introduce indicator variables defined as

$$I_k = \begin{cases} 1, & \text{if } \pi(k) = k \\ 0, & \text{otherwise} \end{cases}$$

for $1 \leq k \leq n$. Let $\mathcal{A}_r$ be the collection of all subsets of $\{1, 2, ..., n\}$ of size r and let $\mathcal{B}_S$ denote all the possible ordered partitions of m into variables $b_s \in \mathbb{N}$ for $s \in S$. We deduce that

$$\begin{aligned} \mathbb{E}\Big(F(\pi)^m\Big) &= \mathbb{E}\left(\left(\sum_{k=1}^{n} I_k\right)^m\right) = \mathbb{E}\left(\sum_{r=1}^{m} \sum_{S \in \mathcal{A}_r} \sum_{\mathcal{B}_S} \prod_{s \in S} I_s^{b_s}\right) \\ &= \sum_{r=1}^{m} \sum_{S \in \mathcal{A}_r} \sum_{\mathcal{B}_S} \mathbb{E}\left(\prod_{s \in S} I_s^{b_s}\right) = \sum_{r=1}^{m} \sum_{S \in \mathcal{A}_r} \sum_{\mathcal{B}_S} \mathbb{P}(I_s = 1, \forall_{s \in S}). \end{aligned}$$

The probability that π has fixpoints at all indices in a given set of size r is again given by the number of permutations where this is true, divided by the total number of permutations. We find

$$\mathbb{P}(I_s = 1, \forall_{s \in S}) = \frac{(n-r)!}{n!}$$

for every set S of size $r \leq n$. Since this number is independent of the partition $\mathcal{B}_S$, its ordering and the choice of S, we can conclude

$$\begin{aligned} \mathbb{E}\Big(F(\pi)^m\Big) &= \sum_{r=1}^{m} \sum_{S \in \mathcal{A}_r} \sum_{\mathcal{B}_S} \frac{(n-r)!}{n!} = \sum_{r=1}^{m} \sum_{S \in \mathcal{A}_r} \left\{ {m \atop r} \right\} \cdot \frac{r! \cdot (n-r)!}{n!} \\ &= \sum_{r=1}^{m} \binom{n}{r} \left\{ {m \atop r} \right\} \cdot \frac{r! \cdot (n-r)!}{n!} \\ &= \sum_{r=1}^{m} \left\{ {m \atop r} \right\} = B_m. \end{aligned}$$

This is exactly what we wanted to show.

□

A very natural question arises: What is the probability of a random permutation to have no fixpoints at all? How about the probability, that it has exactly one fixpoint? What about exactly k fixpoints?

Theorem 4. *For $n \in \mathbb{N}$, let $\pi \in \mathcal{S}_n$ be chosen at random. The probability that π contains exactly k fixpoints as n goes to infinity is given by*

$$\lim_{n\to\infty} \mathbb{P}\Big(F(\pi) = k\Big) = \frac{1}{e \cdot k!}$$

for any fixed $k \in \mathbb{N}_0$.

Again, e denotes Euler's number.

Proof. Let $n \in \mathbb{N}$ and let $\pi \in \mathcal{S}_n$ be chosen at random. We define

$$\mathcal{F}_k(n) = \{\tau \in \mathcal{S}_n \mid F(\tau) = k\}$$

and

$$A_B = \{\tau \in \mathcal{S}_n \mid \forall b \in B, \tau(b) = b\}$$

for $B \subseteq \{1, 2, 3, \ldots, n\}$. By the inclusion-exclusion principle we deduce that

$$|\mathcal{F}_0(n)| = \sum_{|B|=0} |A_B| - \sum_{|B|=1} |A_B| + \sum_{|B|=2} |A_B| - \cdots \pm \sum_{|B|=n} |A_B|$$

where $B \subseteq \{1, 2, 3, \ldots, n\}$. Since $|A_B| = (n - |B|)!$, we conclude

$$\begin{aligned} |\mathcal{F}_0(n)| &= \binom{n}{0} \cdot n! - \binom{n}{1} \cdot (n-1)! + \binom{n}{2} \cdot (n-2)! - \cdots \pm \binom{n}{n} \cdot (n-n)! \\ &= \sum_{r=0}^{n} \left(\binom{n}{r} \cdot (n-r)! \cdot (-1)^r \right) = n! \cdot \sum_{r=0}^{n} \frac{(-1)^r}{r!}. \end{aligned}$$

If we rearrange terms and let n go to infinity, it follows that

$$\lim_{n\to\infty} \mathbb{P}\Big(F(\pi) = 0\Big) = \lim_{n\to\infty} \frac{|\mathcal{F}_0(n)|}{n!} = \lim_{n\to\infty} \sum_{r=0}^{n} \frac{(-1)^r}{r!} = \frac{1}{e}.$$

The last equality is due to the well-known Taylor-Maclaurin series expansion of e^x, evaluated at $x = -1$.

The results for the general number of fixpoints is now very straightforward, as

$$|\mathcal{F}_k(n)| = \binom{n}{k} \cdot |\mathcal{F}_0(n-k)| = \frac{n!}{k!} \cdot \sum_{r=0}^{(n-k)} \frac{(-1)^r}{r!}.$$

Because k is fixed, the term $n-k$ tends to infinity with n and we arrive at the desired result:

$$\lim_{n\to\infty} \mathbb{P}\Big(F(\pi) = k\Big) = \lim_{n\to\infty} \frac{|\mathcal{F}_k(n)|}{n!} = \frac{1}{k!} \cdot \lim_{n\to\infty} \sum_{r=0}^{(n-k)} \frac{(-1)^r}{r!} = \frac{1}{e \cdot k!}$$

for all fixed $k \in \mathbb{N}_0$.

□

3.2 Patterns

Definition 12 (Pattern). *For $n, k \in \mathbb{N}$ where $k \leq n$, let $\pi \in \mathcal{S}_n$ and $\tau \in \mathcal{S}_k$. We say that the pattern τ occurs in π at indices $1 \leq i_1 < i_2 < ... < i_k \leq n$, if and only if the relative orderings of $\pi(i_1), \pi(i_2), \ldots, \pi(i_k)$ and τ are equivalent. Further let $P(\tau, \pi)$ denote the number of different occurrences of τ in π.*

In other words, a pattern is a particular substructure of a permutation. The best way to get familiar with this concept is to consider some examples.

Example 9 (Patterns). Consider the following 3 permutations of length 6:

$\pi \in \mathcal{S}_6$	$P(123, \pi)$	$P(312, \pi)$
341526	6	2
635214	0	3
142635	5	2

In case of the first permutation, the 6 occurrences of the pattern 123 are highlighted below.

$$\underline{\mathbf{34}}\underline{\mathbf{1}}\underline{\mathbf{5}}26 \quad \underline{\mathbf{34}}152\underline{\mathbf{6}} \quad \underline{\mathbf{3}}41\underline{\mathbf{52}}\underline{\mathbf{6}}$$

$$3\underline{\mathbf{4}}1\underline{\mathbf{52}}\underline{\mathbf{6}} \quad 34\underline{\mathbf{152}}\underline{\mathbf{6}} \quad 34\underline{\mathbf{1}}5\underline{\mathbf{26}}$$

These are all 3-element substructures contained in the permutation that have the same partial ordering as 123.

As with the fixpoints before, we want to investigate the distribution of patterns over all permutations by considering random permutations.

Theorem 5 (Expected Number of Pattern Occurrences). *For $n, k \in \mathbb{N}$ with $k \leq n$, let $\pi \in \mathcal{S}_n$ be chosen at random and $\tau \in \mathcal{S}_k$ be fixed. The expected number of occurrences of the pattern τ in π is given by*

$$\mathbb{E}\Big(P(\tau, \pi)\Big) = \binom{n}{k} \cdot \frac{1}{k!}.$$

Proof. Let $n, k \in \mathbb{N}$ such that $k \leq n$ and let $\tau \in \mathcal{S}_k$ be fixed. Now let $\pi \in \mathcal{S}_n$ be chosen at random. Let $\mathcal{A}$ be the collection of all subsets of $\{1, 2, 3..., n\}$ of size k. We define indicator variables

$$I_A = \begin{cases} 1, & \text{if the pattern } \tau \text{ occurs in } \pi \text{ at indices given by } A \\ 0, & \text{otherwise} \end{cases}$$

for all $A \in \mathcal{A}$. It follows that

$$\mathbb{E}\Big(P(\tau, \pi)\Big) = \mathbb{E}\left(\sum_{A \in \mathcal{A}} I_A\right) = \sum_{A \in \mathcal{A}} \mathbb{E}(I_A) = \sum_{A \in \mathcal{A}} \mathbb{P}(I_A = 1).$$

But we know that the probability of τ occurring in π at indices given by any $A \in \mathcal{A}$ is given by the number of permutations where this is true, divided by the total number of permutations. It follows that

$$\mathbb{P}(I_A = 1) = \binom{n}{k} \cdot \frac{(n-k)!}{n!} = \frac{1}{k!}$$

for all $A \in \mathcal{A}$. We therefore deduce that

$$\mathbb{E}\Big(P(\tau, \pi)\Big) = \binom{n}{k} \cdot \frac{1}{k!}$$

for any choice of n, k and τ.

□

Looking at higher moments of patterns turns out to be rather hard. Our previous method using indicator variables begins to break down here as the individual events are not independent anymore. Multiple patterns can overlap at some indices, resulting in a dependency of them occurring. Nonetheless, an asymptotic result can still be found.

Theorem 6 (Asymptotic Moments of Pattern Occurrences). *For $n, k \in \mathbb{N}$, let $\pi \in \mathcal{S}_n$ be chosen uniformly at random and let $\tau \in \mathcal{S}_k$ be fixed. The m-th moment of $P(\tau, \pi)$ as n goes to infinity is asymptotically given by*

$$\mathbb{E}\Big(P(\tau,\pi)^m\Big) \sim \binom{n}{k}^m \cdot \left(\frac{1}{k!}\right)^m$$

for all fixed $m \in \mathbb{N}_0$.

Note that this implies that the m-th moment of pattern occurrences is asymptotically equal to the m-th power of the expected number of occurrences.

Proof. For $n, k \in \mathbb{N}$, let $\pi \in \mathcal{S}_n$ be chosen at random and $\tau \in \mathcal{S}_k$ be fixed. Again we define $\mathcal{A}$ to be the collection of all subsets of $\{1, 2, 3..., n\}$ of size k and

$$I_A = \begin{cases} 1, & \text{if the pattern } \tau \text{ occurs in } \pi \text{ at indices given by } A \\ 0, & \text{otherwise} \end{cases}$$

for all $A \in \mathcal{A}$. It follows that

$$\mathbb{E}\Big(P(\tau,\pi)^m\Big) = \mathbb{E}\left(\left(\sum_{A\in\mathcal{A}} I_A\right)^m\right) = \mathbb{E}\left(\sum_{B\in\mathcal{B}}\prod_{A\in B} I_A\right),$$

where $\mathcal{B}$ denotes the collection of all ordered selections from the set $\mathcal{A}$ with size m and possibly repetition. Using linearity of expectation we deduce that

$$\mathbb{E}\Big(P(\tau,\pi)^m\Big) = \sum_{B\in\mathcal{B}} \mathbb{E}\left(\prod_{A\in B} I_A\right) = \sum_{B\in\mathcal{B}} \mathbb{P}(I_A = 1\ \forall_{A\in B}).$$

Here is where we need to take a different approach than up to now. We can not simply expand into a product of probabilities, because they are dependent. Luckily, the following argument enables us to still make a conclusion in the asymptotic case: Since k is fixed and we are letting n go to infinity, the proportion of elements $B \in \mathcal{B}$ where all I_A for $A \in B$ are independent tends towards 1. When taking the limit, we can discard the cases where there is some dependency, as it shrinks to 0 as n grows.

We can therefore just assume that all I_A are independent. It follows that

$$\mathbb{E}\Big(P(\tau,\pi)^m\Big) \sim \sum_{B\in\mathcal{B}} \prod_{A\in B} \mathbb{P}(I_A = 1).$$

But as we have seen before,

$$\mathbb{P}(I_A = 1) = \frac{1}{k!}$$

for all $A \in \mathcal{A}$. Together with the fact that

$$|\mathcal{B}| = \binom{n}{k}^m \text{ and } |B| = m,$$

we conclude

$$\mathbb{E}\Big(P(\tau,\pi)^m\Big) \sim \binom{n}{k}^m \cdot \left(\frac{1}{k!}\right)^m$$

as n goes to infinity.

□

While long patterns only come up occasionally in applications and are quite hard to count, smaller ones are often more useful. The simplest example for this will be our next topic of concern.

Definition 13 (Inversions). *Let $\pi \in \mathcal{S}_n$. We say that π contains an inversion at indices $1 \le a < b \le n$ if and only if $\pi(a) > \pi(b)$. We denote by $\mathcal{I}(\pi)$ the number of inversions contained in π.*

In other words, inversions are just occurrences of the 21 pattern.

Example 10 (Inversions). Consider the number of inversions of the following permutations of length 5:

$\pi \in \mathcal{S}_5$	$\mathcal{I}(\pi)$
35214	6
12345	0
54321	10

The inversions of the first permutation are again highlighted below.

$$\underline{\mathbf{3}}5\underline{\mathbf{2}}14 \quad \underline{\mathbf{3}}52\underline{\mathbf{1}}4 \quad 3\underline{\mathbf{52}}14$$
$$3\underline{\mathbf{5}}2\underline{\mathbf{1}}4 \quad 3\underline{\mathbf{5}}21\underline{\mathbf{4}} \quad 35\underline{\mathbf{21}}4$$

Another way of thinking about inversions is to think about all the pairs whose order is inverted after we apply the permutation to some list of objects.

It is easy to see that the number of inversions takes its minimal and its maximal value if and only if the permutation is ordered decreasingly or increasingly respectively. Let us now look at the expected number of such inversions.

Theorem 7 (Expected Number of Inversions). *Let $\pi \in \mathcal{S}_n$ be chosen at random. It holds that*

$$\mathbb{E}\Big(\mathcal{I}(\pi)\Big) = \frac{1}{2} \cdot \binom{n}{2}$$

for all $n \in \mathbb{N}$.

This result follows immediately from the more general result about the number of pattern occurrences discussed earlier in theorem 5. Alternatively, consider the following argument:

Proof. We pair up all the elements of S_n, each permutation with its corresponding reversed permutation. The combined number of inversions in each pair is now exactly $\binom{n}{2}$. When choosing a random permutation, we are equally likely to pick either one of every pair. The result follows from this.

□

Using our asymptotic result about moments of pattern occurrences we find that

$$\mathbb{E}\Big(\mathcal{I}(\pi)^2\Big) \sim \mathbb{E}\Big(\mathcal{I}(\pi)\Big)^2 = \frac{n^2 \cdot (n-1)^2}{16} = \frac{1}{16} \cdot n^4 - \frac{1}{8} \cdot n^3 + \frac{1}{16} \cdot n^2.$$

Because inversions are very short patterns, we can also find their second moment explicitly.

Theorem 8 (Second Moment of Inversions). *Let $\pi \in \mathcal{S}_n$ be chosen at random. It holds that*

$$\mathbb{E}\Big(\mathcal{I}(\pi)^2\Big) = \frac{1}{16} \cdot n^4 - \frac{7}{72} \cdot n^3 + \frac{5}{48} \cdot n^2 - \frac{5}{72} \cdot n$$

for all $n \in \mathbb{N}$.

To prove this theorem, we have to do some more work. This is because the events of two inversions can be dependent.

Proof. For $n \in \mathbb{N}$ with $n \geq 4$, let $\pi \in \mathcal{S}_n$ be chosen at random. We will make use of indicator variables defined as

$$I_p = \begin{cases} 1, & \text{if } \pi(a) > \pi(b) \\ 0, & \text{otherwise} \end{cases}$$

for all pairs of indices $p = (a, b)$ with $1 \leq a < b \leq n$. We can now write

$$\begin{aligned} \mathbb{E}\Big(\mathcal{I}(\pi)^2\Big) &= \mathbb{E}\left(\left(\sum_p I_p\right)^2\right) = \sum_p \mathbb{E}\big(I_p^2\big) + 2 \cdot \sum_{p \neq q} \mathbb{E}(I_p \cdot I_q) \\ &= \sum_p \mathbb{P}(I_p = 1) + \sum_{p \neq q} \mathbb{P}(I_p = 1 \wedge I_q = 1). \end{aligned}$$

The first sum is nothing else than the expected value of $\mathcal{I}(\pi)$. Again, since we can not assume $\mathcal{I}_p$ and $\mathcal{I}_q$ to be independent, we can not just expand the probability into a product. Instead, we have to split the sum into different cases which we can evaluate separately. To do so, let $p = (a, b)$ and $q = (c, d)$ where $1 \leq a < b \leq n$ and $1 \leq c < d \leq n$ with $p \neq q$. Looking at all possible arrangements in each case, we find that

$$\mathbb{P}(I_p = 1 \wedge I_q = 1) = \begin{cases} \frac{1}{4}, & \text{if } p \cap q = \emptyset. \\ \frac{1}{6}, & \text{if } b = c \text{ or } a = d. \\ \frac{1}{3}, & \text{otherwise.} \end{cases}$$

Counting how many pairs p and q there are in each case and plugging everything back into the formula, we find that

$$\begin{aligned} \mathbb{E}\Big(\mathcal{I}(\pi)^2\Big) &= \binom{n}{2} \cdot \frac{1}{2} + \binom{n}{2} \cdot \binom{n-2}{2} \cdot \frac{1}{4} + 2 \cdot \binom{n}{3} \cdot \frac{1}{6} + 4 \cdot \binom{n}{3} \cdot \frac{1}{3} \\ &= \frac{1}{16} \cdot n^4 - \frac{7}{72} \cdot n^3 + \frac{5}{48} \cdot n^2 - \frac{5}{72} \cdot n. \end{aligned}$$

Also checking the the cases where $n = 1, 2$ and 3, we conclude that the result holds for every $n \in \mathbb{N}$.

□

Comparing this result to the asymptotic behaviour, we see that the degrees as well as the leading coefficients are indeed equal, just as expected.

3.3 Intervals

The notion of intervals is closely related to notion of patterns.

Definition 14 (Intervals). *Let $n, k \in \mathbb{N}$ with $2 \leq k \leq n-1$ and let $\pi \in \mathcal{S}_n$. We say π contains an interval of length k starting at position a for $1 \leq a \leq n+1-k$ if and only if*

$$\pi(a)\pi(a+1)\cdots\pi(a+k-1)$$

is a permutation of k consecutive numbers. We further define $S(\pi, k)$ to be the number of intervals of length k in π and $S(\pi)$ to be the total number of intervals of any allowed length.

An interval therefore is just a consecutive part of a permutation whose values also cover a consecutive part of the numbers from 1 up to n.

Example 11 (Intervals). Consider the following permutations of length 6:

$\pi \in \mathcal{S}_6$	$S(\pi)$
453612	4
352461	2
351624	0

In the case of the first permutation, the 4 occurring intervals are shown below.

$$\underline{\mathbf{45}}3612 \quad \underline{\mathbf{453}}612$$
$$\underline{\mathbf{4536}}12 \quad 4536\underline{\mathbf{12}}$$

The intervals of a permutation can also be visualised as consecutive segments of an array that stay consecutive whenever the permutation is applied.

Note that segments of length 1 and the entire permutation always satisfy the requirements for intervals. The definition therefore excludes those cases.

Theorem 9 (Expected Number of Intervals on Given Length). *For $n \in \mathbb{N}$, let $\pi \in \mathcal{S}_n$ be chosen at random. It holds that*

$$\mathbb{E}\Big(S(\pi, k)\Big) = \frac{(n-k+1)^2}{\binom{n}{k}}$$

for all $k \in \mathbb{N}$ with $2 \leq k \leq n-1$.

The proof is similar to what we have seen many times before.

Proof. Let $n, k \in \mathbb{N}$ with $2 \leq k \in \mathbb{N} \leq n-1$ and let $\pi \in \mathcal{S}_n$ be chosen at random. We once again define indicator variables as

$$I_a = \begin{cases} 1, & \text{if } \pi \text{ contains an interval of length } k \text{ starting at position } a \\ 0, & \text{otherwise} \end{cases}$$

for $1 \leq a \leq n-k+1$. With linearity of expectancy it follows that

$$\mathbb{E}\Big(S(\pi, k)\Big) = \mathbb{E}\left(\sum_{a=1}^{n-k+1} I_a\right) = \sum_{a=1}^{n-k+1} \mathbb{E}(I_a) = \sum_{a=1}^{n-k+1} \mathbb{P}(I_a = 1).$$

It is not hard to see that

$$\mathbb{P}(I_a = 1) = \frac{(n-k+1) \cdot k! \cdot (n-k)!}{n!} = \frac{n-k+1}{\binom{n}{k}}$$

for all $1 \leq a \leq n-k+1$. It follows from here that

$$\mathbb{E}\Big(S(\pi, k)\Big) = \frac{(n-k+1)^2}{\binom{n}{k}}$$

for any $2 \leq k \leq n-1$.

□

In particular, this theorem implies that the expected number of intervals of length 2 in a random permutation of size n is given by

$$\frac{(n-1)^2}{\binom{n}{2}} = \frac{2 \cdot (n-1)^2}{n \cdot (n-1)} = \frac{2 \cdot (n-1)}{n} = 2 - \frac{2}{n}.$$

Note that as n tends to infinity, this value tends towards 2. For intervals of length greater than 2, it is easy to see that the denominator of the fraction grows faster than the nominator and the value tends to 0 as n grows. This observation motivates the next result.

Theorem 10 (Expected Number of Intervals). *For $n \in \mathbb{N}$, let $\pi \in \mathcal{S}_n$ be chosen at random. It holds that*

$$\lim_{n\to\infty} \mathbb{E}\Big(\mathcal{S}(\pi)\Big) = 2.$$

Note that we can not take the limit first and then sum over all possible lengths of intervals. This is because the number of possible lengths also goes to infinity if we let the size of the permutation grow.

Proof. Let $\pi \in \mathcal{S}_n$ be chosen at random where $n \in \mathbb{N}$. By the previous result and linearity of expectation we know that

$$\mathbb{E}\Big(\mathcal{S}(\pi)\Big) = \mathbb{E}\left(\sum_{k=2}^{n-1} S(\pi,k)\right) = \sum_{k=2}^{n-1} \mathbb{E}\Big(S(\pi,k)\Big) = \sum_{k=2}^{n-1} \frac{(n-k+1)^2}{\binom{n}{k}}.$$

As we have already seen that for $k = 2$ the fraction tends towards 2 as n tends to infinity, we only have to show that the remaining terms vanish as n grows. Taking the first and last term out of the sum, we find that

$$\sum_{k=3}^{n-1} \frac{(n-k+1)^2}{\binom{n}{k}} = \frac{(n-2)^2}{\binom{n}{3}} + \frac{4}{\binom{n}{n-1}} + \sum_{k=4}^{n-3} \frac{(n-k+1)^2}{\binom{n}{k}}.$$

Since all the $n-5$ remaining terms in the summation are of order n^{-2} or smaller, the sum vanishes as n grows. We therefore conclude that

$$\lim_{n\to\infty} \mathbb{E}\Big(\mathcal{S}(\pi)\Big) = \lim_{n\to\infty} 2 + \frac{6\cdot(n-2)}{n\cdot(n-1)} + \frac{4}{n} = 2.$$

We have reached the desired result.

□

Theorem 11 (Second Moment of Intervals of Length 3). *Let $\pi \in \mathcal{S}_n$ be chosen at random. We know that*

$$\mathbb{E}\Big(\mathcal{S}(\pi,3)^2\Big) = \frac{6\cdot n^3 - 16\cdot n^2 - 98\cdot n + 296}{n\cdot(n-1)\cdot(n-2)\cdot(n-3)}$$

for all $n \in \mathbb{N}$ with $n \geq 6$.

The approach we will take in the proof of this result is similar to the one we used to prove theorem 8 about the second moments of inversions.

Proof. For $n \in \mathbb{N}$ with $n \geq 6$, let $\pi \in \mathcal{S}_n$ be chosen at random. We will make use of indicator variables:

$$I_a = \begin{cases} 1, & \text{if } \pi \text{ contains an interval of length 3 starting at position } a \\ 0, & \text{otherwise} \end{cases}$$

for $1 \leq a \leq n-2$. As seen many times before, we write

$$\mathbb{E}\Big(\mathcal{S}(\pi,3)^2\Big) = \mathbb{E}\left(\sum_{a=1}^{n-2} I_a^2 + 2 \cdot \sum_{a<b} I_a \cdot I_b\right),$$

where the second sum is over all pairs of indices $1 \leq a < b \leq n-2$. Using linearity of expectation we arrive at

$$\mathbb{E}\Big(\mathcal{S}(\pi,3)^2\Big) = \sum_{a=1}^{n-2} \mathbb{P}(I_a) + 2 \cdot \sum_{a<b} \mathbb{P}(I_a = 1 \wedge I_b = 1).$$

The first sum is just the expectation of $\mathcal{S}(\pi,3)$ and we easily find

$$\mathbb{P}(I_a = 1) = \frac{n-2}{\binom{n}{3}}.$$

To evaluate the second sum, we again need to split it accordingly to the size of the intersection of the two segments. We deduce that

$$\begin{aligned} \sum_{a<b} \mathbb{P}(I_a = 1 \wedge I_b = 1) = &\sum_{a+2<b} \mathbb{P}(I_a = 1 \wedge I_b = 1) \\ &+ \sum_{a=1}^{n-3} \mathbb{P}(I_a = 1 \wedge I_{a+1} = 1) + \sum_{a=1}^{n-4} \mathbb{P}(I_a = 1 \wedge I_{a+2} = 1) \\ = &\frac{2 \cdot \binom{n-4}{2}^2}{\binom{n}{3} \cdot \binom{n-3}{3}} + \frac{(n-3)^2}{6 \cdot \binom{n}{4}} + \frac{(n-4)^2}{15 \cdot \binom{n}{5}}. \end{aligned}$$

Putting everything together, we conclude

$$\begin{aligned} \mathbb{E}\Big(\mathcal{S}(\pi,3)^2\Big) &= \frac{(n-2)^2}{\binom{n}{3}} + \frac{4 \cdot \binom{n-4}{2}^2}{\binom{n}{3} \cdot \binom{n-3}{3}} + \frac{(n-3)^2}{3 \cdot \binom{n}{4}} + \frac{2 \cdot (n-4)^2}{15 \cdot \binom{n}{5}} \\ &= \frac{6 \cdot n^3 + 2 \cdot n^2 - 260 \cdot n + 656}{n \cdot (n-1) \cdot (n-2) \cdot (n-3)}. \end{aligned}$$

This is what we wanted to show.

□

4 Using Higher Moments

It is now time to talk about theorems which allow us to use the properties of probability distributions found so far. Our main goal is to deduce useful bounds on certain probabilities using higher moments of distributions.

4.1 Using Markov's Inequality

Theorem 12 (Markov's Inequality). *Let $X \in \mathbb{R}^+$ be a random variable. We know that*

$$\mathbb{P}(X \geq a) \leq \frac{\mathbb{E}(X)}{a}$$

for every $a \in \mathbb{R}^+$.

By substituting $a = b \cdot \mathbb{E}(X)$, we get the equivalent inequality stating that

$$\mathbb{P}\Big(X \geq b \cdot \mathbb{E}(X)\Big) \leq \frac{1}{b}$$

for all $b \in \mathbb{R}^+$. Writing the statement this way, it is easier to see what this inequality is actually telling us: For large values of b, the probability, that the random variable X is far away from its expected value, is very small. Also note that this theorem is only useful for values $a \leq \mathbb{E}(X)$ or $b \geq 1$ respectively, as otherwise it just tells us that some probability is smaller or equal to 1, which is already true by definition.

Proof. Let $X \in \mathbb{R}^+$ be a random variable and let $a \in \mathbb{R}^+$. We deduce that

$$\begin{aligned}
\mathbb{E}(X) &= \sum_x \Big(\mathbb{P}(X = x) \cdot X\Big) \\
&= \sum_{x \geq a} \Big(\mathbb{P}(X = x) \cdot X\Big) + \sum_{x < a} \Big(\mathbb{P}(X = x) \cdot X\Big) \\
&\geq \sum_{x \geq a} \Big(\mathbb{P}(X = x) \cdot X\Big) \\
&\geq \mathbb{P}(X \geq a) \cdot a.
\end{aligned}$$

Dividing both sides by a yields the desired result.

□

Note that in the proof given, it is crucial that all values are positive. We can extend this theorem to also take into account higher moments.

Theorem 13 (Extended Markov's Inequality). *Let $X \in \mathbb{R}^+$ be a random variable and let $m \in \mathbb{N}_0$. We know that*

$$\mathbb{P}(X \geq a) \leq \frac{\mathbb{E}(X^m)}{a^m}$$

for every $a \in \mathbb{R}^+$.

The proof is almost identical to the one given above.

Proof. Let $X \in \mathbb{R}^+$ be a random variable , $m \in \mathbb{N}_0$ and $a \in \mathbb{R}^+$. We deduce

$$\begin{aligned}
\mathbb{E}\big(X^m\big) &= \sum_x \Big(\mathbb{P}(X = x) \cdot X^m\Big) \\
&= \sum_{x \geq a} \Big(\mathbb{P}(X = x) \cdot X^m\Big) + \sum_{x < a} \Big(\mathbb{P}(X = x) \cdot X^m\Big) \\
&\geq \sum_{x \geq a} \Big(\mathbb{P}(X = x) \cdot X^m\Big) \\
&\geq \mathbb{P}(X \geq a) \cdot a^m.
\end{aligned}$$

Dividing both sides by a^m yields the result.

□

4.2 Using Chebyshev's Inequality

While Markov's inequality is only useful in very particular cases, the next result is well-known for its broad application.

Theorem 14 (Chebyshev's Inequality). *Let $X \in \mathbb{R}^+$ be a random variable. It holds that*

$$\mathbb{P}\Big(|X - \mathbb{E}(X)| \geq k\Big) \leq \frac{\mathrm{Var}(X)}{k^2}$$

for every $k \in \mathbb{R}^+$.

This result quantifies the connection between the variance and how "spread-out" the distribution is.

Proof. Let $X \in \mathbb{R}^+$ be a random variable and let $k \in \mathbb{R}+$. Using

$$(X - \mathbb{E}(X))^2$$

as our random variable, we will make use of Markov's inequality. We find that

$$\mathbb{P}\Big((X - \mathbb{E}(X))^2 \geq a\Big) \leq \frac{\mathbb{E}\Big((X - \mathbb{E}(X))^2\Big)}{a}.$$

By substituting $a = k^2$ and noting that the numerator on the right hand side is nothing else than the variance of X by definition, we obtain

$$\mathbb{P}\Big((X - \mathbb{E}(X))^2 \geq k^2\Big) \leq \frac{\mathrm{Var}(X)}{k^2}.$$

Because $k > 0$ and $f(x) = x^2$ is an increasing function, we can take the square root of the probability condition and deduce that indeed

$$\mathbb{P}\Big(|X - \mathbb{E}(X)| \geq k\Big) \leq \frac{\mathrm{Var}(X)}{k^2},$$

which is exactly the desired result.

□

To illustrate the power of this theorem, let us use our previous work to deduce some interesting results.

Application 5 (Bound on Fixpoints). Given a random permutation $\pi \in \mathcal{S}_n$ where $n \in \mathbb{N}$ such that $n \geq 2$, we have already shown that the expected number of fixpoints of π is 1. Using Chebyshev's inequality, we are now able to find an upper bound on the probability that π has many more fixpoints than the single one expected. We know that

$$\mathbb{E}\Big(\mathcal{F}(\pi)\Big) = 1,$$

and by definition we find

$$V(X) = \mathbb{E}\Big(\mathcal{F}(\pi)^2\Big) - \mathbb{E}\Big(\mathcal{F}(\pi)\Big)^2 = 2 - 1 = 1.$$

Chebyshev's inequality now tells us that

$$\mathbb{P}\Big(|\mathcal{F}(\pi) - 1| \geq k\Big) \leq \frac{1}{k^2}$$

for any $k \in \mathbb{R}^+$.

Note that for $k \geq 2$ we can get rid of the absolute values, and by additionally substituting $a = k + 1$, we conclude

$$\mathbb{P}\Big(\mathcal{F}(\pi) \geq a\Big) \leq \frac{1}{(a-1)^2}$$

for $a \geq 3$. By plugging in $a = 11$ for instance, we show that

$$\mathbb{P}\Big(\mathcal{F}(\pi) \geq 11\Big) \leq \frac{1}{100},$$

deducing that at least 99 percent of all permutations of any given length have less than 11 fixed points.

Chebyshev's inequality is especially powerful when the variance of the random variable is small. This is the case when the second moment is approximately equal to the square of the expected value. Our asymptotic result concerning the occurrences of patterns in permutations states exactly that.

Application 6 (Bound on Inversions). Given a random permutation $\pi \in \mathcal{S}_n$ for $n \in \mathbb{N}$, we have seen that

$$\mathbb{E}\Big(\mathcal{I}(\pi)\Big) = \frac{1}{2} \cdot \binom{n}{2},$$

where $\mathcal{I}(\pi)$ is again the number of inversions in π. Using Chebyshev's inequality we deduce that

$$\mathbb{P}\left(\left|\mathcal{I}(\pi) - \frac{1}{2} \cdot \binom{n}{2}\right| \geq k\right) \leq \frac{\mathrm{Var}\Big(\mathcal{I}(\pi)\Big)}{k^2}$$

for any $k \in \mathbb{R}^+$. We also remember the following:

$$\mathbb{E}\Big(\mathcal{I}(\pi)^2\Big) = \frac{1}{16} \cdot n^4 - \frac{7}{72} \cdot n^3 + \frac{5}{48} \cdot n^2 - \frac{5}{72} \cdot n$$

and

$$\mathbb{E}\Big(\mathcal{I}(\pi)\Big)^2 = \frac{1}{16} \cdot n^4 - \frac{1}{8} \cdot n^3 + \frac{1}{16} \cdot n^2.$$

The variance of $\mathcal{I}(\pi)$ can now be calculated to be

$$\mathbb{E}\Big(\mathcal{I}(\pi)^2\Big) - \mathbb{E}\Big(\mathcal{I}(\pi)\Big)^2 = \frac{1}{36} \cdot n^3 + \frac{1}{24} \cdot n^2 - \frac{5}{72} \cdot n.$$

If, for example, we let $k = \sqrt{3 \cdot n^3}$, we find that

$$\mathbb{P}\left(\left| \mathcal{I}(\pi) - \frac{1}{2} \cdot \binom{n}{2} \right| \geq \sqrt{3 \cdot n^3} \right) \leq \frac{\mathrm{Var}\Big(\mathcal{I}(\pi)\Big)}{3 \cdot n^3} = \frac{1}{108}$$

in the limit as n goes to ∞. This means that for big n, less then 1 percent of all permutations $\pi \in \mathcal{S}_n$ are farther than $\sqrt{3 \cdot n^3}$ inversions away from the expected value.

Chebyshev's inequality can also be extended to higher moments.

Theorem 15 (Extended Chebyshev's Inequality). *Let $X \in \mathbb{R}^+$ be a random variable and let $m \in \mathbb{N}_0$. We know that*

$$\mathbb{P}\Big(|X - \mathbb{E}(X)| \geq k\Big) \leq \frac{\mathbb{E}\Big(\big|\big(X - E(x)\big)^m\big|\Big)}{k^m}$$

for every $k \in \mathbb{R}^+$.

Note that here the numerator on the right hand side is the extension of the variance to higher powers.

Proof. Let $X \in \mathbb{R}^+$ be a random variable and let $m \in \mathbb{N}_0$. The proof is similar to the one given for the case $m = 2$. Markov's inequality with the random variable

$$\big|\big(X - E(x)\big)^m\big|$$

tells us that for all $a \in \mathbb{R}^+$ we have

$$\mathbb{P}\Big(|(X - \mathbb{E}(X))^m| \geq a\Big) \leq \frac{\mathbb{E}\Big(\big|\big(X - E(x)\big)^m\big|\Big)}{a}.$$

Again setting $a = k^m$ and taking the m-th root, we find that

$$\mathbb{P}\Big(|X - \mathbb{E}(X)| \geq k\Big) \leq \frac{\mathbb{E}\Big(\big|\big(X - E(x)\big)^m\big|\Big)}{k^m},$$

just as desired.

□

We can use this more general theorem to extend one of the previous results.

Application 7 (Stronger Bound on Fixpoints). Given a random permutation $\pi \in \mathcal{S}_n$ for $n \in \mathbb{N}$, let $\mathcal{F}(\pi)$ as usual be the number of fixpoints of π. The expected number of fixpoints is again 1. Chebyshev's extended theorem now tells us that for any $m \in \mathbb{N}$ with $2m \leq n$ we have

$$\mathbb{P}\Big(|\mathcal{F}(\pi) - 1| \geq k\Big) \leq \frac{\mathbb{E}\Big(\Big|\big(\mathcal{F}(\pi) - 1\big)^{2m}\Big|\Big)}{k^{2m}}.$$

We chose the exponent to be even and $k \geq 2$ in order to get rid of the absolute value bars. Expanding the binomial and our results about higher moments of fixpoints we deduce that

$$\mathbb{P}\Big(\mathcal{F}(\pi) - 1 \geq k\Big) \leq \frac{\mathbb{E}\Big(\big(\mathcal{F}(\pi) - 1\big)^{2m}\Big)}{k^{2m}} = \frac{1}{k^{2m}} \cdot \sum_{t=0}^{2m} \binom{2m}{t} \cdot B_t.$$

Similarly, we now substitute $a = k + 1$ and find that

$$\mathbb{P}\Big(\mathcal{F}(\pi) \geq a\Big) \leq \frac{1}{(a-1)^{2m}} \cdot \sum_{t=0}^{2m} \binom{2m}{t} \cdot B_t$$

for all $a \geq 3$. Using for instance $m = 3$ with this new formula, we find that

$$\mathbb{P}\Big(\mathcal{F}(\pi) \geq a\Big) \leq \frac{1 + 6 \cdot 1 + 15 \cdot 2 + 20 \cdot 5 + 15 \cdot 15 + 6 \cdot 52 + 203}{(a-1)^6} = \frac{877}{(a-1)^6}$$

for permutations of length greater or equal to 6. Plugging in $a = 8$ for instance, we find that

$$\mathbb{P}(\mathcal{F}(\pi) \geq 8) \leq \frac{877}{7^6} \approx 0.0075,$$

and we conclude that more than 99 percent of all permutations of some fixed length greater or equal to 6 have less than 8 fixpoints. Since all permutations of length less than 6 clearly also have less than 8 fixpoints, we can extend the conclusion to permutations of all lengths.

5 Probabilistic Proofs

In this chapter, we will see 4 non-probabilistic theorems which have elegant probabilistic proofs that illustrate the wide range of application of the concepts and strategies discussed earlier on.

5.1 Independent Sets

This first result comes from graph theory. A graph is nothing else than a collection of so-called nodes, where every pair of nodes is either connected or not. It is often of interest to find a large set of nodes inside a graph, where no pair is connected. This is what we call an independent set. For every node v, we call the number of nodes, it is connected to, its degree. We will denote the degree of v by $\deg(v)$.

Theorem 16 (Caro-Wei). *Given a graph where V is the set of nodes, we can find a subset of V with size at least*

$$\sum_{v \in V} \frac{1}{\deg(v)+1},$$

where no two nodes are connected.

In the first chapter we have already proven a similar result using alteration. For this stronger version we need a different approach. Consider the following greedy algorithm: We pick a random node from the graph and put it in our set. Then we remove the node itself and all nodes that are connected to it from the graph. We repeat this process until there are no nodes left. It is easy to see that no two nodes in the set we built are connected. We will make use of a random permutation to decide, in which order to pick the nodes, and then look at the expected size of our set.

Proof. For $n \in \mathbb{N}$, let G be a graph on n nodes and let $\pi \in \mathcal{S}_n$ be chosen at random. We number the nodes from 1 to n and perform the algorithm mentioned above by picking the nodes in the order given by π and skipping nodes we have already deleted. Let us define

$$I_k = \begin{cases} 1, & \text{if the algorithm picks the } k\text{-th node} \\ 0, & \text{otherwise} \end{cases}$$

for all $1 \leq k \leq n$. Let X denote the number of nodes that end up in our set.

Using linearity of expectation we write

$$\mathbb{E}(X) = \mathbb{E}\left(\sum_{k=1}^{n} I_k\right) = \sum_{k=1}^{n} \mathbb{E}(I_k) = \sum_{k=1}^{n} \mathbb{P}(I_k = 1).$$

In order for the k-th node v_k to be picked by the algorithm, all nodes connected to it have to be located further back in the permutation. The only restriction on the ordering of the $\deg(v_k) + 1$ nodes in question is that v_k appears at the start. The probability of that happening is given by

$$\mathbb{P}(I_k = 1) = \frac{1}{\deg(v_k) + 1}.$$

Summing this probability over all nodes, we find that indeed

$$\mathbb{E}(X) = \sum_{v \in V} \frac{1}{\deg(v) + 1},$$

and we conclude that there must be a choice of π such that

$$X \geq \sum_{v \in V} \frac{1}{\deg(v) + 1}.$$

This independent set has the desired size.

□

From this theorem we can deduce another result that only uses the average degree of the graph. The following theorem will be necessary for the proof.

Theorem 17 (Jensen's Inequality). *Given a convex function defined on some real interval A and any sequence of n values $x_1, x_2, \ldots, x_n \in A$ for $n \in \mathbb{N}$, we know that*

$$f(\lambda_1 x_1 + \lambda_2 x_2 + \cdots + \lambda_n x_n) \leq \lambda_1 f(x_1) + \lambda_2 f(x_2) + \cdots + \lambda_n f(x_n),$$

where the λ_k are positive real numbers satisfying $\lambda_1 + \lambda_2 + \cdots + \lambda_n = 1$.

A function is convex if and only if its second derivative is greater or equal to 0 at every point. What Jensen's inequality tells us is that, given a convex function evaluated after taking some weighted average, the value we get can not be greater than the value we get when taking the weighted average after evaluating the function.

Proof. Let f be any convex function defined on the real interval A. We will prove the theorem by induction on n. The case $n = 1$ is trivial, and for $n = 2$ the statement reduces to

$$f(t \cdot x_1 + (1-t) \cdot x_2) \leq t \cdot f(x_1) + (t-1) \cdot f(x_2).$$

This is equivalent to the fact that, considering the graph of the function, the connecting line any two points lies above the rest of the graph. This is a direct consequence of convexity. The case $n = 2$ handled, assume now we know the theorem holds for all $n < m$ where $m \geq 3$ and consider the case $n = m$. We let

$$x_1, x_2, \ldots, x_m \in A \text{ and } \lambda_1, \lambda_2, \ldots, \lambda_m \in \mathbb{R}^+ \text{ with } \lambda_1 + \lambda_2 + \cdots + \lambda_m = 1.$$

Using the assumption for $n = m - 1$, we find that

$$\begin{aligned} &\lambda_1 f(x_1) + \lambda_2 f(x_2) + \cdots + \lambda_{m-1} f(x_{m-1}) + \lambda_m f(x_m) \\ &\geq \Lambda \cdot f\left(\frac{\lambda_1 x_1 + \lambda_2 x_2 + \cdots + \lambda_{m-1} x_{m-1}}{\Lambda}\right) + \lambda_m f(x_m), \end{aligned}$$

where $\Lambda = \lambda_1 + \lambda_2 + \cdots + \lambda_{m-1}$. Since $\Lambda + \lambda_m = 1$ we can now apply the theorem for $n = 2$ and conclude that

$$\begin{aligned} &\lambda_1 f(x_1) + \lambda_2 f(x_2) + \cdots + \lambda_{m-1} f(x_{m-1}) + \lambda_m f(x_m) \\ &\geq \Lambda \cdot f\left(\frac{\lambda_1 x_1 + \lambda_2 x_2 + \cdots + \lambda_{m-1} x_{m-1}}{\Lambda}\right) + \lambda_m f(x_m) \\ &\geq f(\lambda_1 x_1 + \lambda_2 x_2 + \cdots + \lambda_{m-1} x_{m-1} + \lambda_m x_m), \end{aligned}$$

just as desired.

□

Equipped with Jensen's inequality we are now ready for a slightly weaker but more convenient corollary of the Caro-Wei theorem.

Theorem 18 (Turán). *Let $n \in \mathbb{N}$. Given a graph on n vertices with average degree d, we can find a set of nodes with size at least*

$$\frac{n}{d+1},$$

such that no two nodes are connected.

Note that this fraction is a big improvement over $n/2d$, the result we found in the first chapter.

Proof. For $n \in \mathbb{N}$, let G be a graph on n nodes with average degree d. Note that the function

$$f(x) = \frac{1}{x+1}$$

is convex in the interval $(0, \infty)$. The Caro-Wei theorem tells us that there is some set of nodes in G of size X such that no two are connected and

$$X \geq \sum_{v \in V} \frac{1}{\deg(v)+1} = \sum_{v \in V} f\Big(\deg(v)\Big),$$

where V is again the set of vertices of G. Applying Jensen's inequality with all weights being equal to $1/n$, we find that

$$\frac{X}{n} \geq \frac{1}{n} \sum_{v \in V} f\Big(\deg(v)\Big) \geq f\left(\sum_{v \in V} \frac{\deg(v)}{n}\right) = f(d) = \frac{1}{d+1}.$$

Multiplying both sides by n, we get the desired result.

□

5.2 Sum-Free Sets

Let us now have a look at a beautiful result from number theory. We introduce the following term, describing a set of numbers:

Definition 15 (Sum-Free Set). *A set A is called sum-free if and only if there do not exist elements $x, y, z \in \mathcal{A}$ such that $x + y = z$.*

Note that in the definition, the 3 numbers do not necessarily have to be distinct.

Example 12 (Sum-Free Sets). The following sets are sum-free:

$$\{1, 4, 7, 10\},\ \{1, 10, 100, 1000\},\ \{5, 6, 7, 8\}$$

On the other hand, the sets

$$\{3, 4, 5, 6\},\ \{4, 6, 9, 10\}$$

are not sum-free, because $3 + 3 = 6$ and $4 + 6 = 10$.

Given any set of natural numbers, it is very interesting to think about how big its biggest sum-free subset is. The following theorem is very surprising:

Theorem 19 (Lower Bound on Sum-Free Subsets). *Given a set of natural numbers A, we can always find a sum-free subset of size at least $\frac{1}{3} \cdot |A|$.*

To prove this result, we will make use of a lemma.

Lemma 5. *Let A be a set of real numbers. Define*

$$B = \{x \cdot a \ \forall a \in A\}$$

for any real $x \neq 0$. It holds that A is sum-free if and only if B is sum-free.

This should seem very intuitive as the condition of being sum-free is linear and the real number we multiply with can just be factored out. We are now ready for the proof of the theorem.

Proof. Let A be a set of natural numbers and $x \in (0, 3)$ be chosen at random. We define

$$B = \{x \cdot a \ \forall a \in A\}$$

and let C be the set of all elements $b \in B$ with the following property:

$$\lfloor b \rfloor - 1 \text{ is divisible by 3.}$$

Using indicator variables and linearity of expectancy, it is not hard to see that

$$\mathbb{E}\big(|C|\big) = \frac{1}{3} \cdot |A|,$$

since b is uniformly distributed over $[0, 3 \cdot k]$ for some $k \in \mathbb{N}$ and

$$\mathbb{P}\Big(\lfloor b \rfloor - 1 \text{ is divisible by } 3\Big) = \frac{1}{3}$$

for all $b \in B$. We can therefore find $x \in (0, 3)$, such that

$$|C| \geq \frac{1}{3} \cdot |A|.$$

Finally note that the set C is sum-free due to its definition. It follows by the lemma that the corresponding subset of A we get by dividing every element by x, is also sum-free. We have therefore found a sum-free subset of A with the desired size.

□

5.3 Intersecting Subsets

The following theorem from combinatorics is about the maximal number of subsets of a given set, where every subset is of some fixed length and every pair of subsets has at least 1 common element.

Theorem 20 (Erdős-Ko-Rado). *Let $n, k \in \mathbb{N}$ with $n \geq 2k$. Let H be a collection of pairwise intersecting subsets of $\{1, 2, \ldots, n\}$, all of size k. It follows that*

$$|H| \leq \binom{n-1}{k-1}.$$

Note that if $n < 2k$, then every pair of subsets will be intersecting.

Another crucial observation is that if we arrange the numbers from 1 to n on a circle in some order, the maximal number of contiguous arcs of length k of this circle which belong to H is at most k.

Proof. Let $n, k \in \mathbb{N}$ with $n \geq 2k$. Suppose H is a collection of pairwise intersecting subsets of $X = \{1, 2, \ldots, n\}$, all of size k. Choose a permutation $\pi \in \mathcal{S}_n$ at random and, also at random, $m \in X$. Let us fix π and consider now the set

$$A = \{\pi(m+1), \pi(m+2), \ldots, \pi(m+n)\},$$

where addition is modulo n, meaning that we wrap around the permutation when we reach the end. By the observation made earlier,

$$\mathbb{P}(A \in H) \leq \frac{k}{n}.$$

This is because, as we vary m, the set A is going through all continuous arcs of a circle defined by π. Looking at the big picture we also see that

$$\mathbb{P}(A \in H) = \frac{|H|}{\binom{n}{k}}.$$

Combining the two results, we conclude that

$$|H| \leq \frac{k}{n} \cdot \binom{n}{k} = \binom{n-1}{k-1}.$$

□

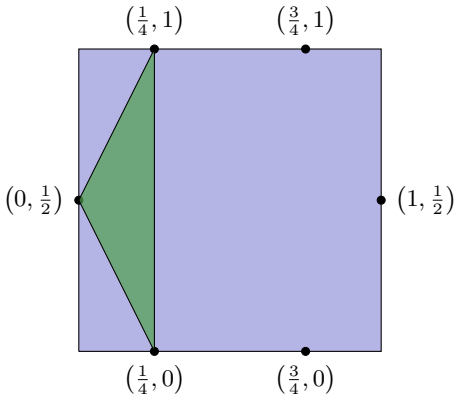
Note that in the special case where all the subsets share a common element, this maximal value is actually achieved.

5.4 Heilbronn Triangle Problem

Our last result is from geometry. Given $n \in \mathbb{N}$, the Heilbronn triangle problem is asking the following question: How should we arrange n points inside the unit square, such that the area of the smallest triangles defined by 3 points is maximised. The unit square is just the square with side length 1.

Example 13 (Heilbronn Triangle Problem). When given 6 points, the best way to arrange them in the unit square is as follows:

Heibronn Triangle Problem with 6 Points



The origin is chosen to be in the lower left corner of the square. In this way, the smallest triangle has area

$$A = \frac{1}{8},$$

which is the highest value achievable. For a larger number of points, the optimal arrangements will also contain points in the interior of the unit square, not only on its boundary.

Answering this question explicitly by giving a construction of such a set of points turns out to be very hard for large n. This is where the probabilistic argument comes into play.

Theorem 21 (Lower Bound on Area of Triangles). *It is possible to arrange more than n points inside the unit square, such that every triangle formed by 3 points has an area greater than or equal to*

$$\frac{1}{100 \cdot n^2}$$

for every $n \in \mathbb{N}$.

To prove this theorem we will use the method of alteration discussed in section 2.7 earlier on. When picking $2n$ points inside the square at random, we wish to show that the number of "bad" triangles is less than n. Removing one point from all of those triangles will give the desired result.

Proof. For $n \in \mathbb{N}$, let us select $2n$ points inside the unit square independently at random. Let us define X to be the number of bad triangles with area less than

$$\epsilon = \frac{1}{100 \cdot n^2}.$$

We can use indicator variables to deduce that

$$\mathbb{E}(X) = \binom{2n}{3} \cdot \mathbb{P}(T \text{ is bad}),$$

where T is a random triangle. Let a, b, c denote the vertices of the random triangle and let s be the distance between a and b. For T to be a bad triangle, the distance from the point c to the line connecting a and b can not be greater than $2\epsilon/s$. We find that c has to lie withing a rectangle around this line of width $4\epsilon/s$ and length at most $\sqrt{2}$, as this is the maximal distance inside the unit square. This results in the conditional probability

$$\mathbb{P}(T \text{ is bad} \mid s) \leq \frac{4\sqrt{2}\epsilon}{s}.$$

Given two random points inside the unit square, the probability that their distance is in the interval $[t, t + \Delta t]$ can be bounded by observing the following: For this to be true, one of the two points has to lie within a ring-shaped region around the other point of area

$$\pi \cdot (t + \Delta t)^2 - \pi \cdot t^2.$$

Ignoring the parts where this region lies outside the unit square, we write

$$\mathbb{P}(s \in [t, t + \Delta t]) \leq \pi \cdot (t + \Delta t)^2 - \pi \cdot t^2 = 2\pi t \cdot \Delta t + \pi \cdot \Delta t^2.$$

If we now make Δt be very small, this is gives upper bound for the probability that s is very close to t.

Taking the limit as Δt goes to zero, we find that

$$\mathbb{P}(s \in [t, t+dt]) \leq 2\pi t \cdot dt.$$

Combining the two probabilities and integrating over all possible distances, we deduce:

$$\mathbb{P}(T \text{ is bad}) \leq \int_0^{\sqrt{2}} \mathbb{P}(s \in [t, t+dt]) \cdot \mathbb{P}(T \text{ is bad} \mid s) = \int_0^{\sqrt{2}} 8\pi\sqrt{2} \cdot \epsilon \; dt = 16\pi \cdot \epsilon.$$

Finally, remembering the definition of ϵ, we are able to deduce that

$$\mathbb{E}(X) = \binom{2n}{3} \cdot \mathbb{P}(T \text{ is bad}) = \binom{2n}{3} \cdot \frac{16\pi}{100 \cdot n^2}$$

$$= \frac{16\pi \cdot (2n-1) \cdot (n-1)}{150 \cdot n} < \frac{32\pi \cdot n}{150} \approx 0.67 \cdot n < n.$$

We conclude that there must be some way of selecting $2n$ points in the unit square, such that less then n triangles are bad. Removing one point from every bad triangle leaves us with more than n points, where every triangle has area greater than ϵ.

□

6 Conclusion

We have looked at many different aspects of the probabilistic method. Doing so, we have seen how probability theory can even be used to prove purely deterministic theorems. We conclude that the probabilistic method can be applied to a wide range of problems, not only from combinatorics, but also from many other fields of study. The proofs given turn out to be surprisingly short and beautiful. Having analysed properties of permutations, we have found connections of their distributions with other important mathematical concepts as for example Euler's number and the Bell numbers.

I hope to have achieved not only to convey, in a straightforward way, an idea of why the probabilistic method works, but also my fascination with this topic.

Some of the results I would like to go on investigating are the "Lovász-Local-Lemma", the "Janson Inequality", the "1-2-3 Theorem", and various results from the so-called "Ramsey Theory". Also worth looking at are the concepts of random graphs, pattern avoiding permutations, probability generating functions, dependency graphs, as well as the conditions on distributions to be uniquely determined by their moments. Diving deeper into the matter and looking at applications of the probabilistic method in higher level mathematics and modern-day research would lie beyond the scope of this paper.

7 Acknowledgements

I would like to offer my special thanks for their priceless support to:

Eric Fitze, teacher of mathematics at MNG Rämibühl, who agreed to supervise this paper. Thank you for motivating me to choose a challenging topic and for the freedom you allowed me. Your uncomplicated way and your constructive feedback always kept me going. Working with you was a most enjoyable experience.

Prof. Dr. Valentin Féray, professor for mathematics at the University of Zürich, who introduced me to the topic. Your expert knowledge helped me to narrow down the scope of my paper and to expand my mathematical horizon. Thank you for your time and for everything you taught me. It was a great honour and pleasure.

Prof. Dr. Thomas Kappeler, professor for mathematics at the University of Zürich, whom I got to know through the outreach programme "Junior Euler Society" a long time ago. You shared my enthusiasm for the probabilistic method as a topic for my paper and recommended Valentin Féray. Thank you for offering me the unique opportunity to work in a University setting.

Katharina Blarer, who spent many hours reading through my work, correcting mistakes and – last but not least – serving me cups of hot tea while I was working. Your opinion as well as your linguistic competence were very valuable. Thank you for encouraging me all the way through and for being such a supportive mother.

Thomas Imbach, my father, with whom I spent a lot of time talking about my ideas for this paper. Doing so helped me to order my thoughts and to gain an outside perspective. Thank you for being such a great listener and for believing in me from the start.

8 Bibliography

Alon, N., Spencer, J. H. (2008). *The Probabilistic Method*, 3rd ed. Hoboken, NJ: John Wiley & Sons, Inc.

Balachandran, N. (2010). *The Probabilistic Method in Combinatorics*. Pasadena, CA: California Institute of Technology.

Bevan, D. (2015). *Permutation patterns: basic definitions and notation*. Glasgow, Scotland: University of Strathclyde.

Chen, E. (2014). *Expected Uses of Probability*. Plano, TX: Mathematical Reflections, Issue 6.

Corteel, S., Louchard, G., Pemantle, R. (2006). *Common intervals in permutations*. Strasbourg, France: Discrete Mathematics and Theoretical Computer Science, DMTCS, vol. 8, pp. 189-214.

The following websites were of great use while doing my research:

www.cut-the-knot.org/Probability/ProbabilisticMethod.shtml

wikipedia.org/wiki/List_of_probabilistic_proofs_of_non-probabilistic_theorems

wikipedia.org/wiki/Random_permutation_statistics