Das empirische Gesetz der groSSen Zahlen verstehen

Martin Wabnik

Probability Theorie

The Basics

This section aims to present probability theory reduced to simple principles. It deals exclusively with material that can be taught in general education schools.

What is probability theory about? From a certain point of view, one could say: it is about making predictions for the future. With the help of probability theory, we can say how likely different futures are.

When we work with probability theory, we have a random experiment and know the *probability of an event* E. If we perform the random experiment several times that is, n times we obtain a *relative frequency* $h_n(E)$ of the event E. The whole art of probability theory now lies in determining the *probability of a relative frequency* $\mathbf{P}(h_n(E))$.

If we toss a fair coin 10 times, we want to know, for example, how likely it is to get exactly 5 H ($H \equiv$ *heads* $\equiv$ head). Which means the relative frequency of H equals $\frac{1}{2}$. Or we roll a die three times and want to know how likely it is that the relative frequency of the event "rolling a 6" is at least $\frac{1}{3}$.

Now, many people find this question contradictory because we cannot know what will happen in the future. Nobody can predict the future! In addition, the statement of a probability can be understood as something uncertain or speculative. Therefore, we have to face the following fundamental question:

What do we know about random experiments
that we have not yet performed?

If we use our standard model, in which we draw balls from a box, and we call the set of all balls the population and the drawn balls the samples, then we can phrase the question as follows:

What do we know about the samples
that we have not drawn yet?

We do not know what the outcome of a single random experiment will be, but we can specify which possible combinations of outcomes exist if we repeat the random

experiment several times. For this we use combinatorics. Combinatorics holds today and as we assume also in the future, and there is nothing uncertain about it. We can state exactly which and how many combinations of outcomes there are and what relative frequencies they have. In this way, we can assign probabilities to the different relative frequencies. At first, we will achieve this by simple counting. So we can answer the fundamental question stated above as follows:

We know how likely the different relative frequencies are.

This statement about probability has nothing uncertain about it, because it is a statement about numbers of combinations. A probability statement is therefore a statement about existing possibilities. Mathematics says nothing about what will actually happen.

We can even go one step further and reduce probability theory to an extremely simple main theorem: We will find that there are many combinations whose relative frequencies of outcomes are close to the probabilities of the outcomes, and that there are few combinations whose relative frequencies of outcomes are far away from the probabilities of the outcomes.

So we have the **Main Theorem of Probability Theory**:

Most relative frequencies are similar to the probability.

If we start from our standard model of drawing balls from a box with blue and red balls, we can state this theorem even more clearly. Since we draw balls from a population (all the balls in the box) and the possible combinations of blue and red balls are called samples, we can write:

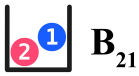
Most samples are similar to the population.

This main theorem will guide us toward understanding probability theory and the relationship between relative frequency and probability. We will be able to understand mathematically why, when tossing a coin 100 times, we expect a relative frequency of the outcome H of about 0.5: because there are many more combinations with relative frequencies of H near 0.5 than there are other combinations. Then we no longer need to speculate whether, after many repetitions of the experiment, there might be some strange kind of convergence and the coin might not be able to do what it wants, but we will simply assign a probability to each possible relative frequency. In this way, we will know how likely each future is.

We live in a world in which
1) relative frequencies are important, and
2) there are extremely many more combinations of outcomes with relative frequencies that are similar to the probabilities of the outcomes than there are other combinations of outcomes, when we repeat a random experiment many times.

In many areas, it has proven useful to bet on those combinations for the future of which there are very many. That is what makes the success of probability theory.

Probabilities of Relative Frequencies



Let us start with the simplest random experiment imaginable: We have the box B_{21} with one blue and one red ball. We draw randomly with replacement and with order. The probability for *blue* is, just like the probability for *red*, equal to 0.5.

When we draw a ball, we record the color and number of the ball and put it back. The notation shown on the left can then be understood as one possible outcome of fourfold random drawing with replacement and with order from B_{21} or also as an ordered sample of size 4 from the population B_{21}.

Now we can write down the possibilities of double, triple, quadruple, etc. draws, organized by relative frequencies. For comparability reasons, we here decide to illustrate only even numbers of draws.

For simplicity, we will almost always call the drawn sequences of blue and red balls samples, rather than outcomes of multiple draws.

Even without knowledge of combinatorics, students can recognize the following illustrations as complete lists of all possible samples of size 2, 4, 6, etc. from B_{21}.

Abb. 1 Samples of size 2

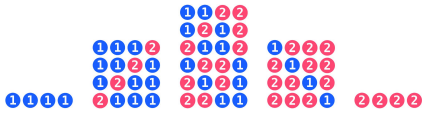


Abb. 2 Samples of size 4

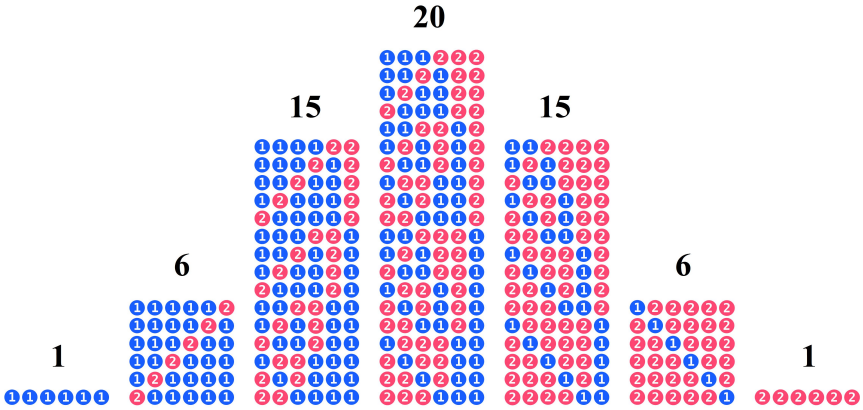


Abb. 3 Samples of size 6

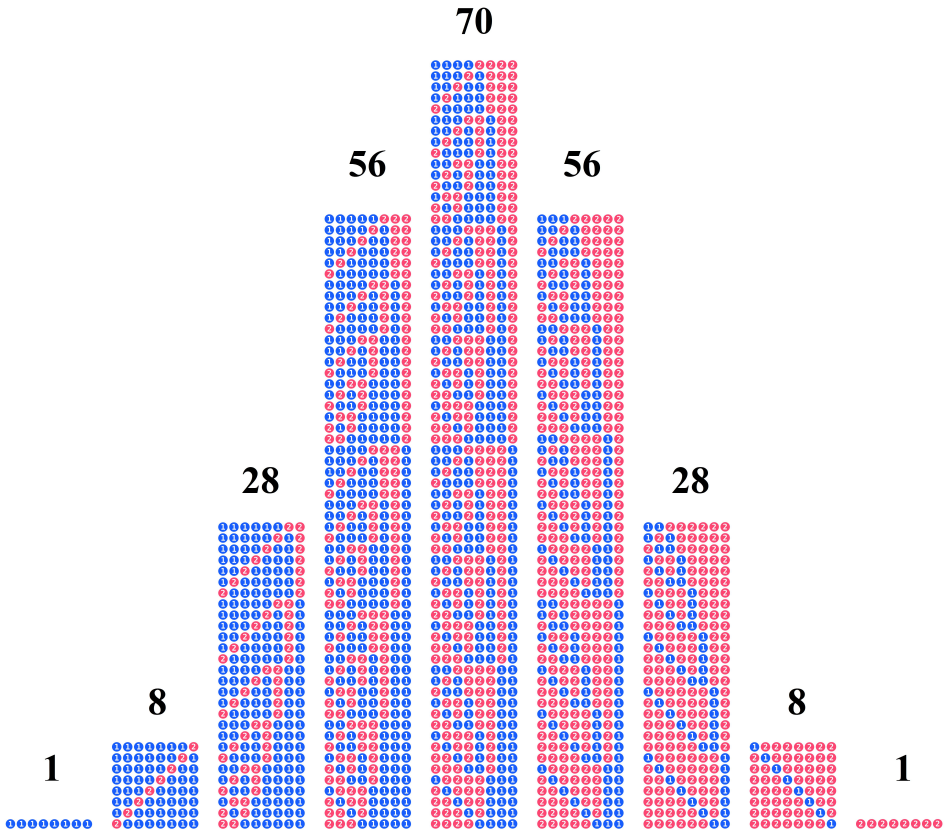


Abb. 4 Samples of size 8

We can calculate the probability that the relative frequency of *red* in a sample of size 6 is equal to $\frac{1}{3}$ by dividing the number of samples with this relative frequency by the total number of possible samples, that is

$$\mathbf{P}\left(h_6 = \frac{1}{3}\right) = \frac{15}{1+6+15+20+15+6+1} = \frac{15}{64} \approx 0.234$$

In this way, for a given sample size, we can assign a probability to each possible relative frequency of red balls.

Notes on Didactics

1) In the introductory lessons of probability theory, the expression $\mathbf{P}\left(h_6 = \frac{1}{3}\right)$ will probably be replaced by a more intuitive one, which is not a problem for the further content development.
2) Apart from fraction arithmetic, nothing else is required as prior knowledge.
3) With more than two balls in the box and for larger sample sizes, we will not get very far by drawing the balls manually, but the principle of determining probabilities of relative frequencies will not change.

Relationship between Relative Frequency and Probability

If we convert the diagrams with blue and red balls into bar charts, we can place the resulting charts with sample sizes up to 10 side by side and compare them.
For better clarity, the four outer bars are colored blue and the middle bar is colored red.

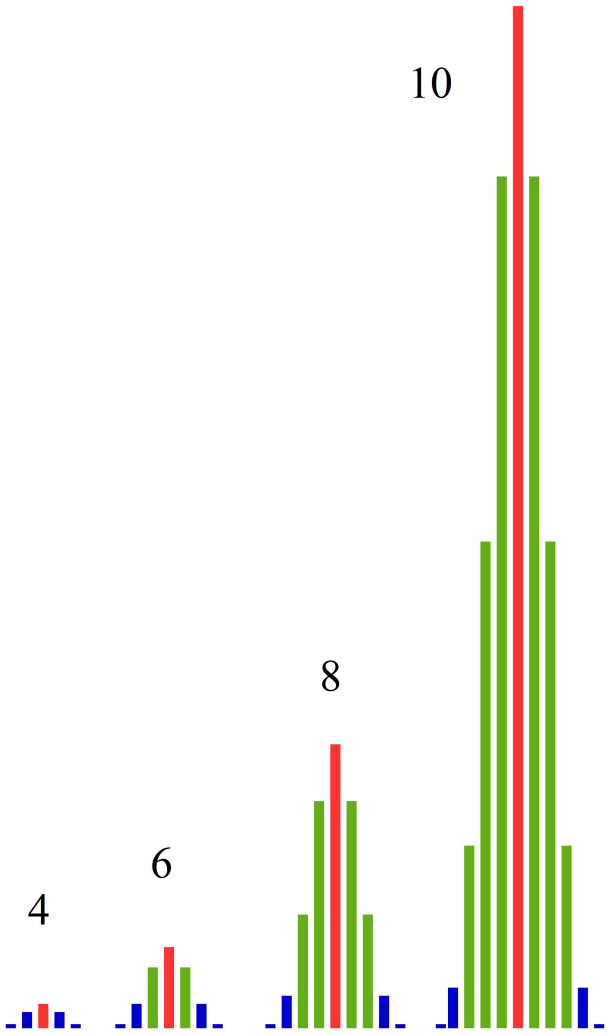


Abb. 5 Counts of samples of size 4, 6, 8, and 10

Two things can be observed from this diagram:
1) There are more samples in the middle than at the edges.
2) These differences become larger as the sample sizes increase.

One conclusion is that the probability for a middle outcome becomes larger the more often we perform the experiment. And since the middle contains *the* outcomes whose relative frequencies of red balls are close to 0.5, it follows: The probability that the relative frequency of *red* is close to the probability of *red* increases the more often we perform the experiment.
Or formulated even more simply:

It becomes increasingly likely that
the relative frequency of red balls is close to 0.5.

More generally:

As the number of repetitions of the experiment increases, it becomes increasingly likely that the relative frequencies of an event approach its probability.

And this formulation may also be useful:

There are more combinations whose relative frequencies of an event are close to the probability of the event than there are other combinations. These differences become larger the more often we perform the random experiment.

In terms of statistics, this means: The larger the sample size, the greater the probability that the sample is similar to the population where ßimilar"here means that the relative frequency of an event in the sample is close to the probability of that event.
In other words: The larger the sample size, the greater the probability that the relative frequency is close to the population proportion.

Let us look at another random experiment:
If we have one blue and two red balls in the box, we observe a similar pattern only with the difference that now the counts *of* the samples whose relative frequencies of red balls are close to $\frac{2}{3}$ increase faster than the other samples.

Abb. 6 Samples of size 2 from B_{32}

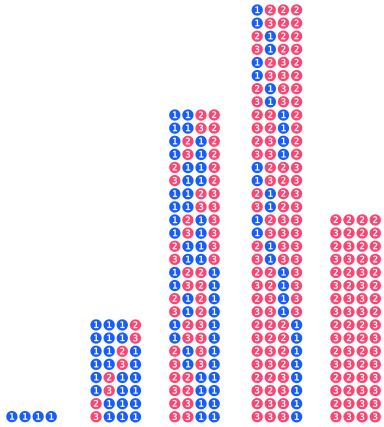


Abb. 7 Samples of size 4 from B_{32}

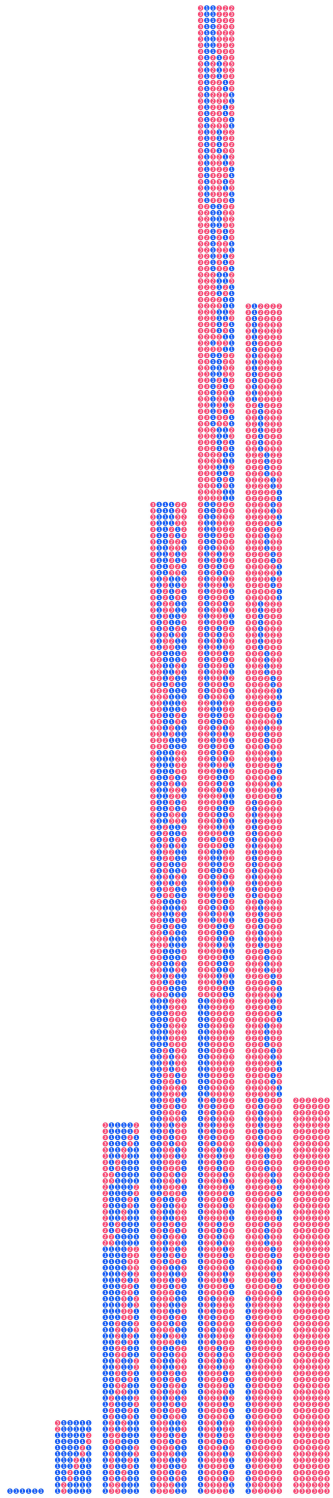


Abb. 8 Samples of size 6 from B_{32}

And another example is shown: We draw from B_{51} and the clustering of the relative frequency occurs at or near $\frac{1}{5}$.

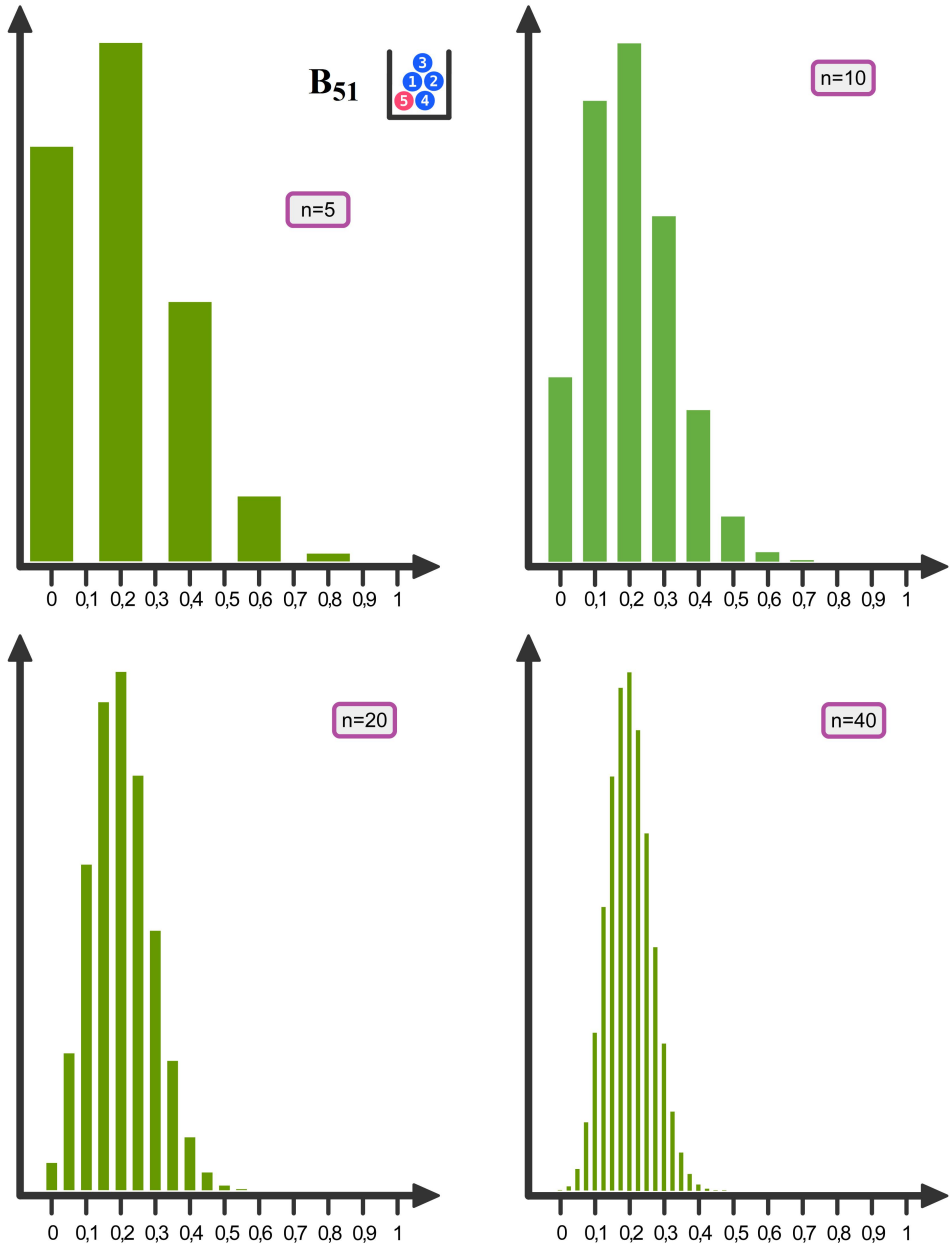


Abb. 9 Samples from B_{51}

Note on Didactics
Even if students cannot calculate the individual counts of combinations without knowledge of combinatorics, such diagrams make sense in introductory probability lessons, because they can provide a good intuition for the actual relationships between relative frequencies and probabilities. The illustrated counts can, in this case, serve as purely observational facts without computational proof.

The Galton Board

There is an excellent way to make the relationship between relative frequency and probability literally tangible: The Galton Board!

In the previous section, we considered the basic experiment of drawing a ball from the box B_{21}, which contains one blue and one red ball. The probability for the blue ball is equal to the probability for the red ball, 0.5. If we then replace the ball and draw again, we find the same probabilities again.

It is similar with the Galton Board. A ball falls from the top onto the first round wooden peg. The ball can then fall either to the right or to the left. If the peg is positioned exactly in the middle, the probability of falling to the right is 0.5. The probability of falling to the left is also 0.5.

Regardless of whether the ball falls to the right or left, it hits another centrally positioned peg, at which it falls to the right with probability 0.5 and to the left with probability 0.5, and so on.

If the ball remains in the middle bin, it has previously fallen to the right and to the left equally often. This corresponds to an outcome with as many blue as red balls from our previous experiment. For the ball to remain in a bin to the right of the center, it must have fallen to the right more often than to the left. For the rightmost bin, the ball must have deflected to the right at every peg.

If we let many balls fall through the arrangement of pegs, we typically obtain a ball pattern like the one shown in the photo.

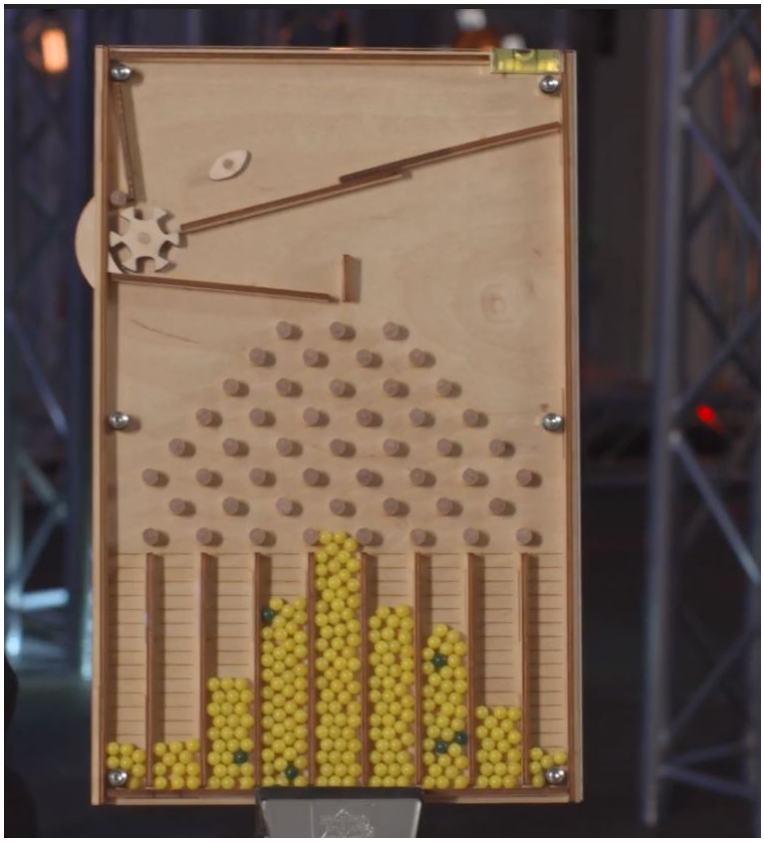


Abb. 10 Galton Board: When a ball hits a peg, it can fall to the left or to the right. Both possibilities have a probability of 0.5. The same situation occurs at the next level. Shown is a Galton Board with eight levels.

If we distinguish the possibilities *left* and *right* at each level of this Galton Board, the 8-tuples corresponding to a ball falling into one of the middle bins have a relative frequency of *right* of approximately 50%. We observe that more balls are in the middle than at the edges. But why is this so? It is not because the balls want to balance falling left and right, nor because they must obey a law of chance, but quite simply because *more paths lead to the middle than to the edges*. Once again, we can see that the relationship between relative frequency and probability is a matter of combinatorics, not of convergence.

For this insight, it is initially sufficient to count the paths of the first levels. This process is illustrated in the following figure. Here six paths lead to the middle and only one to the right edge.

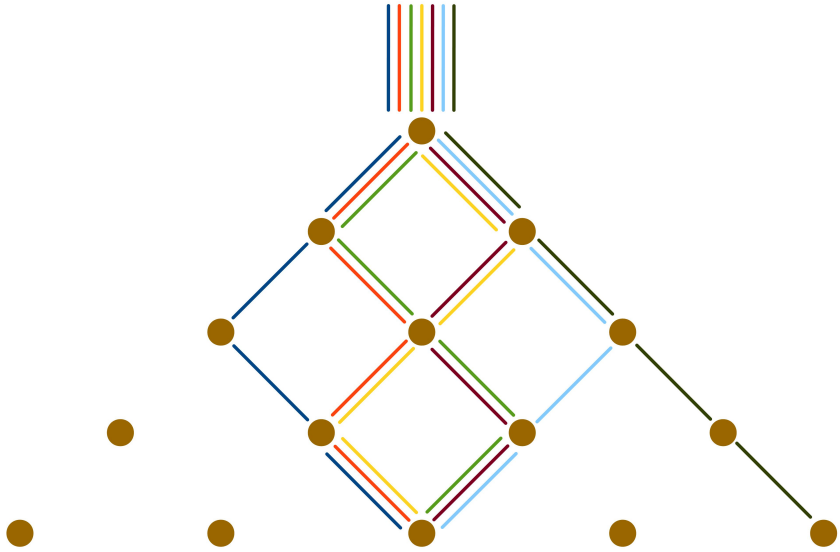


Abb. 11 Galton Board, more paths to the middle

Even without investing further mathematics, we can already see what the main misunderstanding regarding the empirical law of large numbers is: The relative frequency does not have to approach the probability even after very many repetitions of the experiment. Applied to the Galton Board, this would mean that a ball could no longer fall to the right or left after a certain number of levels. We would have to physically install a barrier so that a ball that has already fallen to the right many times cannot fall to the right again, because after all, the ball does not know its position. Even though we have only seen an eight-level Galton Board, it is technically clear: Even if we add arbitrarily many levels to the Galton Board, the relative frequency can always deviate as far as imaginable from the probability. At each level, the ball always has the possibility, for example, to fall to the right.

Pascal's Triangles

If we want to know how many paths lead to each bin without using combinatorics, we can use Pascal's Triangle. This triangle is built level by level from top to bottom, by adding a 1 on each side at the outer edges on every level, and forming the inner numbers as the sum of the numbers directly above to the left and right. This corresponds exactly to the path of a ball through the Galton Board: To hit a peg or fall into a bin, the ball can come from the top left or the top right. The sum of the paths from the top left and top right is the number of paths that lead to this peg or the corresponding bin.

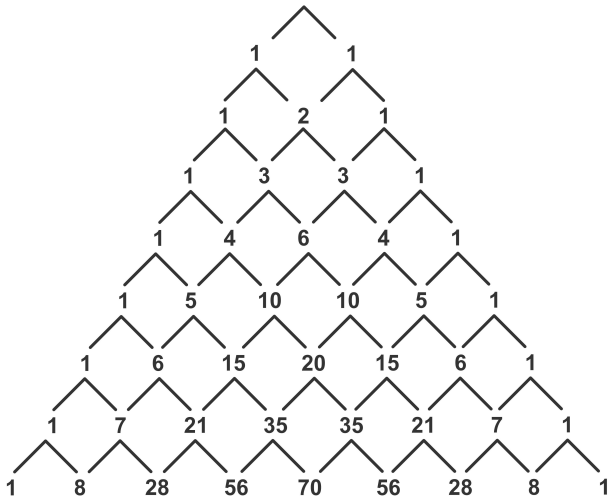


Abb. 12 Pascal's Triangle

Applied to the experiment of drawing a ball from B_{21}, it could look, for example, like this:

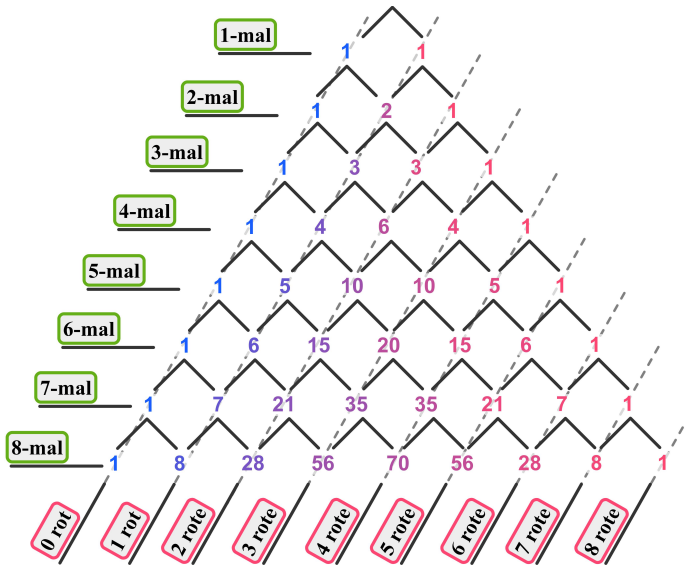


Abb. 13 Pascal's Triangle, drawing balls from B_{21}

From the bar charts we constructed for repeated draws from B_{21}, we saw that there are more combinations in the middle than at the edges. We could also observe that this trend becomes stronger the more we draw. With Pascal's Triangle, we can additionally see why this is the case: Because only a single 1 is added at the outer edges from level to level, the counts of combinations at the edges increase only slowly. In the middle, however, already large numbers are summed, resulting in even larger

numbers.

Extended Pascal's Triangles

If we draw from B_{32}, we can determine the counts of possibilities entirely without combinatorics using an extended Pascal's Triangle.

To motivate this method, let's look at how we can proceed with repeated draws from B_{32} to list all combinations of blue and red balls.

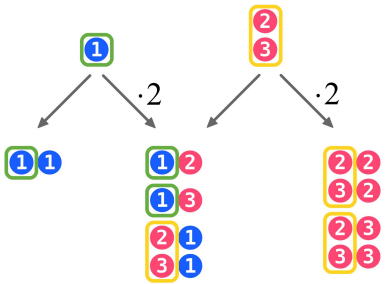


Abb. 14 Extended Pascal's Triangle Step from 1 to 2

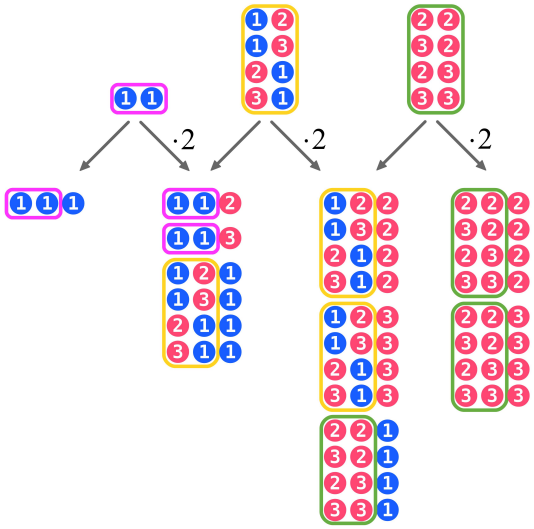


Abb. 15 Extended Pascal's Triangle Step from 2 to 3

There are many ways to list all combinations of balls from B_{32}. To obtain, for example, all 3-combinations with exactly one red ball, we can
1) add the available red balls on the right to all pairs that contain no red ball, and then
2) add the blue ball on the right to all pairs with exactly one red ball.
Continuing this systematic procedure, we obtain the following extended Pascal's Triangle.

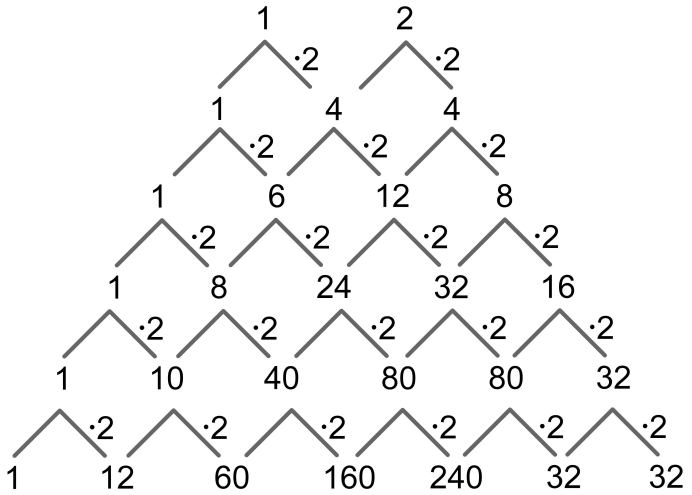


Abb. 16 Extended Pascal's Triangle

Specifically for drawing from B_{32}, we can also design the extended Pascal's Triangle as follows:

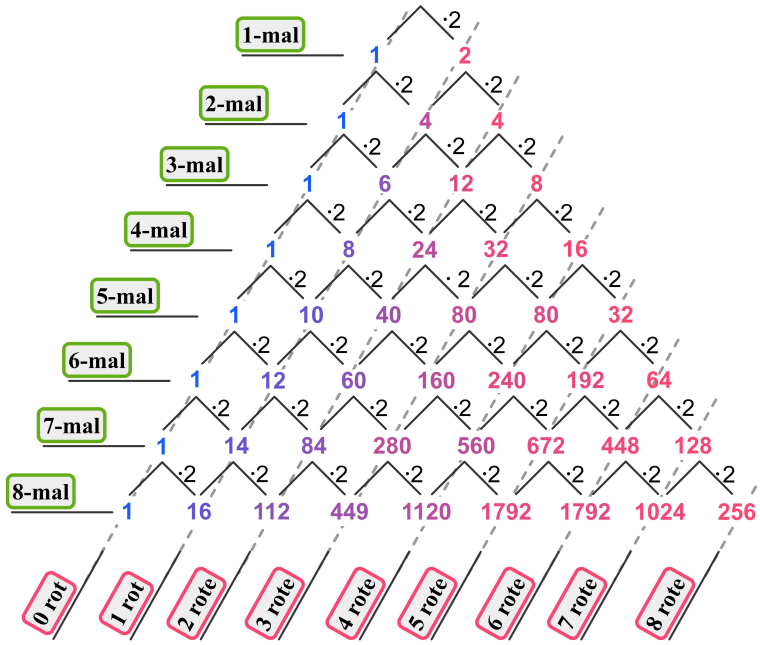


Abb. 17 Extended Pascal's Triangle, drawing from B_{32}

Here too, we can read off various regularities concerning the relationship between relative frequency and probability. For example, the clustering now occurs further to the right of the center, which is not surprising since on the right side twice as many combinations are added as on the left side. We also see that the counts grow much faster than when drawing from B_{21}. Many people who see these numbers for the first time are surprised by how quickly the counts of combinations grow, even though there is önlyöne more ball in the box.
Finally, the development of the counts of combinations when drawing from $B_{10;7}$ will be illustrated with an extended Pascal's Triangle.

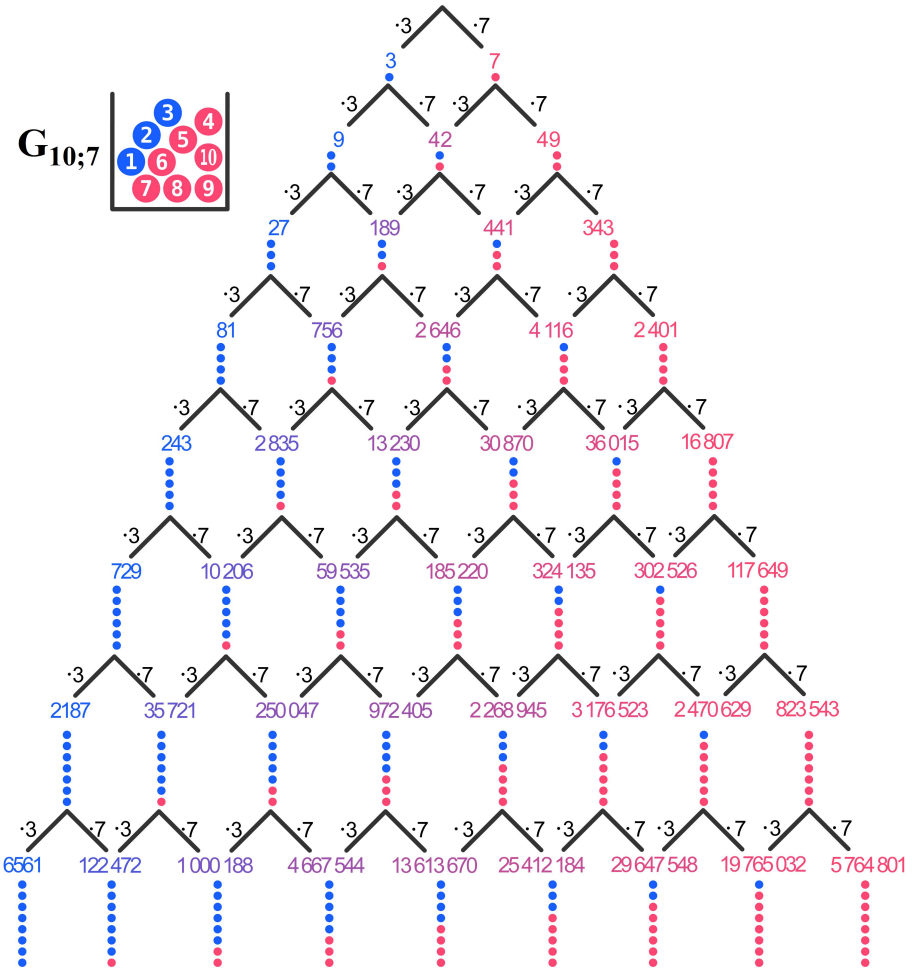


Abb. 18 Extended Pascal's Triangle, drawing from $B_{10;7}$

Consequences

No restriction of relative frequency. From Figure 4, p. 5, we can draw an important insight: We do not only have many samples in the middle, but also a sample consisting entirely of blue balls on the far left and a sample consisting entirely of red balls on the far right. Therefore, although it is more likely to draw a sample from the middle, we cannot exclude the two samples at the edges just as the random experiment cannot, because it knows nothing about our diagram. The random experiment simply produces one of the possible samples, without regard to what we humans consider usual or unusual. Each individual sample has the same probability, and so a sample from the edge can also be produced by the random experiment.

The situation does not fundamentally change even if the samples have larger sizes, e.g., 100 or 1000. While the two edge samples become increasingly unlikely relative to the samples in the middle, every possible sample still has the same probability. So why should a sample from the middle be ällowed"while a sample from the edge is too unlikely?

As a consequence: Although it becomes more likely to draw a sample from the middle as the sample size increases, we cannot exclude a single one of the possible samples. Therefore, even for very large sample sizes, samples can be drawn whose relative frequencies of red balls are far from the middle. Thus, there can be no talk of the relative frequency *having to* approach the probability or stabilize.

This applies to all other fillings of the box with blue and red balls as well: as long as there is at least one blue and one red ball in the box, a purely blue or purely red sample can also be drawn, in which case the relative frequency of *blue* or *red* is maximally far from the probability.

No restriction of the population. If we infer from the sample to the population, we can analogously exclude no populations as the source of the sample: if half the balls in a box are blue and the other half red, we can draw a sample consisting only of blue or only of red balls regardless of the sample size. To draw a sample with only blue balls, it is sufficient that there is just one blue ball in a box and, e.g., the other 99 balls are all red. Such a sample is not less likely than any other sample and therefore can be drawn just like any other. So if we have only blue balls in a sample and 100 balls in the box, there could be between 1 and 100 blue balls in the box.

Thus, we cannot rely on the fact that, if the sample is large enough, the relative frequency of red balls in the sample will already be similar to the population proportion of red balls, i.e., the probability of *red*. Humans have no way to exclude a population proportion based on a sample (except for the trivial case that there must be at least one blue and one red ball in the box if both blue and red balls appear in the sample).

Of course, after ënoughrepetitions of the experiment, one can simply take the relative frequency as an estimate for the population proportion, hoping that the relative frequency will be close to the population proportion. But it is certainly more logical and informative to assign probabilities to the different possible population proportions based on the observed relative frequency that is exactly what is intended in direct inferential statistics.

Maximum Likelihood Estimator. The established method in mathematics, which is sometimes claimed to „equate“ the relative frequency with the probability, is called the Maximum Likelihood Estimator, but it has a completely different logic: Among all possible populations, the one is sought from which the observed sample can be drawn with the highest probability. This has nothing to do with the false statement that the probability (in our case the proportion of red balls in the box) must be close to the relative frequency (the proportion of red balls in the sample). A closer examination of the Maximum Likelihood Estimator is given in the section on the basic

questions of statistics.

Weak Law of Large Numbers illustrated. What we have seen in the diagrams are the essential statements of the Weak Law of Large Numbers, which is correct and proven. We could gain these insights already after just a few repetitions of the experiment. Given this, why would one still need a vaguely formulated and additionally false claim that after many repetitions the relative frequency must approach the probability?

In simple terms, the Weak Law of Large Numbers states: If we define an (possibly very small) interval around the probability of an event, then it becomes increasingly likely, as the number of repetitions increases, that the relative frequency lies within this interval. Moreover, it not only becomes more likely, but the probability even converges to 1. This means: If we divide the number of combinations that fall within the interval by the total number of combinations, the result approaches 1 more and more as we repeat the experiment often enough. Although the following diagrams do not prove this insight, they strongly suggest it.

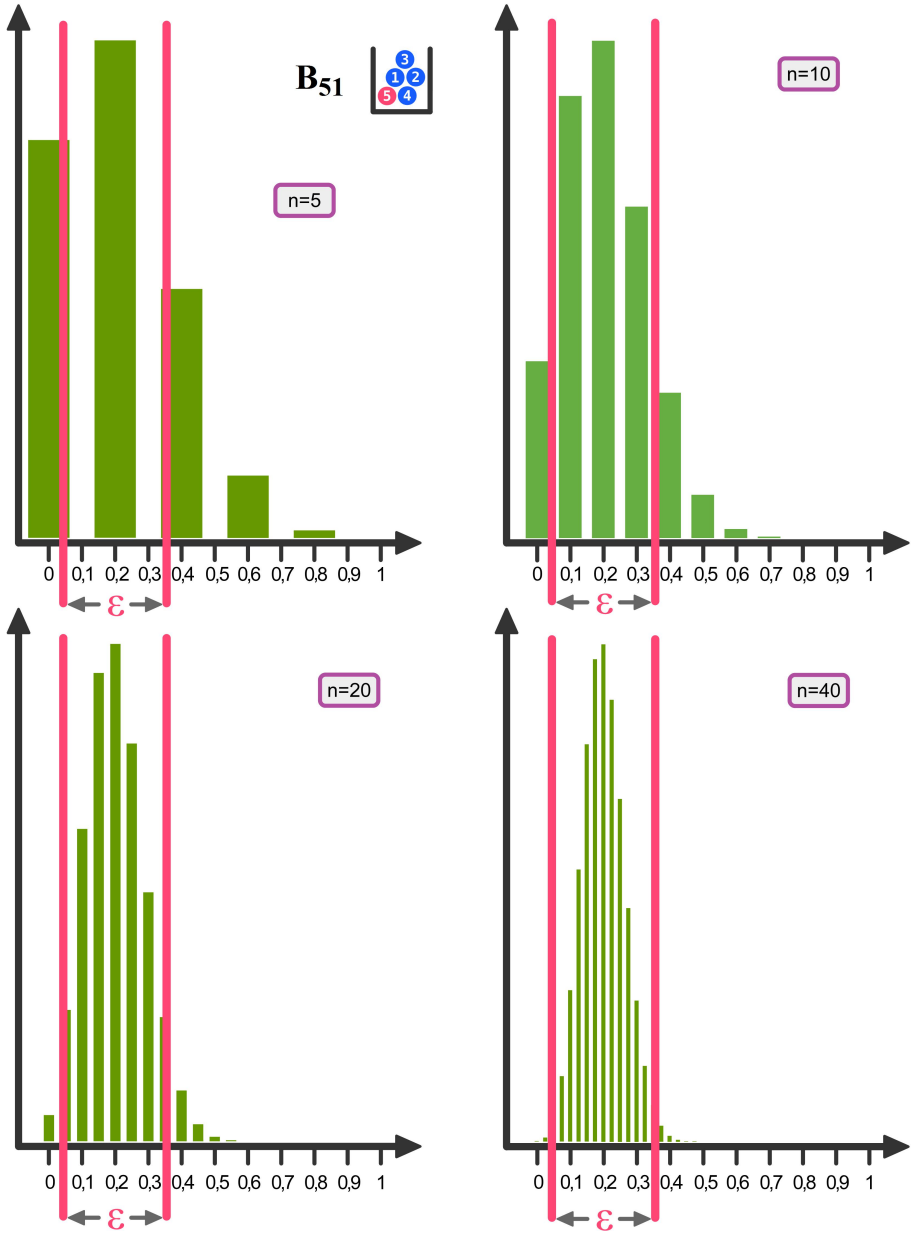


Abb. 19 Weak Law of Large Numbers illustrated

Misinterpretation. A „popular“ misinterpretation of the Weak Law of Large Numbers should be pointed out: This law says nothing about what will actually happen if we repeat a random experiment many times. For example, it does not claim that if we flip a coin frequently, the relative frequency of H will eventually fall within a chosen interval around 0.5. Instead, the Weak Law of Large Numbers tells us something about what proportion of theoretically possible outcomes lies within or outside

a chosen interval for a given number of repetitions. It is therefore a statement about the set of all theoretically possible outcomes, not a prediction of which outcomes will occur in actual repetitions of the coin toss. The two statements have nothing to do with each other.

Looking at many trials vs. looking at combinations. In connection with the empirical Law of Large Numbers, it is repeatedly emphasized that it is certainly correct if the random experiment is repeated often enough. This often leads to a strange and entirely unmathematical argument: Even if, after a certain number of trials, an undesirable relative frequency occurs, the experiment is supposedly just not repeated enough times. So if the relative frequency of H in a coin toss is too high, one simply has to toss the coin more times, and then the relative frequency will eventually settle around 0.5. This reasoning is almost the opposite of empiricism: Even if one empirically observes that a law does not hold, it still holds, because the result at hand does not count.

If we want to toss a coin, for example, 100 times, mathematics can tell us something about which outcomes are possible. We can organize these possible outcomes, classify them, or check which of these outcomes fall within a chosen interval and which do not. What we cannot do, however, is predict which outcome will occur if we toss the coin 100 times. Mathematics, for example, knows nothing about what will happen if we have tossed the coin 50 times and H has appeared too often. The point of randomness is that we *do not* know this.

Binomial distribution instead of convergence. The discussion about the meaning of the empirical Law of Large Numbers and the possible convergence of relative frequencies is not without a certain humor: In middle school, students are told that the relative frequency must converge, only to explain to the same students a few years later, using binomially distributed random variables, that this exactly does not happen.

The typical bar charts that illustrate the probability distributions of binomially distributed random variables are basically the same as those we initially saw by drawing blue and red balls. We always have a scale from far left to far right and somewhere usually in the middle a point where the possible outcomes accumulate or where the probabilities are particularly high. At the edges, there are far fewer possible outcomes and the probabilities are much lower. The more often a trial is conducted or the longer the Bernoulli sequence is, the more pronounced this difference becomes. However, even after a very large number of trials, we never reach the point where part of the scale remains empty. There are always outcomes at the far left or far right of the scale, and of course, these outcomes can occur. But why then tell middle school students that there are outcomes too unlikely to occur?

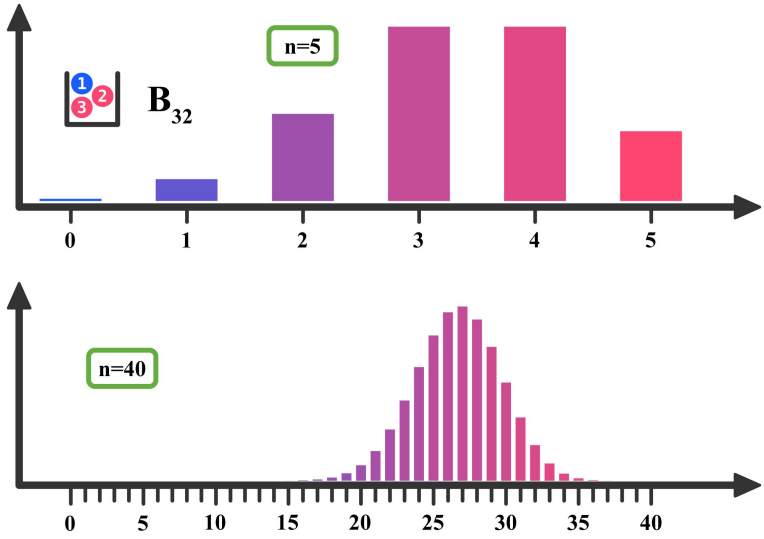


Abb. 20 Binomial distributions

Does the Probable Happen More Often Than the Improbable?

We humans bring order to our lives and daily routines by „seeing“ regularities in the events around us. One of these is the object concept: A child must first learn that the ball that rolled behind a door is not gone, but still exists there and can be retrieved at any time.

When organizing our daily lives, we follow various principles: We choose *the* regularity that is simplest, most practical, safest, etc. Instead of understanding the ball as an object that exists behind the door, we could also assume that the ball ceases to exist whenever we do not see it and only re-enters existence when and perhaps because we look at it.[1] However, this way of thinking is likely much more complicated than the object concept and therefore finds few adherents.

The rule that the probable happens more often than the improbable, and also the reverse rule, namely that something that happens frequently is likely and something that happens rarely is unlikely, belongs to the regularities that structure our lives. Perhaps some people also respond with hostility to any question that might call this principle into doubt.

Let us nevertheless look at what mathematics says. We cannot mathematically speculate about what will happen in the future, but if we specify how many times we want to perform a random experiment, we can calculate. If we, for example, plan to toss a coin 1000 times in the future, we can already say which 1000-tuples are

[1] It can be very productive to discuss with students whether aliens would use mathematics. What if aliens had a completely different concept of reality than we do and believed that nothing exists when one closes their eyes and the world is recreated when one opens their eyes assuming aliens have eyes. But one does not even need to engage in literally in this case otherworldly considerations: Would we have the same mathematics if humans could effortlessly recognize quantities? If we knew how many sheep are on the meadow without counting? We can recognize a single black sheep in a large flock of white sheep at a glance. Would we have different mathematics if we could not do that?

possible. We also know all possible relative frequencies of H and can give their probabilities. We can also order the 1000-tuples by changes in color or by any other criterion and determine the probabilities of the relative frequencies generated according to these criteria.

What we absolutely cannot do, however, is exclude a particular 1000-tuple from occurring. Each has the same probability, because each 1000-tuple occurs only once in the set of all possibilities. Therefore, there are no regularities in these 1000-tuples that influence the occurrence of certain outcomes, just as the coin during the trials does not follow laws that could affect the relative frequency of, for example, H.

The concept that the probable occurs more often than the improbable also assumes that coins, dice, etc., always behave similarly. If a coin has shown H in approximately 50% of the trials so far, it should do so in the future and even more so in the long run. This principle is strong causality: Similar causes lead to similar effects. (As opposed to weak causality: Only exactly identical causes lead to identical effects.) Humans could hardly manage daily life if we did not believe in these principles.

Randomness, however, spoils this „calculation“: If from now on we perform exactly $10,000$ coin tosses in our lives, we could get H $10,000$ times without the coin or our tosses being special in any way. In other words, it could happen that from now on, every time we toss a coin, we get only H for no deeper reason. Then what is probable namely achieving a relative frequency of H of about 0.5 simply would not happen. Thus, the probable does not necessarily happen more often than the improbable.

Are the Causes Probable? The situation is similarly catastrophic regarding the causes of random outcomes. If we draw balls from a box with replacement and after many trials we have drawn very few blue and very many red balls, this does not mean that the box contains mostly red balls. So, if something happens frequently like drawing red balls this does not mean that the probability for it must be high. The cause can also be extremely unlikely. As long as we do not have the possibility to open the box and look inside, we humans have no way to exclude certain proportions of blue or red balls in the box (except for 0 and 1).

In short, the regularities that the probable happens more often than the improbable and that what happens frequently is likely collapse as soon as we specify how often a random experiment is to be performed. Then we can even indicate how probable the completely improbable is. On the other hand, mathematics can also make us notice that everything that happens is unique, even though we are often not aware of this in everyday life.