An Intuitive Course in Probability (and Statistics)

for data science

Caio Velasco

BSc in Mechanical Engineering, Federal University of Rio de Janeiro

Master in Public Policy, University of California Los Angeles

June 2023

“When you explain a why, you have to be in some framework that allows something to be true. Otherwise, you’re perpetually asking why. ”

Richard Feynman, Theoretical Physicist

Contents

List of Figures

The Real Motivation

Even though I studied a semester of *Probability and Statistics* at the Engineering School back in Brazil and a similar course at UCLA, I realized that I hadn't actually mastered the subject. I knew how to solve some things and had a decent understanding of some complex concepts, but at some point, things got a bit more complicated and I started to struggle. I noticed I had some gaps in my knowledge, but I didn't know where they came from. Thus, I decided to take a step back and start from zero.

It was based on this experience that I found the solution and created the **Intuitive Course in Probability (and Statistics)**, *for data science*. This is what I believe:

- No matter what your goal is in the world of data, **it's crucial** you understand what lies behind the mathematical framework of probability theory.

I kindly invite you to ask yourself the following question: *Do I know what is described below?*

> The **Statistical Model** was developed with the objective of formalizing *chance regularity* patterns (i.e., the *systematic information*) present in observed data. The mathematical formalization (based on probability theory) underlying this model seeks to capture this systematic information, which is generated by *stochastic mechanisms*. Therefore, to make a decision under uncertainty in a logical and consistent manner, we need to establish a (modeling) framework for *stochastic phenomena* and move from there.

Well, if your answer is **no**, take the course :)

I believe you will break some barriers if you dedicate some time to learn this.

Chapter 1

Probability

In this chapter, the main objective is to learn the concept of **Probability**.

We will start by understanding what it means for something to be *Stochastic*, a concept that underlies the world of statistics. Then, we will learn about the idea of *Statistical Regularity*, a **crucial** characteristic that is **observed** in situations we believe to be of the "stochastic" type (note that the word *observed* is quite important and widely used). By learning about these concepts, you will realize that Probability is nothing more than an attempt to mathematically "tame" that statistical regularity in a coherent manner.

You will use these concepts to build a **Statistical Model**, and for that, you need to know how to build a **Probabilistic Model**.

With that, you will be ready to understand thoroughly what is an *Experiment* in Statistics.

This chapter was mostly based on the book *Probability Theory and Statistical Inference* by Aris Spanos [1], one of my favorites. I will try to keep it simple and concise. I will also make references to other sources if needed.

Let's get started.

1.1 What is a "Deterministic" Phenomenon?

A phenomenon is *deterministic* when its future is **not** uncertain.

A simple example is the mathematical function, which is a rule with certain properties (*all elements in the domain* ***must*** *be covered and you* ***cannot*** *associate the same element in the domain with multiple elements in the co-domain* - you should not forget this, but most people do..). Therefore, if you know the rule, then as soon as you know the input (the *argument* of the function), you also know the output (the *value* this function *takes*).

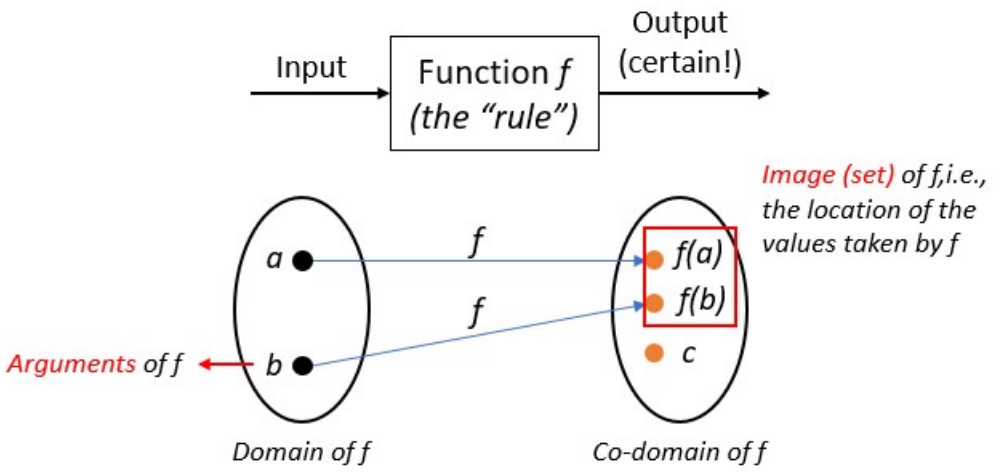

Figure 1.1: The Function Blackbox

If we take a function f with the rule $f(x) = 2x$ and evaluate the value this function takes for $x = 3$, you know *with certainty* or *for sure* that the function will take the value $f(3) = 6$. There is no uncertainty about the future in this process. (also note the terms being used, they are the language here).

Another way to put is: a situation is *deterministic* when the previous state can *with certainty* predict the next state. In other words, given that we have an initial state, we can always predict any future state of interest because the path to get there is deterministic.

Do not forget: Determinism relates to **certainty** in predicting the future.

1.2 What is a "Stochastic" Phenomenon?

Let's define what we mean by *stochastic phenomena* and introduce the concept of *statistical regularity*.

Definition - Stochastic Phenomenon and Statistical Regularity

A **stochastic phenomenon** is a situation where the data exhibit patterns of *statistical regularity*. It is also called by [1] as *chance regularity*. Okay, but what does that mean?

The data we observe in the real world **exhibit certain patterns**, called *patterns of statistical regularity*, and what we want to do is to mathematically model these patterns. Constructing this model is precisely what we will use to draw conclusions about what is happening when uncertainty is present.

Pay close attention now, as the whole construction of the concept of probability is based on the following idea:

> The concept of probability is built on top of the property that the data we observe exhibit certain regular patterns.

Great, but how can something uncertain exhibit regular behavior?

To answer this question, we need to break down the concept into two parts:

$$chance + regularity$$

By breaking it into two parts, we can differentiate the idea behind each and understand how the idea of *uncertainty* (the chance of something happening) is connected to a certain **pattern**:

The real world presents uncertainty, and of course, we don't know about the future. However, this notion is closely linked to the fact that at the **individual** level, we have uncertainty, but at the **aggregate** level, we notice a certain pattern. (This is **very important**. See below).

Intuition - Statistical Regularity

Consider the experiment: Rolling two dice and observing the sum.
You know (with certainty) the results: $2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12$
But where is the *uncertain* part, and where is the *regular* part?

Individual Level

If you roll the two dice for the first time, you won't know what will be the result. If you roll them for the second time, you still won't know. If you roll them for the thousandth time, you still won't know.
So, at the *individual* level, we note the idea of **chance** or **uncertainty** in the data.

Aggregate Level

Now, if you roll the dice numerous times, **keep track** of all results, and plot them on a histogram-type graph (the one that aggregates the **frequency** of equal results in an experiment - see more below), you will be able to observe a certain pattern.
Therefore, at the *aggregate* level, we observe a certain **regularity** in the data.

Note that this figure has an interesting property, the *aggregation* of data.

In this particular case, the aggregation rule was a simple count of each equal result that appeared in the 100 attempts. For example, if the $sum = 7$ appeared 17 times, this histogram of *relative frequencies* has a height (y-axis) of 17%, or 0.17.

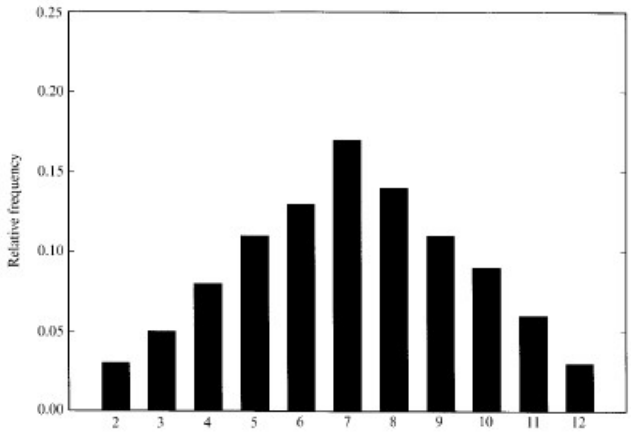

Figure 1.2: Histogram of the sum of two dice rolled 100 times

Note: the **Histogram** can be presented in two ways:

- showing the *absolute frequency* on the y-axis, which represents the count of each result
- or showing the *relative frequency* on the y-axis, which represents the percentage of each result in the whole, in this case, out of the 100 attempts.

  Either way, the histogram presents the idea of **aggregation**.

Note: The histogram was just an example to help us visualize the patterns of *chance regularity* and justify the whole story of Probability. There are other ways to visualize such patterns, and it is precisely in these other forms that **extremely** important concepts arise, such as:

- the concept of **Distribution** of the data, which refers to how the results of the experiment are distributed after several trials, which makes us believe that there is a "law" that is stable enough to guide this format (which can be displayed by the histogram).
- the concept of **Independence** refers to the (non) influence that a previous result (or previous results) exerts on the following ones. In this

case, the way to visualize it is to notice that if you cover a **a little piece** inside the red window in the figure below and try to predict the next result, you won't be able to predict it. If you cover up another little piece, you still won't be able to predict it. Therefore, knowing the information that has not been covered does not help *predict* the next result. In this case, we say that the results of the experiment are *independent*.

- and the concept of **Homogeneity**, which refers to a property exhibited by the data that makes us believe that when we slide that red window to the right, the vertical variation found in the results is practically constant. When it is not, we say that the data exhibits *heterogeneity*.

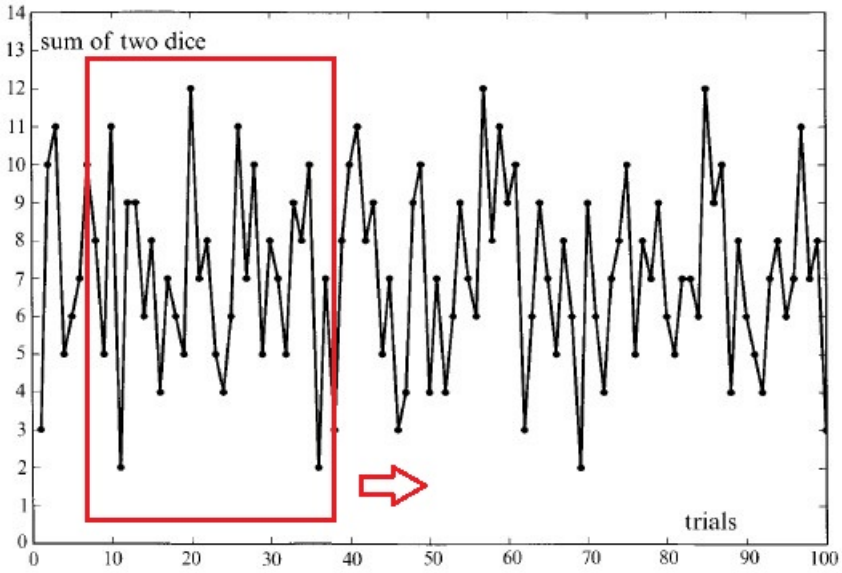

Figure 1.3: Chart of the chronological sequence of the sum of 100 dice rolls

Did you notice that we touched on the subject of *frequency* by the introduction of the *histogram*, right? Did you also notice that we extracted very important concepts like *Distribution* and *Independence* from the same subject? Furthermore, did you notice that we focused on the idea of **relative** frequency?

Well, that's where we begin to unravel the **intuition** behind Probability.

Consider the same experiment: rolling two dice and observing the sum.

However, let's focus on the structure of the data behind the sum, that is, let's observe what appears on the two dice faces (aka, the sample space, which we will learn more about later).

Notice in the figure below, that it is more *likely* for the sum to be 4 than to be 12, as there is an inherent property in this experiment that some outcomes are more **frequent** than others. In this case, we could say that it is *three times more likely* to get a sum of 4 than a sum of 12.

If we go back to the histogram, you'll see that this idea of frequency, in the form of *relative* frequency, is closely related to the intuitive meaning of probability in the sense of being *proportional to the whole*.

The histogram, in this case, conveys the idea of the *relative frequency* with which the results of the experiment appear and gives us an indication that **the more we roll the dice, the easier it becomes to see the pattern of chance regularity** that seems to underlie the explanation of the concept of probability. This is **at the heart** of the definition of Probability, as we will see in a bit.

However, although the histogram helps us understand the observed stochastic phenomenon of chance regularity, it does not formalize the concept of probability. After all, we can only see the counts of equal results after several trials of an experiment. To understand this, let's start with the table below.

	1	2	3	4	5	6
1	(1,1)	(1,2)	(1,3)	(1,4)	(1,5)	(1,6)
2	(2,1)	(2,2)	(2,3)	(2,4)	(2,5)	(2,6)
3	(3,1)	(3,2)	(3,3)	(3,4)	(3,5)	(3,6)
4	(4,1)	(4,2)	(4,3)	(4,4)	(4,5)	(4,6)
5	(5,1)	(5,2)	(5,3)	(5,4)	(5,5)	(5,6)
6	(6,1)	(6,2)	(6,3)	(6,4)	(6,5)	(6,6)

4

12

Figure 1.4: Table with all possible results of the dice faces when rolled together

In this table, it becomes clear that there is a possible connection between the histogram and Probability. It is precisely in this attempt to capture the chance of each outcome among all possible outcomes that the idea of evaluating how likely it is (**or how frequent it is**) to observe a certain outcome arises after **many** trials. The reason why *many* is in bold is because this is at the core of the definition. We will get there.

1.3 What is "Probability"?

Notice that if the possibility of getting a sum of 4 appears more often (i.e., with higher frequency) than a sum of 12, it is very likely that the probability of the sum being 4 is lower than that of the sum being 12, as shown in red in the figure below. This comparison is made for all possible outcomes when we roll the dice numerous times and observe the pattern that emerges when we count the obtained results and plot them in a histogram-like graph.

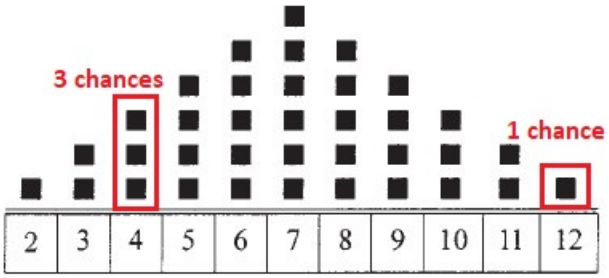

Figure 1.5: Graph that attempts to capture the idea of the probability distribution of our experiment

Intuition - Probability and its long run behavior

The concept of **Probability** is closely linked to the idea of the **relative frequency** with which the results of a certain experiment occur as we repeat that experiment **many** times, group the identical results and count them, and finally compare their proportions in all the experiment trials.

To be more precise, the reality is that the essence of Probability is not only embedded in the *frequency* with which identical results appear after several trials of an experiment but it is also related to a certain physical structure behind the experiment, which we could refer to as the natural *symmetry* or *geometry* of the problem in question.

In the case of rolling two dice, the geometry of the dice is a good indicator of why we see the results we see in the table above, i.e., it is a very good reason to argue why we observe that some sums occur more frequently than others.

The last important point in this section is that our objective is to present the path that mathematical modeling has taken to model such stochastic situations, starting from the observation of stochastic phenomena, capturing their *statistical regularity*) with the concept of probability (and our friend, the probability distribution) and formalizing the whole experiment with the **Statistical Model** based on Probability Theory (which we introduced on the "Real Motivation" page at the beginning of this course).

It is within this formalization that the concepts of *Probability, Probability Distribution, Probability Space, Random Variable, Independence, Probabilistic Model, Parameters, Estimation, Specification, Identification*, and so on, emerge.

This formalization via statistical models (including one of its parts, the probabilistic models) is done within a *Modeling Framework*.

This (modeling) *framework* is known as **Probability Theory**, a subject taught in mathematics, statistics, and physics, but which is generally not covered in depth even in Engineering, where it is usually called Probability and Statistics, which is presented in a more condensed form, which ends up leaving aside the essence that we present here and contributes to the motivation that I described at the beginning: I had to go back from scratch and go deeper mathematically in order to master the subject.

Probability theory is a modeling framework that aims to mathematically formalize the *statistical regularity*, a characteristic present in the mechanisms that constitute the "essence" of *stochastic phenomena*. We'll get there.

And that's what we will learn in this course :)

Note: Frequentists vs. Bayesians
I won't go into detail about the different **probability interpretations**[1] in the world of statistical scholars. However, it is worth noting that it is precisely because of the relationship between *Probability* and *Frequency* that we say that we are in the realm of *Frequent Probability*, also known as *Classical Probability*. The latter seems to be even more connected to the natural symmetry (or geometry) of a problem, an idea that came first in history. The other side of the story is the Bayesians, but it's too early to delve into that. For now, let's **finally** stick to the definition of the concept of Probability. See below.

[1] See Wikipedia

Finally, let's finally define Probability, within the world of Frequentists/Classics.

Definition: Long-Run Relative Frequency (aka Probability!)

The **Probability** of an event is defined as the *limit* of the relative frequency of its outcomes in a very large number of hypothetical trials.

The **limiting** idea refers to a context where we look at the probability of an event **AS IF** we were to repeat the experiment many times, hypothetically. This is the Frequentist (or Classical) Interpretation of Probability.

Just to look at an example, see the figure below, where a coin was tossed

many times and the relative frequency of Heads was evaluated so that we could observe what its probability could be. See the picture below. [2]

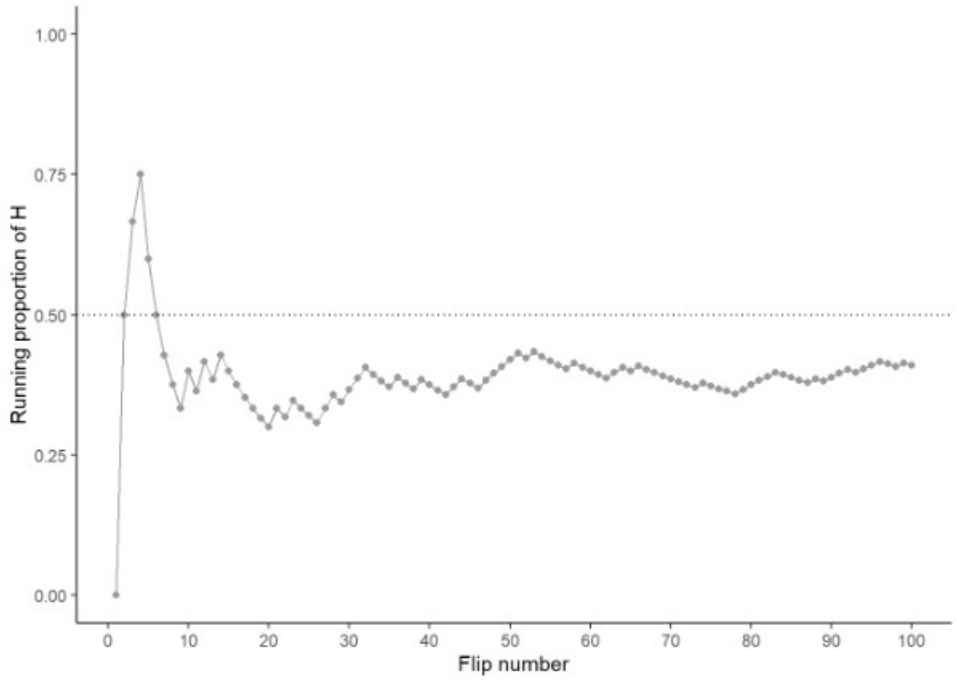

Figure 1.6: Coin tossed many times and its running proportion of Heads

In the picture above, one can notice that as we flip the coin more and more, the *limiting* behavior of getting Heads approaches 0.5 or 50% or $\frac{1}{2}$.

Okay, now you've learned what the concept of Probability is and noticed that the Probability Distribution came out of the statistical regularity present in the stochastic phenomena that the statistician wants to tame.

The idea now is to understand how the statistician will train these phenomena and for that, we will take another step towards the **Statistical Model**. This next step will be to understand what a **Probability Model** is.

[2] See Source

Chapter 2

Probabilistic Models

2.1 Formalizing the Probability Model

Bibliography

[1] Aris Spanos. *Probability Theory and Statistical Inference*. Wiley, 2003.