Urns and Trees: A Minimalist Guide to Probability for Frequentist Statistics & Machine Learning

Hanti Lin
University of California, Davis
ika@ucdavis.edu

February 20, 2026

Rough Draft. Please Do Not Cite.

Abstract

I'd like to present to you a short path through the garden of probability theory—leading to a place where you can begin to grasp the core ideas of frequentist statistics and machine learning. I hope this path will spark your interest and inspire you to explore other parts of the garden as well. For pedagogical purposes, each section introduces only one key idea and has exactly one page.

Contents

1 Motivation from Frequentist Hypothesis Testing

Let me walk you through an example of statistical inference and do some philosophy.

1.1 An Empirical Problem, and an Inference Method

Suppose that an urn contains exactly three balls, black or white, and not all of them have the same color. We ask: What is the proportion of black balls, written θ? Since not all balls have the same color, there are only two potential answers—or competing hypotheses:

$$\begin{aligned} H &: \quad \theta = 1/3 \\ H' &: \quad \theta = 2/3 \end{aligned}$$

To figure out the true answer, we are going to shake the urn well, draw a ball from it, observe its color, and put it back—and repeat this $n = 3$ times. Sorry, only three times (at least for now). The above specifies a particular empirical problem.

Now, how to make an inference based on observations? Let's consider this inference method:

An Example of an Inference Method: M^*

- Infer that $\theta = 1/3$ if the observed relative frequency of getting a black ball is closer to $1/3$ than to $2/3$—that is, if the observed relative frequency is less than one half.
- Infer that $\theta = 2/3$ if the opposite holds.
- Suspend judgment otherwise.

We want to assess this inference method by examining its *reliability* for obtaining the true hypothesis—or, to be more precise, its *varying* reliability across the relevant possible scenarios. There are two (coarse-grained) possible worlds: the world in which $\theta = 1/3$, and the world in which $\theta = 2/3$. Let's consider them in turn.

1.2 Reliability in a Possibility World

Consider *Scenario A*, in which the number of observations is $n = 3$, and the proportion of black balls is $\theta = 1/3$, so there are one black ball and two white balls in the urn. The possible observational outcomes are sequences of balls of length $n = 3$, represented as the branches of the tree in Figure 1. A black edge between two nodes means

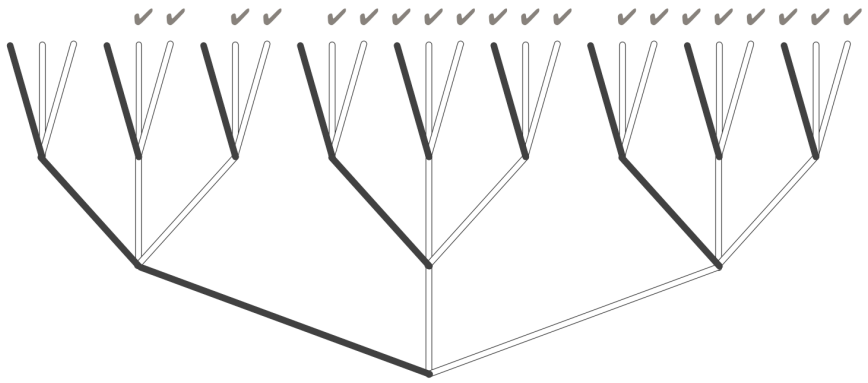


Figure 1: In Scenario A, there are 27 possible branches (or data sequences), among which the 20 checkmarked ones are the inputs that would prompt method M^* to output the true hypothesis.

getting a black ball; a white edge, a white ball. Each node branches out into three directions, corresponding to the three colored balls in the urn: black, white, and white, respectively.

Enter probabilities. Since we always shake the urn well before drawing a ball, those outcomes, or data sequences of length $n = 3$, have equal physical probabilities. Of those 27 data sequences, 20 of them have a less-than-one-half relative frequency of getting a black ball—as indicated by the 20 checkmarks in the figure—and those data sequences are exactly the inputs that would prompt the inference method M^* to output (or infer) the hypothesis $\theta = 1/3$—the true hypothesis in this scenario. So, in this scenario, the physical probability that M^* outputs the true hypothesis is equal to $20/27 \approx 74\%$. This physical probability can then be used to define, or measure, the reliability of M^* in this scenario.

1.3 Clarification

A bit of housekeeping. First, at this point, you might be wondering what physical probability is, or how the term 'physical probability' should be interpreted. This is an important metaphysical—or semantic—question. But hold that thought: we'll return to it shortly. For now, let's stay focused on the epistemological issues.

Second, it should be noted that, in general, the simple arithmetic technique of counting and division, as used here, is not enough to calculate the needed probabilities in many cases, such as when data sequences do not have equal probabilities or when the number of sequences is infinite (countably or uncountably). Don't worry—in those cases, probability theory will take care of it. The point of the present setting is to reveal the essence of frequentist statistics by stripping away unnecessary mathematics.

Third, if a 0.74 reliability—as the physical probability of getting the truth—is not good enough, we can increase the sample size to $n = 15$ and obtain a more-than-0.9 reliability for the same inference method M^*. But then there will be a tree that is much harder to draw, with $14,348,907$ $(= 3^{15})$ branches, of which $13,082,880$ ones are marked with a checkmark, so that $\frac{13,082,880}{14,348,907} \approx 0.91$.

1.4 How Reliability Varies across Possible Scenarios

We have calculated the reliability of inference method M^* in Scenario A, and can easily do it for the other scenario—*Scenario B*—in which $\theta = 2/3$ and $n = 3$, using the very same arithmetic technique. Omitting the (entirely repetitive) details, let me just report the result: in Scenario B, the physical probability that M^* outputs the true hypothesis ($\theta = 2/3$) is also $20/27 \approx 0.74$.

Now we can plot how the reliability of M^* varies across possible scenarios, as depicted in Figure 2, in which I also include a reliability-scenario plot for another inference method, M, which is defined to (dogmatically) output the hypothesis $\theta = 1/3$ irrespective of the input.

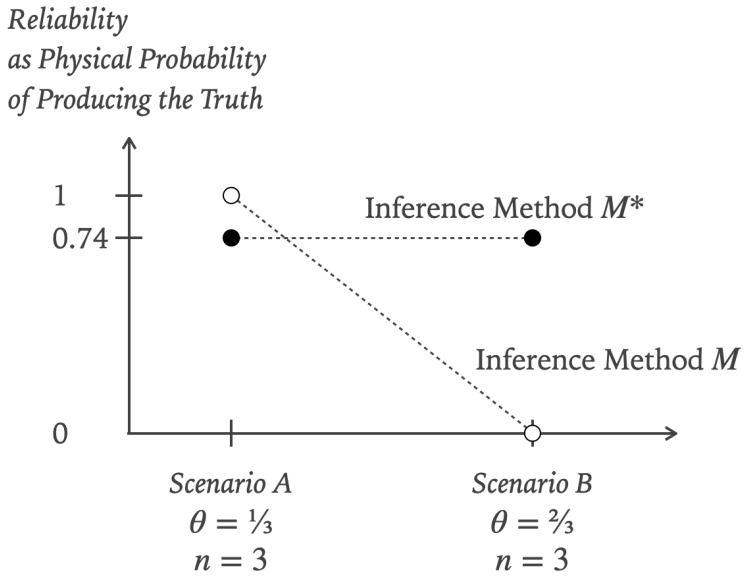


Figure 2: The reliability-scenario plots of two inference methods

Here is an epistemological question: in the context of the present empirical problem, is one of the two inference methods justified—and if so, which one? In frequentist statistics, this question is answered by stating and applying some evaluative standards, possibly also defending those standards and arguing against other standards.

1.5 Standards for Evaluating Inference Methods

The following provides some initial examples of evaluative standards, which are relatively easy to define:

Definitions (Guaranteed Reliability, Maximin Reliability)

- An inference method is said to *guarantee a reliability of at least* $1 - \alpha$ if and only if, in every scenario under consideration, its reliability is $1 - \alpha$ or higher, where α is a positive real number.
- An inference method is said to *have maximin reliability* if, and only if, its worst reliability across the scenarios under consideration is at least as high as the worst reliability of any other inference method.

As the underlines suggest, the first standard is defined by quantifying over possible scenarios; the second standard is more complex—it even quantifies over possible inference methods.

Frequentist statisticians have developed a lot more evaluative standards. Those standards are distinctively "frequentist":

What Are Frequentist Standards?

- A *frequentist standard* is defined as a condition on reliability-scenario plots—it rules out some reliability-scenario plots and, thereby, rules out the associated inference methods.
- The idea is that a frequentist standard is meant to assess inference methods by examining how their respective reliabilities vary across a range of scenarios.

But what range of scenarios? It may be the set of the scenarios under consideration, or those compatible with the background assumptions of one's context of inquiry.

1.6 From Evaluative Standards to Epistemological Theses

Something needs to be stated explicitly in order to go from mere definitions of evaluative standards to substantive judgments of justified inference. This point is particularly important for epistemologists: while statisticians always explicitly define evaluative standards (such as maximin and, as we will see very soon, low significance level), the intended epistemological judgments often need to be reconstructed from their practices.

Let me begin with some naïve examples. Here is a thesis that employs a clearly defined evaluative standard to make a judgment about justified inference:

> *Thesis 1.* The standard of maximin reliability is necessary for an inference method to be justified at least in the problem context at hand.

If we accept this thesis, we can conclude that the dogmatic method M is not justified. If, furthermore, we also accept the following:

> *Thesis 1'.* The standard of maximin reliability is sufficient for an inference method to be justified at least in the problem context at hand.

then we can conclude that inference method M^* is justified—provided that we can prove that M^* does achieve maximin reliability. If, instead, we accept the following

> *Thesis 2.* The standard of a guaranteed reliability of at least 0.9 is necessary for an inference method to be justified in the problem context at hand.

then, unfortunately, none of those two inference methods is justified. In this case, actually no inference method is justified with the fixed sample size $n = 3$, and we have to bump up the sample size to $n = 15$ or higher in order to make it possible to achieve a guaranteed reliability of at least 0.9.

1.7 The Frequentist Way of Thinking

The above are just some particularly simple (and naïve) examples of epistemological theses—ones that connect frequentist evaluative standards to justified inference. They are just potential, partial answers to a group of epistemological questions that characterize the distinctively frequentist way of thinking in statistics:

Distinctively Frequentist Questions

- Which frequentist standard—or what combination of such standards—counts as a necessary and/or sufficient condition for an inference method to be justified?
- And, justified under what scope? Justified universally across all problem contexts, or in a particular problem context (with a given set of competing hypotheses and possibly a fixed sample size)?

Frequentist statisticians—or perhaps frequentist philosophers of statistics—are largely united by the idea that thinking about those questions is key to justification of scientific inference, even though they often have disagreements on how to answer those questions.

I have been intentionally *not* presenting frequentist statistics in the standard way—even avoiding presenting the particular epistemological theses underlying the pioneers of frequentist statistics such as Fisher (1925), Neyman & Pearson (1933), and the 19th century scientist and philosopher C. S. Peirce. The reason is simple: while the details of their views differ dramatically, I want to focus on what is shared by, and even core to, their views.

1.8 Summary: The Structure of Key Concepts

Let me conclude this tutorial with a diagram that summarizes the dependencies among the concepts introduced above, as depicted in Figure 3.[1] Perhaps this conceptual net-

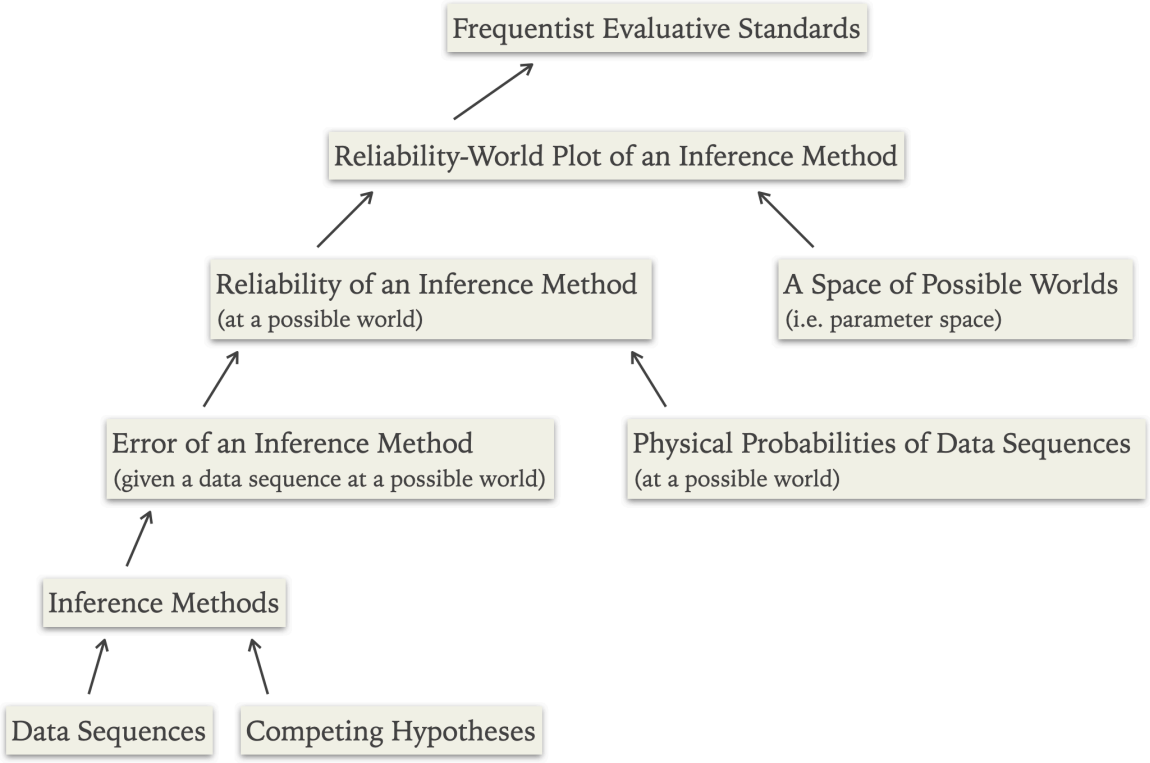


Figure 3: A roadmap of key concepts and their dependencies in frequentist statistics

work has never been drawn in any textbooks or research articles. But if I am right, it expresses a common ground shared by many frequentist statisticians and philosophers sympathetic to frequentist statistics. It is also a common ground that highlights places where disagreements may arise, giving rise to several issues. One such source of controversy is interpretational, concerning how 'physical probability' should be understood. As noted above, however, this is not our focus here. I will turn the spotlight to the top of this conceptual network: What frequentist standard should we use to evaluate inference methods?

[1]To clarify, this diagram intentionally omits some dependencies that are not important in the present context and that, if made explicit, would clutter the picture—such as the missing arrow from "Data Sequences" to "Physical Probabilities of Data Sequences".

2 Understanding Probability

I will not really define or analyze the concept of physical probabilities here. But it might be helpful to clarify this concept.

2.1 Interpretations of Probability: Physical, Subjective, or Epistemic

How to start understanding a concept? I guess it is not easy to find a satisfactory definition or analysis of the concept of dogs. But no worries—a good starting point for learning a concept is by way of examples, such as the dogs that our parents point us to. So let me give a paradigm example of a possible event that has, say, a 1/3 physical probability of occurrence.

> **Example (1/3-Urn with One Draw)**. Suppose that an urn contains exactly one black ball and two white ones. Ann is going to shake the urn well and draw a ball from it. There are three possible outcomes for Ann: getting a black ball, getting one of the two white balls, and getting the other white ball. Those three possible outcomes are equally probable. So, the possible outcome *Ann gets a black ball* has a physical probability of occurrence whose value is 1/3.

Clarifications are in order.

First of all, the 1/3 probability is *physical*, rather than *subjective* or *epistemic*. More specifically, this probability is not *subjective*, as it is independent of the belief that Ann actually has. Moreover, this probability is not *epistemic*, as it is independent of the belief that Ann should have, may have, or is justified in having. Instead, this probability is physical in that it exists objectively in the physical world (or the possible physical world represented by this scenario), and its value depends solely on certain physical facts in this scenario—the makeup of the physical system and the physical process that leads to the occurrence, or failure of occurrence, of the event.

2.2 Measuring a Rational-Valued Probability

Now, compare the following two cases:

- *Case 1: A Coin Model.* There is a coin whose bias, or probability of landing heads, is $\theta = 1/3$, and we are going to toss it just once.
- *Case 2: An Urn Model.* There is an urn with 3 balls, of which exactly 1 is black, and we are going to draw a ball after shaking the urn well, so that all the three possible outcomes are equally probable.

Case 2 can be used as a model of Case 1: the event of drawing a black ball from the urn can be used to model the event of the coin landing heads. Then Case 2 provides a qualitative yardstick that we can use—by counting—to measure the numerical value of a certain probability in Case 1: $\theta = 1/3$. This explains how a *numerical* quantity may arise from counting objects that hold a certain *qualitative* (equivalence) relation, such as "equally long", "equally probable", or the like. That said, there remains a metaphysical question:

> Is it that the coin in Case 1 can be accurately modeled by the urn model in Case 2 *because* the probability in Case 1 has a numerical value $1/3$? Or is it the other way round, so that the probability in the coin in Case 1 has a numerical value $1/3$ *because* the coin in Case 1 can be accurately modeled, and thus measured, by the urn model in Case 2?

Although I will leave this metaphysical issue unanswered here, those sympathetic to measurement theory tend to think that it is the latter.

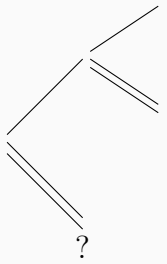
2.3 Merging Branches (Forming a Marginal Distribution)

Extend the above comparison as follows:

- *Case 1: A Coin Model.* There is a coin whose bias, or probability of landing heads, is $\theta = 1/3$, and we are going to toss it n times. So there is a binary tree, with 2^n data sequences, over which probabilities are distributed non-uniformly.
- *Case 2: An Urn Model.* There is an urn with 3 balls, of which exactly 1 is black, and we are going to draw a ball n times—always after shaking the urn well, and always with replacement. So there is a ternary tree, with 3^n data sequences, over which probabilities are distributed uniformly.

Case 2, again, can be used to accurately model Case 1. This can be done by *merging* some of the branches of the ternary tree in Case 2 to form a binary tree—a duplicate of the binary tree in Case 1. Now, please design a diagram with $n = 2$ and explain to yourself how you would merge branches by using the following hint:

After merging, the ternary tree for the urn model becomes:



The idea is that some branches as possibilities can be merged to form a single branch as a *more coarse-grained* possibility.

Now consider the sequence $s = \texttt{01}$ with $n = 2$, where `1` means landing heads, corresponding to drawing a black ball, and `0` means landing tails, corresponding to drawing a white ball. Once the above diagram is completed, the probability of each coarse-grained branch, say sequence $s = \texttt{01}$ can be identified with the proportion of (equally probable) fine-grained branches that get merged into the coarse-grained $s = \texttt{01}$. It is 2 out of 9. This probability can also be written $\left(\frac{2}{3}\right)\left(\frac{1}{3}\right)$, but why? The ternary tree of the urn model has many (equally probable) branches, of which $\left(\frac{2}{3}\right)$ get merged into sequences of the form `1x`, of which $\left(\frac{1}{3}\right)$ get merged into the sequence $s = \texttt{10}$. So, this sequence has a probability $\left(\frac{2}{3}\right)\left(\frac{1}{3}\right)$. Think about how this formula can be generalized to $\left(\frac{a}{b}\right)^k \left(1 - \frac{a}{b}\right)^{n-k}$ for an urn with many balls to be drawn repeatedly.

2.4 Measuring an Irrational-Valued Probability

Now, why does this physical probability have this particular value 1/3? Can this be explained by something more fundamental, such as the qualitative, non-numerical relation of *being equally probable*, and *being more probable*? Yes, in this scenario, each of the three outcomes is as probable as the others, which suggests, implies, explains, or constitutes the fact that each has a 1/3 probability of occurrence. This might help make the numerical value 1/3 less mysterious than it appears to be.

Now, think about this numerical claim: "The physical probability of an event E (say, a coin's landing heads) is equal to $\frac{1}{\sqrt{2}}$, whether or not E is repeatable." Since E might be unrepeatable (possibly because the coin will be destroyed soon after), I suggest that this numerical claim be understood as follows.

> The physical probability of an event E is equal to, say, $\frac{1}{\sqrt{2}}$ $(= 0.7071\cdots)$ iff the following holds (where an $^{m}/_{n}$-urn is a well-shaken urn in which there are n balls in total, m of which are black):
>
> - E is more probable than drawing a black ball from a $^{7}/_{10}$-urn;
> - E is more probable than drawing a black ball from a $^{70}/_{100}$-urn;
> - E is more probable than drawing a black ball from a $^{707}/_{1000}$-urn;
> - E is more probable than drawing a black ball from a $^{7071}/_{10000}$-urn;
>
> $\vdots$
>
> - for any fraction $^{m}/_{n}$ in the above,
>
> E is less probable than drawing a black ball from an $^{m+1}/_{n}$-urn.

That is, the quantitative aspect of physical probability might be reducible to the qualitative relations of *being more probable* and of *being less probable* together with (hypothetical, merely counterfactual?) urn models.

Note that this is at most just a metaphysical account—an account of the nature of the numerical values of physical probabilities. It does not offer anything epistemological; in particular, it does not tell us how to measure or estimate those numbers.

Two methodologies are used above to understand physical probabilities. Let me explain them in turn.

2.5 Methodology I: Understanding Quantities by Qualities and Counting

What does it mean that the bias of a coin—its probability of landing heads—equals 1/3? It cannot mean that, if the coin were to be tossed many times, the frequency of landing heads *would* be equal, or close, to 1/3. For this frequency *could* deviate significantly from 1/3—just with a small but *nonzero* probability that is less than or equal to $2\,e^{-2n\epsilon^2}$ (according to Hoeffding's inequality).

Fortunately, there is another way to understand physical probabilities. On the propensity account, the physical probability of a possible outcome of a stochastic process is a number that measures the strength of the tendency, disposition, or propensity that the process possesses for producing that outcome (rather than any other possible outcomes). Metaphysicians have widely discussed the nature of tendencies, dispositions, or propensities, and I have nothing interesting to say about that here. Instead, my focus here is placed on where the number comes from—why the strength of a propensity of getting a particular outcome is measured by this rather than that particular number.

Here is the guiding principle in use, borrowed from representational measurement theory (not to be confused with measure theory):

> Understand quantities in terms of (i) qualities and (ii) counting.

For example, a rod r is 1/3 meters long just in case, if three additional rods each as long as r were concatenated to form a longer rod, then this longer rod would be exactly as long as the chosen unit, 1 meter. The quality in question is a binary relation, "as long as", and we count three in this case.

2.6 Methodology II: Understanding by Modeling

Instead of defining the concept of physical probabilities and the concept of stochastic processes, I would like to propose that they be understood with the methodology of modeling—following a standard methodology in natural sciences: *understanding a target by successive approximate modeling*.

An analogy might be helpful. Physicists do not define all the concepts that concern the hydrogen atom. Instead, they build progressively refined models to better understand the hydrogen atom:

- Bohr's model introduced quantized circular orbits, which Sommerfeld refined by allowing elliptical motion corrections.
- Schrödinger's wave model replaced orbits with probability clouds, and Pauli's refined model added electron spin and spin-orbit coupling. Within non-relativistic quantum mechanics, further refined models incorporated the leading-order relativistic correction to the electron's kinetic energy, yielding a more accurate model of fine structure.
- Dirac's relativistic model unified quantum mechanics and special relativity, predicting fine structure and antimatter from first principles.
- But only the refined models of quantum electrodynamics could model precision effects like the Lamb shift and the anomalous magnetic moment.
- Today's most refined models incorporate nuclear recoil, proton structure, and hyperfine interactions.

This demonstrates how the hydrogen atom is progressively better understood through a hierarchy of increasingly precise and conceptually unified models. What I did above is just adapting this methodology to understanding distributions (or measures) of physical probabilities.

In an earlier example, we imagined a physical process that has a probability of $\frac{1}{\sqrt{2}}$ to produce event E. I propose that this aspect of this physical process be understood by modeling it with progressively more accurate urn models, which do not represent other aspects of the physical process. Something like this might be the best we can do to understand probabilities and many things in the world—or do you have a better way to understand the hydrogen atom?

2.7 More on Urn Models

Urn models are used here to provide some paradigm examples of physical probabilities. We might think that the above descriptions of urn models are too sketchy and need additional details, such as these: 'There will be no peeking when drawing a ball.' 'The balls are equal in size.' We might need more examples to gain enough understanding before we can proceed to learn or develop a theory that employs this concept. And when the task of theory development is pursued to a certain point, we might need to revisit the stock of paradigm examples and revise it. And we might, or might not, eventually think that it is too classical to capture physical probabilities and needs to be replaced by a quantum mechanical example. So, what we take as paradigm examples might be revised, expanded, or replaced as we go along—as is often the case in an *ongoing* process of clarifying a concept.

2.8 Frequentist Statistics Possibly Without Frequencies

At this point, you might be wondering what physical probabilities are, or how the term 'physical probability' should be interpreted. This is an important metaphysical—or semantic—question. Perhaps physical probabilities (quantitative or qualitative) are one of these:

- long-run (hypothetical) frequencies (Neyman 1955),
- propensities (tendencies, dispositions) (Popper 1959),
- Humean chances (Lewis 1987)
- primitive physical states posited in science (Sober 2000: sec. 3.2).

For focus, I will set aside this metaphysical/semantic debate and concentrate on epistemological issues. Suffice it to note the following:

> The probabilities used in *frequentist* statistics might—or might not—be best understood as long-run *frequencies*, as there are notable alternatives already developed in philosophy. The term 'frequentist' in 'frequentist statistics' is therefore quite misleading—but unfortunately too entrenched to change.

3 Probability Theory

I want to give you a shortest path to a tool we will use very often: *Hoeffding's inequality.*

3.1 Hoeffding's Inequality for Urn Models

Consider an urn containing b balls in total, of which exactly a are black—the total number of colors does not matter for now. We are going to draw a ball n times with replacement (*and always after a good shake*). So, the possible outcomes are sequences of balls of length n—there are b^n possible sequences (*and they are equally probable*).

Now, of the possible data sequences of length n, what is the proportion of those whose relative frequency of black balls is a "good" estimate of the true proportion of black balls in the urn—differing from the true proportion a/b by less than ϵ? (*That is, under the assumption of always giving the urn a good shake, what's the probability of obtaining such a "good" sequence?*)

Here is a partial answer: the proportion of such "good" sequences is large as long as n, the number of draws, is large. This partial answer can be stated more precisely as follows, where $\mathbb{P}\big(|X_{\text{freq}} - a/b| < \epsilon\big)$ denotes the proportion in question—and you should think about why the notation is designed that way:

$$\mathbb{P}\left(\left|X_{\text{freq}} - \frac{a}{b}\right| < \epsilon\right) \quad \geq \quad 1 - 2\,e^{-2n\epsilon^2}$$

Without the parenthesized clauses in the above, this is a purely combinatorial result in arithmetic and algebra—just a matter of counting. When the parenthesized clauses are taken seriously, this is an important theorem in probability theory.

3.2 Different Versions

The inequality presented above has some variants. One variant is entirely equivalent, just switching the attention from a large probability of something good to a small probability of something bad:

$$\mathbb{P}\left(\left|X_{\text{freq}} - \frac{a}{b}\right| \geq \epsilon\right) \quad \leq \quad 2\,e^{-2n\epsilon^2}$$

Another variant is strictly weaker, replacing the exponential decay $2\,e^{-2n\epsilon^2}$ by a much slower decay $1/(4n\epsilon^2)$, yet the proof technique goes from calculus to mere algebra

$$\mathbb{P}\left(\left|X_{\text{freq}} - \frac{a}{b}\right| \geq \epsilon\right) \quad \leq \quad \frac{1}{4n\epsilon^2}$$

We will see another version of Hoeffding's inequality below—not a mere reformulation or a weakening, but a truly probabilistic generalization. In fact, the rest of this tutorial section on probability is mostly about giving you just enough background to understand that version of Hoeffding's inequality.

3.3 Coin Models

In addition to urn models, there is another classic type of probability model: coin models. Consider a coin whose probability of landing heads equals a certain fixed real number θ. Suppose that coin tosses are, in a sense, independent. We are to toss this coin $n = 5$ times.

Now, what is the probability of obtaining a data sequence $s = \texttt{01101}$, where `1` means landing heads and `0` means tails? Answer: this probability equals $\theta^3(1-\theta)^2$.

Then, what is the probability of obtaining that data sequence $s = \texttt{01101}$ or any other one at which the frequency of heads is equal to $3/5$? Answer: $C_3^5\,\theta^3(1-\theta)^2$, namely $\frac{5!}{3!\,2!}\,\theta^3(1-\theta)^2$.

Suppose that we have a coin model with a certain bias, say, $\theta = \frac{1}{\sqrt{2}} = 0.7071\cdots$. While a coin model can be used to represent some aspects of many physical systems, it can be better understood in light of some simpler models, such as urn models, that successively approximate the coin model in question—in the way suggested in Sections 2.4-2.6.

The main theorem in probability theory to be presented below, Hoeffding's Inequality, will be stated for coin models. But before that, we need more preliminary concepts.

3.4 Discrete Probability Distributions

In the examples above, probabilities are distributed over the branches of a tree as data sequences. They are examples of this concept:

Definition (Discrete Probability Distribution). A *discrete probability distribution* is a function that assigns real numbers—nonnegative and summing to 1—to a countable set of possible outcomes $x_0, x_1, \ldots, x_i, \ldots$ (such as data points or data sequences); put formally, it is a function p taking this form:

- $p(\,\cdot\,) : \{x_0, x_1, \ldots, x_i, \ldots\} \to \mathbb{R}$,
- $p(x_i) \geq 0$ for each i.
- $\sum_i p(x_i) = 1$.

In case you are wondering what countability means: a set S is called *countable* iff the elements of S can be enumerated by using just a finite set of natural numbers or using all natural numbers. In the latter case, S is said to be *countably infinite*.

Exercise 1: Think about how a discrete probability distribution can be approximated to high accuracy by an urn model: drawing a ball just once from an urn that has enough balls with enough colors. Hint 1: It is like using rational numbers to approximate any given real number with any given desired accuracy. Hint 2: Each x_i corresponds to a unique color. Hint 3: Sorry, an urn model has only finitely many balls by definition—this is the difficult part.

3.5 Constructing (Some) Probability Measures

Given a discrete probability distribution p that assigns probabilities (numbers) to the possible outcomes in Ω, it can be used to construct more stuff as follows:

- An *event* E is formalized as a set of possible outcomes taken from $\Omega = \{x_0, x_1, \ldots\}$. I.e. events are formally subsets of Ω.
  E.g. if $E = \{x_3, x_7\}$, E denotes the event in which the outcome we get is x_3 or x_7.
- While p only assigns probabilities to outcomes, we can extend the assignment from outcomes to events and talk about the probability of an event E, written $\mathbb{P}(E)$.

The value of $\mathbb{P}(E)$ is determined by p according to the following formula:

$$\mathbb{P}(E) \quad = \quad \sum_{x \in E} p(x)\,.$$

This $\mathbb{P}(\,\cdot\,)$ is called the *probability measure* constructed from p.

Intuitive understanding (the blue part below) can then be reinforced by formal reasoning (the red part below):

The probability that the outcome we get is either x_3 or x_7

$=$ $\mathbb{P}($ the outcome we get is either x_3 or x_7 $)$

$=$ $\mathbb{P}(E)$, where $E = \{x_3, x_7\}$

$= \sum_{x \in \{x_3, x_7\}} p(x)$

$= p(x_3) + p(x_7)$

$=$ the probability that we get x_3 + the probability that we get x_7

3.6 Variables

The victors of World War II could have been the Allies, and could have been the Axis. That is:

1. There are at least two possible outcomes of World War II.
2. 'The victors of World War II' is a variable, which can take different values in different situations—i.e. at different outcomes.
3. The value of 'the victors of World War II' is the Allies at one possible outcome (the actual one).
4. The value of 'the victors of World War II' is the Axis at another possible outcome (a counterfactual one).

Now, replace historical outcomes by outcomes in the coin toss experiment with $n = 5$:

1′. There are at least two possible outcomes of the coin toss experiment, such as `11100` and `01000`.

2′. X, which is short for 'the frequency of `1`s', is a variable, which can take different values at different possible outcomes.

3′. The value of X ('the frequency of `1`s') is $3/5$ at the first possible outcome `11100`.

4′. The value of X ('the frequency of `1`s') is $1/5$ at the second possible outcome `01000`.

Note the parallelism between 1-4 and 1′-4′. This gives you an idea of what variables are as commonly used in natural and social sciences, but *not* in pure mathematics.

> What's a variable as used in natural and social sciences, rather than pure mathematics? Roughly and intuitively speaking, such a variable can be expressed by an expression taking this form:
>
> *the so-and-so*.
>
> This is called a definite description in linguistics.

3.7 Random Variables

A random variable is, well, a variable that is random.

- A *variable X—in real applications*—can be intuitively expressed by a definite description 'the so-and-so' (as seen above), which can take different values or refer to different objects at different possible outcomes.
- Variable X is called a *random variable* when we wish to emphasize that the actual outcome is generated through a stochastic process and, thus, the actual value of X is determined stochastically.

A bit of notation for using random variables:

- '$X(\omega) = x$' means that the value of X at an outcome ω is equal to x. This is why—*in formal treatment*—a random variable X is defined as a function that maps each possible outcome ω in Ω to a value $X(\omega)$.
- '$X = x$' means that we (turn out to) obtain an outcome at which the value of X is equal to x.

Make sure that you can make sensible use of the language of probability theory. For example:

- $\mathbb{P}\big(X = x\big)$ makes sense.
- $\mathbb{P}\Big(\{\omega \in \Omega : X(\omega) = x\}\Big)$ makes sense, too, and means the same as the previous one.
- $\mathbb{P}\big(X(\omega) = x\big)$ makes no sense.

You will see more examples below.

3.8 Calculating Probabilities with Random Variables

Whenever you see a probabilistic expression $\mathbb{P}(\ldots)$, look for the element whose value is stochastically generated, and make it explicit. Once you have identified it as a random variable X, the meaning of this probability $\mathbb{P}\big(\ldots X \ldots\big)$ is transparent and its value may be calculated:

$\mathbb{P}\big(\ldots X \ldots\big)$

$= \mathbb{P}\big($the value of random variable X satisfies a certain condition $c\big)$

$= \mathbb{P}\big($we get one of the outcomes at which the value of X satisfies condition $c\big)$

$\overset{!!!}{=}$ the sum of the probabilities of the "marked" outcomes (or data sequences), where an outcome ω is "marked" iff the value of X at ω satisfies condition c.

The last of the three equalities, as flagged with exclamation marks, represents a key calculation method. It is exactly this equality that gets formalized in probability theory, as demonstrated in the red part below:

$\mathbb{P}\big(\ldots$ random variable $X \ldots\big)$

$= \mathbb{P}\big($the value of random variable X satisfies a certain condition $c\big)$

$= \mathbb{P}\big($we get one of the outcomes at which the value of X satisfies condition $c\big)$

$= \mathbb{P}\big($the set of the outcomes at which the value of X satisfies condition $c\big)$
i.e. $\mathbb{P}\big($the event that the value of X satisfies condition $c\big)$

$= \mathbb{P}\Big(\{\omega \in \Omega : X(\omega)$ satisfies condition $c\}\Big)$
where $X(\omega)$ denotes the value of X at outcome ω

$= \sum_{\omega \in \Omega:\ X(\omega) \text{ satisfies condition } c} p(\omega)$
where p is the probability distribution used to construct $\mathbb{P}$

$=$ the sum of the probabilities of the "marked" outcomes (or data sequences), where an outcome ω is "marked" iff the value of X at ω satisfies condition c.

Note, again, how formal reasoning (the red part) reinforces intuitive understanding (the blue part).

3.9 Hoeffding's Inequality for Coin Models

While a very simple version of Hoeffding's inequality has been presented above in terms of urn models, here is a more general version:

Let $\mathbb{P}$ be a probability measure constructed from a probability distribution over the tree of binary data sequences with length n. For each $i = 1, \ldots, n$, X_i is a binary random variable understood as follows: $X_i = \texttt{1}$ means that the i-th datum obtained is `1`; similarly for $X_i = \texttt{0}$. Suppose that the *IID* assumption holds (so that this is basically the case of n independent tosses of a coin that has a fixed bias):

- (*Independence*) Those n random variables are independent, in the sense that $\mathbb{P}(X_1 = x_1, \ldots, X_n = x_n) \;=\; \mathbb{P}(X_1 = x_1) \cdot \ldots \cdot \mathbb{P}(X_n = x_n)$.
- (*Identical Distribution*) The probabilities of getting a datum 1 are identical to each other and equal to, say, θ; namely, $\mathbb{P}(X_i = 1) = \theta$ for each i.

Since the exact details of this $\mathbb{P}$ depend on parameter θ and sample size n, let's make this explicit by rewriting $\mathbb{P}$ as $\mathbb{P}^n_\theta$. Then we have:

$$\mathbb{P}^n_\theta \left(\left| \frac{1}{n}\big(X_1 + \cdots + X_n\big) \;-\; \theta \right| \;<\; \epsilon \right) \quad \geq \quad 1 - 2\, e^{-2n\epsilon^2}.$$

Exercise 2: Interpret the $\mathbb{P}^n_\theta(\ldots)$ above using the ideas from the previous subsection. Make sure that you identify the random variable and imagine marking some branches of the tree.

Exercise 3: Interpret $\mathbb{P}^n_\theta\,(X_1 + \cdots + X_n \;=\; x)$ using the ideas from the previous section. Use your interpretation to argue, intuitively, that $\mathbb{P}^n_\theta\,(X_1 + \cdots + X_n \;=\; x) \;=\; \frac{n!}{x!\,(n-x)!}\,\theta^x\,(1-\theta)^{n-x}$, which is known as the *binomial* distribution (over $n+1$ coarse-grained possible outcomes, as x can be 0, 1, ..., or n).

3.10 Excursion: Understanding by Modeling

Recall a methodology explained in Section 2.6: understanding a target by successive approximate modeling.

Exercise 4: Now, try to understand the stochastic process underlying the probability measure $\mathbb{P}_\theta^n(\ldots)$ on the previous page and make it less mysterious—by letting it be approximately represented by an urn model with a large number of balls and two colors. The crux is to represent the event of landing heads by the event of drawing a black ball, and tails by a white ball, and then use the urn-model version of Hoeffding's inequality, which was described above in purely arithmetic, non-probabilistic terms.

3.11 Construction vs. Axiomatization?

Any function $\mathbb{P}(\,\cdot\,)$ constructed from a discrete probability distribution actually falls under a very big category: it is a type of function called *probability measure*—a function satisfying the so-called *axioms of probability* (more on them soon).

For now, suffice it to know the following:

- We need the definition of probability measures in its full generality only when discrete probability distributions (and even continuous ones) cannot be used to generate enough probability measures for our use. Thankfully, this does not often happen in this course.
- That said, the notation of $\mathbb{P}(\,\cdot\,)$ will be used very often.
- Any theorems about probabilities we will use here can be derived from the axioms of probability along with appropriate definitions of relevant concepts.
- It is an oversimplification to say that mathematics proceeds from axioms and definitions, although some mathematics textbooks have been written to create that impression. A more accurate picture, albeit still simplified, is this: people—including the best mathematicians of all time—do mathematics, then formalize their practice by some axioms and definitions, and then do more mathematics, and the new practice gets formalized, too—and repeat. So let's begin with *doing* mathematics.

3.12 Axioms of Probability

Well, take a quick look at the definition of probability measures, although you don't really need to understand the full details (yet):

Definition (Probability Measure). A *probability measure* is a function $\mathbb{P}(\,\cdot\,)$ that assigns real numbers to some (possibly not all) events as sets of possible outcomes in a nonempty space Ω, such that the following conditions—known as the *axioms of probability*—are satisfied:

- **Axiom 0 (Sigma-Algebra)** The set of the events that $\mathbb{P}(\,\cdot\,)$ assigns numbers to, written $\mathcal{A}$, is a so-called *sigma-algebra* over Ω—namely, $\mathcal{A}$ contains Ω and $\varnothing$, and is closed under complement and countable union.

- **Axiom 1 (Nonnegativity)**
  $\mathbb{P}(E) \geq 0$ for any E in $\mathcal{A}$.

  Axiom 2 (Normalization)
  $\mathbb{P}(E) = 1$ if $E = \Omega$, the event that must happen.

- **Axiom 3 (Countable Additivity)**
  $\mathbb{P}(E_0 \cup E_1 \cup \ldots) \;=\; \mathbb{P}(E_0) + \mathbb{P}(E_1) + \ldots,$
  for any sequence of mutually disjoint events $E_0, E_1, \ldots$ in $\mathcal{A}$.

Let's not worry too much about Axiom 0 if you are not familiar with such terms as 'closed' or 'countable union'. For any function $\mathbb{P}(\,\cdot\,)$ constructed from a discrete probability function p, I can assure you that this $\mathcal{A}$ does satisfy the requirement of Axiom 0; in fact this $\mathcal{A}$ turns out to be the set of *all* subsets of Ω.

Exercise 5: Prove that any $\mathbb{P}(\,\cdot\,)$ constructed from a discrete probability distribution p satisfies Axioms 1-3, so that it is indeed a probability measure.

3.13 Some Derived Rules in Probability Theory

Here you go:

1. **The Law of Total Probability** (Simple Version)
   $\mathbb{P}(A) \;=\; \mathbb{P}(A \cap B) + \mathbb{P}(A \cap \overline{B})$

2. **The Law of Total Probability** (Finite Version)
   If $\{B_1, \ldots, B_k\}$ is a finite partition of Ω,
   then $\mathbb{P}(A) \;=\; \mathbb{P}(A \cap B_1) + \cdots + \mathbb{P}(A \cap B_k)$

3. **The Complement Principle**
   $\mathbb{P}(\overline{A}) \;=\; 1 - \mathbb{P}(A)$

4. **The Inclusion-Exclusion Principle**
   $\mathbb{P}(A \cup B) \;=\; \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$

5. **The Union Bound**
   $\mathbb{P}(A \cup B) \;\leq\; \mathbb{P}(A) + \mathbb{P}(B)$

6. **The Intersection Bound**
   If $\mathbb{P}(A) \geq 1 - \delta_1$ and $\mathbb{P}(B) \geq 1 - \delta_2$,
   then $\mathbb{P}(A \cap B) \geq 1 - (\delta_1 + \delta_2)$

Exercise 6: Begin with using countable additivity to prove 1, which gives you a hint for proving 2. Then use countable additivity and normalization to prove 3. Then use some axioms of probability and whatever rules you have derived to prove 4, which helps you prove 5, which in turn helps you prove 6. Hint: always try to draw some Venn diagrams to see whether they might help. Finally, guess what would be the full version of the Law of Total Probability, and prove it.

3.14 More Distributions: Discrete vs. Continuous

We have many types of probability models or distributions, from simple to complex:

- *Urn Model*, or *Urn-Based Probability Distribution*: a function $p(x)$ that assigns rational numbers—nonnegative and summing to one—to a finite set of outcomes/colors, which (as you may recall) can be used to construct a probability measure $\mathbb{P}(\,\cdot\,)$ as follows:

  $$\begin{aligned}&\mathbb{P}(\text{an event } E \text{ as the set of outcomes that would make } E \text{ happen})\\ &= \textstyle\sum_{x \in E} p(x)\end{aligned}$$

- *Discrete Probability Distribution*: a function $p(x)$ that assigns real numbers—nonnegative and summing to one—to a finite or countably infinite set of outcomes, which (as you may recall) can be used to construct a probability measure $\mathbb{P}(\,\cdot\,)$ as follows:

  $$\begin{aligned}&\mathbb{P}(\text{an event } E \text{ as a set of outcomes})\\ &= \textstyle\sum_{x \in E} p(x)\end{aligned}$$

- *Continuous Probability Distribution*: a function $p(x) : \mathbb{R} \to \mathbb{R}$ as a curve on the XY-plane whose area under the curve is 1, which can be used to construct a probability measure $\mathbb{P}(\,\cdot\,)$ as follows:

  $$\begin{aligned}&\mathbb{P}(\text{an event } E \text{ as an interval [a, b] on the } X\text{-axis})\\ &= \text{the area under the curve between the two vertical lines } x = a \text{ and } x = b\\ &= \int_a^b p(x)\,dx\end{aligned}$$

Let's not worry about how this last probability measure $\mathbb{P}$ assigns numbers to a sigma-algebra of subsets of the real line $\mathbb{R}$.

4 Application: Returning to Hypothesis Testing

4.1 Recall $\theta = 1/3$ vs. $\theta = 2/3$

Recall this empirical problem:

- Competing Hypotheses:
  - H_0: The bias of the coin is $\theta = 1/3$.
  - H_1: The bias of the coin is $\theta = 2/3$.
- Data Sequences: binary sequences (of heads or tails) of finite length.
- Background Assumptions:
  - IID data.
  - $\theta \in \Theta = \{1/3, 2/3\}$.

Recall the inference method M^* discussed above, formally defined as follows:

- $M^*(x_1 \dots x_n) = H_0$ if
  - the frequency of `1`s in data sequence $x_1 \dots x_n$ is closer to $1/3$ than to $2/3$,
  - i.e., if $(x_1 + \cdots + x_n)/n < 0.5$.
- $M^*(x_1 \dots x_n) = H_1$ if
  - it is the opposite,
  - i.e., if $(x_1 + \cdots + x_n)/n > 0.5$.
- $M^*(x_1 \dots x_n) =$ `?` otherwise.

Notation: $x_1 \dots x_n$ is shorthand for finite data sequence $(x_1, \dots, x_n)$ of length n; X_1 is a specific random variable, and x_1 is an unspecified value that X_1 might take.

Exercise 7: Use Hoeffding's inequality to find a sample size that guarantees that the probability that [M outputs the true hypothesis] is at least 95%—guaranteed in the sense that this probability is at least 95% no matter which hypothesis on the table is true. Put formally, our current goal is to find an n such that

$$\begin{aligned} \mathbb{P}_\theta^n\Big(M^*(X_1, \dots, X_n) = H_0\Big) &\geq 0.95 \quad \text{if } \theta = 1/3, \\ \mathbb{P}_\theta^n\Big(M^*(X_1, \dots, X_n) = H_1\Big) &\geq 0.95 \quad \text{if } \theta = 2/3. \end{aligned}$$

Hint: $2e^{-2n(\cdots)^2} \leq 0.05$.

4.2 Standards: Uniform vs. Pointwise Convergence

The above exercise can be generalized to show that the inference method discussed above achieves this standard:

Definition (Uniform Convergence for Identification). An inference method M for an empirical problem is said to achieve the standard of *uniform (stochastic) convergence for identification* if and only if

for any threshold of high probability $1 - \delta$ less than 1,

there exists a sample size N such that,

for all possible worlds θ in the Θ given in the empirical problem (i.e. those compatible with the background assumptions of the empirical problem),

$\mathbb{P}_\theta^n\Big(M \text{ outputs the hypothesis true at } \theta\Big) \geq 1 - \delta$, for any $n \geq N$.

We will soon see that the above standard is too high to be achievable in some empirical problems—in that case, we have no alternative but to lower the bar. Here is a relatively low standard:

Definition (Pointwise Convergence for Identification). An inference method M for an empirical problem is said to achieve the standard of *pointwise (stochastic) convergence for identification* if and only if

for any threshold of high probability $1 - \delta$ less than 1,

for all possible worlds θ in the Θ given in the empirical problem (i.e. those compatible with the background assumptions of the empirical problem),

there exists a sample size N such that,

$\mathbb{P}_\theta^n\Big(M \text{ outputs the hypothesis true at } \theta\Big) \geq 1 - \delta$, for any $n \geq N$.

The higher standard demands a "critical mass" N that works uniformly across the board—for all possible worlds θ in the parameter space Θ. In contrast, the lower standard allows the critical mass to vary: each possible world θ can have its own critical mass N_θ that works for θ—possibly not for other worlds.

4.3 Hypothesis Testing: $\theta = 1/2$ vs. $\theta \neq 1/2$ (I)

Now switch to another empirical problem, which is the same as the previous one except for the following

- Competing Hypotheses: $H_0 : \theta = 1/2$ vs. $H_1 : \theta \neq 1/2$.
- Background Assumption: $\theta \in \Theta = [0, 1]$.

Then we have:

> **Impossibility Result (No Free Lunch in Hypothesis Testing)**. In the hypothesis testing problem of $\theta = 1/2$ vs. $\theta \neq 1/2$, no inference method achieves the standard of uniform convergence (for identification).

To illustrate, consider the inference method M defined below:

- $M(x_1 \dots x_n) = H_0$ if the frequency of `1`s in data sequence $x_1 \dots x_n$ differs from $1/2$ by less than $\epsilon = 0.001$.
- $M(x_1 \dots x_n) = H_1$ otherwise.

Exercise 8: Show that this inference method fails uniform convergence. Hint 1: Imagine an extremely large n (but keep it fixed), and figure out the relation between the following two quantities.

$$\sum_{x:\left|\frac{x}{n}-\frac{1}{2}\right|<0.001} \frac{n!}{x!\,(n-x)!}\,\theta^{x}\,(1-\theta)^{n-x} \qquad \text{for } \theta = 1/2$$

$$\sum_{x:\left|\frac{x}{n}-\frac{1}{2}\right|\geq 0.001} \frac{n!}{x!\,(n-x)!}\,\theta^{x}\,(1-\theta)^{n-x} \qquad \text{for } \theta \to 1/2 \text{ (but } \theta \neq 1/2)$$

Hint 2: Understand the second quantity as a function of θ. Hint 3: The failure of uniform convergence for the above inference method generalizes quite straightforwardly for all inference methods once you consider the following quantity:

$$\sum_{(x_1,\dots,x_n)\,\in\,A_0} \theta^{\,x_1+\cdots+x_n}\,(1-\theta)^{n-(x_1+\cdots+x_n)}\,,$$

where A_0 is a subset of $\{0,1\}^n$, used to define when an inference method outputs (or accepts) hypothesis H_0.

4.4 Hypothesis Testing: $\theta = 1/2$ vs. $\theta \neq 1/2$ (II)

Fortunately, it is not hard to prove that pointwise convergence is achievable.

EXERCISE 9: Prove the claim. Hint 1: Consider inference methods M of the following form.

- $M(x_1 \ldots x_n) = H_0$ if the frequency of `1`s in data sequence $x_1 \ldots x_n$ differs from $1/2$ by less than $\epsilon_n = 1/n^{(\cdots)}$.
- $M(x_1 \ldots x_n) = H_1$ otherwise.

Hint 2: Don't forget our old friend: $2e^{-2n\epsilon^2}$.

5 Loose End

Probability theory has been introduced as a tool to pursue, in a small way, an inferential task: hypothesis testing. Generalizing to other inferential tasks is conceptually straightforward, even when the required mathematics becomes sophisticated.

Whereas *hypothesis testing* concerns thinking probabilistically about two hypotheses at a time, *model selection* often concerns a countable collection of hypotheses, or models, at a time; and this includes selecting not only purely statistical models, but also probabilistic causal models. By contrast, *point estimation* deals with a very large number of hypotheses at once—typically a continuum of them, such as the unit interval $[0, 1]$ as the set of possible biases of a coin. In that setting, each value in $[0, 1]$ represents a hypothesis on the table.

Regression (including nonparametric regression) can be viewed as point estimation "on steroids": one estimates not a point in $[0, 1]$, but an element of a far more abstract space—for instance, a curve in a (possibly very large) space of functions $y = f(x)$ on the XY-plane. If we modify regression by upgrading the X-axis to a space of possible images, and downgrading the Y-axis to just two categories ("Yes, it's a cat" vs. "No, it's not"), we move from regression to a classic example of *classification*. The task of classification encompasses not only classifying images, but also classifying a word (or token) as a continuation of a given string of words—the basic idea underlying large language models.

These five inferential tasks—hypothesis testing, model selection, point estimation, regression, and classification—admit a fairly unified treatment in frequentist statistics and machine learning theory, but that will have to be the next story. I hope, however, that you are now motivated to learn more probability theory and to think probabilistically more often.