What to Believe and How Much

A short guide to statistical thinking



JOSE BLANCA

What to Believe and How Much

A short guide to statistical thinking

JOSE BLANCA

Contents

A note on why and how this book was written 1
This is not a pipe 2
Understanding the p-value 13
What went wrong with the p-value 20
Maximum likelihood 29
Inverse probability: the probability we actually want 35
Bayesian inference 42
The frequentist alternative 50
The Bayesian defense 55
A pragmatic truce 62
Pluralism and craft 66

A note on why and how this book was written

For many years I have taught bioinformatics and used statistics in the laboratory, and over that time I've seen the same two problems again and again. The first is that many people who use statistics don't really understand what they're doing. The second is that statistics can give a false sense of objectivity: a p-value in a paper makes a result look scientific, whether or not the reasoning behind it holds up.

I think that understanding these tools is essential for anyone who uses them: biologists, psychologists, doctors and many others. And yet most people don't understand them. Partly, that's because the ideas are genuinely difficult. But it's also because statistics is usually taught as a set of recipes, without the conceptual framework that explains what each method does and why. This book tries to provide that framework.

There's also a more personal reason. When I started, I was a Bayesian, and I wanted to explain why the frequentist approach and the p-value were wrong. I still think that the way the p-value is usually used is wrong. But along the way I stopped being a Bayesian. Today I'd call myself a pluralist, and in a way this book traces that journey.

This book was written, as we say in Spanish, *a cuatro manos*: with four hands. The ideas and the perspective are mainly mine, and I'm responsible for the final result, including any errors. But Claude Opus 5.5, an AI model made by Anthropic, has worked with me throughout as an editor, a teacher, a fact-checker and an English corrector. Without its contributions, this book wouldn't have been possible.

This is not a pipe

What do you think a p-value means?

Imagine that you are researching the effectiveness of some therapy: you treat some people, leave a control group untreated, and then compare the results for the two groups. After doing the data analysis you get a p-value of 0.01, well below the standard 0.05 threshold. Now think about what this p-value means. Which of these statements are true?

1. You have absolutely disproved the null hypothesis (that there is no difference between the population means).
2. You have found the probability of the null hypothesis being true.
3. You have absolutely proved your experimental hypothesis (that there is a difference between the population means).
4. You can deduce the probability of the experimental hypothesis being true.
5. You know, if you decide to reject the null hypothesis, the probability that you are making the wrong decision.
6. You have a reliable experimental finding, in the sense that if the experiment were repeated a great number of times, you would obtain a significant result 99% of the time.

Pick your answers before reading on.

The right answer is that all six are false. Don't worry if you didn't get it: few people really understand the meaning of the p-value. The p-value is not the probability that you are wrong when you reject the null hypothesis. Nor is it the probability that the results are due to chance, or one minus the probability that your hypothesis is true. In fact, the p-value is not the probability of any hypothesis being true or false. So what is it? It is something that has

to do with the data that you would expect to see if the null hypothesis were true. We'll see exactly what later on.

Michael Oakes gave these statements to British academic psychologists in 1986. In 2002, Heiko Haller and Stefan Krauss gave them to three groups from six German universities: 44 psychology students who had passed their statistics courses, 39 psychology professors and lecturers, and 30 people who taught statistics and null hypothesis testing. Every single student believed at least one of the statements was true. So did 90% of the professors and lecturers, and 80% of the statistics teachers. The most common mistake was statement 5, the idea that the p-value is the probability of being wrong. About 70% of each German group believed it, and 86% of the British psychologists. Similar studies with academic psychologists in Spain, Italy and Chile have found the same errors. In Israel, Ruma Falk and Charles Greenbaum asked students the same kind of question. Beforehand, the students had read a classic article warning against these very mistakes, and "none of the statements is correct" was offered as an explicit option. Even so, 87% chose at least one wrong interpretation.

Why bother?

Statistics is fundamental to modern science, and the p-value is one of the most common statistical tools. Yet most scientists do not understand what they're doing when they apply it. Most of us learned statistics as a collection of recipes. You have two groups, you run a t-test. You have a contingency table, you run a chi-square. If the number that comes out is smaller than 0.05, you are happy, you put an asterisk over the bar in your plot and you move on. And if somebody asks you what that number means… well, let's hope nobody asks.

The physicist Richard Feynman had a name for this: *cargo cult science*. After the Second World War, some Pacific islanders who had seen planes land full of goods built runways and control towers out of bamboo, and waited for the planes to come back. Everything looked right, but the planes never landed. Doing statistics by recipe is a lot like that. We follow the ritual (collect the data, run the test, check the threshold) without understanding what the ritual is for. But science is not about collecting data and running tests to get results. It is about understanding the world. And this is not only a theoretical problem. The misuse of statistical tools, and of the p-value in particular, has become a real practical problem for science. It is one of the

causes of the replication crises, alongside others such as publication bias, career incentives, and studies too small to detect anything. Understanding what we are doing won't fix all of that, but it is a necessary start.

If you marked any of the statements in the quiz as true, and you think that understanding what we do matters, this book may be useful to you. I won't explain how to run any test. What I want is to understand what we are doing when we do statistics, and why we do it the way we do. I'll warn you in advance: a lot of what follows is philosophy. Behind the p-value, and behind any other statistical tool, there is philosophy. And there are deep philosophical disagreements that statisticians have been fighting about for a century.

Gamblers and insurers

Let's start at the beginning.

Probability theory was born at the gambling table. The first people to think seriously about probabilities wanted to know their chances at cards and dice. Gerolamo Cardano wrote a book on games of chance around 1564 (in Latin, if you fancy some light reading). Galileo was asked about dice around 1620. In 1654 Pascal and Fermat exchanged letters about how to split the stakes of an unfinished game, and Huygens published a small treatise in 1657. All of them asked the same kind of question: if I know how the game works, what's the probability of each outcome? That's a kind of p(data | hypothesis).

Then, in 1713, Jacob Bernoulli asked the reverse question: if I've observed some frequencies, how sure can I be about the underlying chance? That's the inverse probability, the one we actually want, already showing up at the beginning of the 18th century. An example of this inverse probability: if I'm playing dice with a friend, given the outcomes I've seen, what's the probability that my friend is cheating? Also quite early, De Moivre found that the bell curve, the normal distribution, appears as an approximation to the binomial when you toss a coin many times.

In parallel, the first statistics appeared. The word itself comes from *state*: originally, statistics meant collecting data about the state, such as how many people live here, how much land and wealth there is… and, of course, how much tax can be collected. Not the most romantic origin, but a practical one. And those first statistics were about people being born and people dying. In 1662 John Graunt turned London's weekly death records into the first

statistical tables. In 1693 Edmond Halley, the comet one, used a mortality table to calculate the price of life annuities. Who do you think was very interested in knowing the probability of people dying or getting sick? Insurers. If I sell this person a life insurance policy, how much is it going to cost me? Insurers were among the first to work with statistics.

In 1710, John Arbuthnot noticed that in London, for 82 consecutive years, more boys than girls had been christened. If it were a matter of pure chance, 50-50 each year, the probability of that would be $(½)^{82}$, a ridiculously small number. That's arguably the first significance test in history: the 50-50 hypothesis was rejected. Arbuthnot's conclusion, by the way, was that this was evidence of Divine Providence.

Models

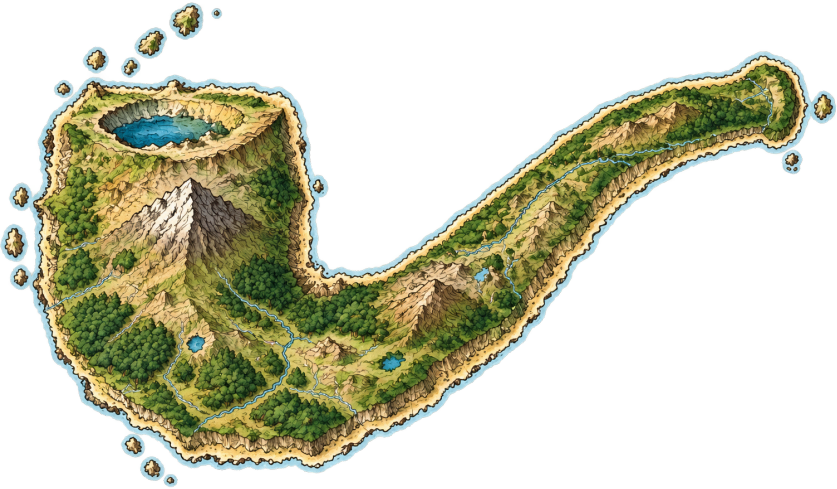
What are we doing when we do statistical inference? In its most basic form: we have some data and we want to end up with models of the world that approximately match the data.

Imagine that we go out into the street and measure the height of everybody who walks by. That's our data, a long list of numbers. Then we plot the distribution of those heights, perhaps one for women and another for men, and we conclude: this is how human height is distributed. That distribution is our model. We started with data and we ended up with an idea about how the world is. At least, that's how it looks. We'll see later that the model didn't come from the data alone.

This is not a pipe

There is a famous painting by René Magritte. It shows a pipe and, below it, the caption *Ceci n'est pas une pipe*, "this is not a pipe". The first reaction is: what do you mean it's not a pipe? Of course it is a pipe! Well, no. It is a painting of a pipe.

This might look like a silly joke, but it is one of the most important ideas of this whole book, and I'll keep coming back to it. There is reality, and there are the representations, the models, the ideas that we make of reality. They are not the same thing. Another way of saying it: **the map is not the territory**.



This is not an island. It's a map of one.



The territory.

A map is a simplified representation. It does not show every stone. It may show a river, but when you get there you might find only a dry riverbed.

This idea, that the model is not reality, is behind some of the deepest disagreements that statisticians have had for more than a century.

We confuse reality with our ideas about reality all the time, in our everyday experience of the world. If I tell you "there's a chair over there", is there a chair? Most people would say yes, obviously. But a physicist could say that there's a bunch of atoms. Who is right? Both. "Chair" and "atom" are categories that we use to represent reality, and they are good categories because they are useful. If I tell you there's a chair, you can sit down. If I tell you there's a bunch of atoms, you won't know what to do with that information, but a chemist will. In that sense, chairs and atoms are both real: they are useful representations of reality itself. But they are maps, not the territory. Many of you may think this is a pointless discussion, of interest only to philosophers who don't care about practical questions. But unless you take it seriously, you'll have a hard time understanding statistics and what we can do with it: understand disease, test a new drug, or, as we'll see later in the book, even help win a war.

What makes a good model

Science does not even aim to create exact models of reality; scientific models ignore many aspects of reality on purpose. When you studied the ball rolling down the inclined plane in physics class, you probably ignored the rotation of the ball and the air resistance. Was the model wrong? Strictly speaking, yes. Was it useful? Very. You went to the lab and made some predictions with your simplified model. If they more or less matched the measurements, that's all you wanted.

A good model, then, is an analogy that works:

- it represents something about the real world, and its predictions can be compared with what we observe,
- but it is simple enough to be understood and used.

Any of our ideas about the world, any hypothesis, theory or common idea, should be like this: a simplification that works. A statistician, George Box, put it like this: *all models are wrong, but some are useful.*

There are two key properties of models. First, **a model makes predictions**. If I study Newton's physics, I can predict how long the ball will take to fall. Second, **those predictions can be checked**. Keep these two properties in mind, because they will come back. Like the map and the territory, this might seem simple, but remember it: it is one of the foundations of statistical

and scientific thinking. I can go to the lab, measure, and see whether the model fits the data or not. If I imagine that the door is over there, I can try to walk out of the room. I may be right, or I may walk into a wall. As simple as that.

Statistical models

We have been talking about scientific models. Statistical models are a special kind of scientific model: they include some idea of randomness.

Imagine that I have a coin in my hand. Not a real coin: an imaginary one. I can imagine that this coin has a probability of 0.5 of landing heads. That's a typical statistical model. Assuming that model, could I predict how many heads I would get if I tossed the coin 100 times? Not exactly. I can't predict the result of a single experiment. But I can predict the distribution of outcomes if I repeat the experiment many times: sometimes I'll get 50 heads, sometimes 45, sometimes 48, sometimes 55. So from a statistical model I can calculate the distribution of outcomes I expect to see: how probable it is to get 50 heads, or 45, or 70.

Notice the direction of that calculation. We start from the model and get the probability of the data. The model tells us how probable the data are, not how probable the model is. Keep this in mind: swapping the two is one of the most common errors with p-values.

On top of that, statistical models usually have parameters. I could imagine a coin that lands heads with some unknown probability h, which might be 0.5, or 0.4, or 0.3. Finding the value of h that fits the data is one of the typical things we'll want to do. This is called **estimation**.

Models usually also make assumptions: for instance, that each toss is independent of the previous one.

We don't have direct access to reality. We only have a proxy: the data we gather through experiments and observations. With that data we can estimate the parameters of our models, check whether they fit the world, and compare them.

Models don't come from data alone

Where do the models come from? We need some method to create them, and here there is an important point: models can't be created from the data alone.

Remember the heights example. It looked as if the model had come straight out of the data. But look at everything we brought to the data before we had a single measurement. We chose to measure height, and not shoe size or eye colour. We decided that splitting the data by sex was a good idea. And when we said "this is how height is distributed", we were probably already imagining the familiar bell curve. The data didn't tell us any of that. We did.

This is analogous to what happens with your senses. You open your eyes, light comes in, and a cat appears in your mind. You might think that the image of the cat is created from the light alone, but it is not. It is created by a very complex machine, your visual system, which uses that light. Without the visual system there is no cat in your mind. This might seem a mere detail, until you think about the times your visual system fools you.

This is not a new idea. Philosophers of science call it the *theory-ladenness* of observation: we never observe "raw" data, because what we look at, what we measure and how we interpret it already depend on our ideas about the world. Machine learning has met the same problem from the engineering side and calls it *inductive bias*. A learning algorithm needs some built-in assumptions about what kind of pattern to look for, and without assumptions, learning from data is simply impossible. So, to build models from data you need a method, and every method carries assumptions.

What statistics is for

Statistics does different things with data.

The first is to **describe** the data. The simplest thing we can do is summarize the data. How many heads and how many tails came up? What's the distribution of heights? This is descriptive statistics, and for that we can use, for instance, charts. This is extremely useful, and you should never skip it. In a way, the plots and summaries are already simple models of the data.

The second is to **build** models. Once we have chosen the kind of model, say a coin that lands heads with some probability h, building it mostly means estimating: adjusting its parameters until the model fits the data as well as possible. What is h? How big is the effect of the therapy? And how uncertain are we about those values? Estimation can also be seen as choosing between models. A coin with $h = 0.5$, another with $h = 0.4$, another with $h = 0.3$: they are all competing models that differ only in the value of one parameter, and estimating means picking the one that fits the data best.

The third is to **check** models against the data. Checking is testing whether a model fits the data. To do it, we compare the model's predictions with the data, and the p-value is one way of doing that.

Finally, once we have a model we trust, we can **use** it to make predictions about the world: how many heads we'll get in the next 100 tosses, or how a new patient will respond to the therapy. Prediction is one of the main reasons for building models in the first place, and it is most of what machine learning does.

A philosophy teaser

There are two big philosophical ideas and one trap behind the ideas that we are going to explore.

The first idea is that there is a deep **asymmetry between rejecting and accepting** a hypothesis. Rejecting is more straightforward than accepting: we can see millions of white swans and still be wrong to accept that all swans are white, while just one black swan is enough to reject the hypothesis.

The trap is thinking that the probability of having a disease given a positive result in a check-up is the same as the probability of a positive result if I am ill. They sound similar. They are radically different. The first is the probability of the hypothesis given the data; the second is the probability of the data given the hypothesis. The first one is called an **inverse probability**.

And the second philosophical idea is related to the probability of the hypothesis: **can a probability represent a degree of belief?** If I say "after the experiment, I'm 90% sure the drug works", does that 90% mean anything?

This second idea deserves a short preview, because it has caused a lot of debate, and it is the disagreement that underlies two different approaches to statistical analysis. Imagine that I have a real coin. I can say it has some probability of landing heads, and I can measure it: I toss it many, many times and compute the proportion of heads. This approach seems quite uncontroversial. But what about this idea: "I think the probability that it will rain tomorrow is 0.8". How do you measure that? Tomorrow only happens once. It will rain or it won't. Where's the frequency? The frequentists' position was that hypotheses should not have probabilities attached to them, and that you can't measure the probability of something that only happens once.

Bayesians, on the other hand, fully accept that we can assign probabilities to hypotheses.

The problem is that if you accept the frequentists' position, you can never talk about the probability of a hypothesis. A hypothesis is true or false, just like it either rains tomorrow or it doesn't. All you can say is: if the hypothesis were true, I'd expect to see this. That's why traditional statistics is so focused on the p-value, which is a probability of the data, not of the hypothesis. If hypotheses can't have probabilities, p(data | hypothesis), the probability of the data given the hypothesis, is all that's left, and confusing it with p(hypothesis | data) is the trap. Meanwhile, every weather forecaster on TV happily tells you that the probability of rain tomorrow is 80%, and they seem to believe it means something. This discussion may seem theoretical and of no practical consequence, but it will come back again and again, and it has deep implications for how we do statistics, and even for how we decide whether a cancer treatment works.

A book to read alongside this one

If you want a good introduction to statistics, I recommend David Spiegelhalter's *The Art of Statistics: Learning from Data* (2019). It is not a list of recipes. It tries to explain what statistics is about, and it does it very well. It is the perspective of a practising, experienced statistician, full of real examples. My perspective will be a bit more philosophical, so I think the two complement each other.

Further reading

- Spiegelhalter, D. (2019). *The Art of Statistics: Learning from Data*. Pelican.
- Gigerenzer, G. (2004). Mindless statistics. *The Journal of Socio-Economics*, 33, 587–606. The source for the quiz and its results.
- Haller, H. & Krauss, S. (2002). Misinterpretations of significance: a problem students share with their teachers? *Methods of Psychological Research Online*, 7(1).
- Oakes, M. (1986). *Statistical inference: a commentary for the social and behavioural sciences*. Wiley.
- Falk, R. & Greenbaum, C. W. (1995). Significance tests die hard: the amazing persistence of a probabilistic misconception. *Theory & Psychology*, 5(1), 75–98.

- Badenes-Ribera, L. et al. (2016). Misconceptions of the p-value among Chilean and Italian academic psychologists. *Frontiers in Psychology*, 7, 1247.
- Greenland, S. et al. (2016). Statistical tests, P values, confidence intervals, and power: a guide to misinterpretations. *European Journal of Epidemiology*, 31, 337–350.

Understanding the p-value

Take a coin and toss it 100 times. You get 60 heads. Is the coin fair? Run the exact binomial test and the software gives you p = 0.057. Just above 0.05, so it's not significant. Does that mean the coin is fair? That it's almost unfair? And what does that 0.057 actually mean?

You have used p-values, and you will probably use them for the rest of your scientific life. So try to answer that last question. The most common answer I get is: "it's the probability that the hypothesis is true". If you told that to Ronald Fisher, the man who popularized the p-value, he would not be happy. There is no probability of any hypothesis in a p-value. None.

Don't feel bad. As we saw with the quiz, almost nobody understands the p-value, including many of the people who teach it. Part of the problem is that its definition is a tongue-twister: the p-value is the probability of obtaining data at least as extreme as the data observed, if the null hypothesis were true.

P-values, like other statistical concepts, are not easy ideas. Fisher worked at an agricultural research station, among scientists who needed practical methods, not mathematical proofs. So he wrote books for them, full of worked examples and with very little theory: *Statistical Methods for Research Workers* and *The Design of Experiments*. They were enormously successful, and many of their readers, and later many textbooks, used them as recipes to be applied rather than ideas to be understood. What you were taught in your statistics course is, more or less, those recipes. Let's try to understand them.

Am I seeing what I expect to see?

The p-value rests on two ideas:

1. If I assume a hypothesis, a model of the world, I can calculate what I expect to see.

2. If what I see is very different from what I expected, I reject the hypothesis.

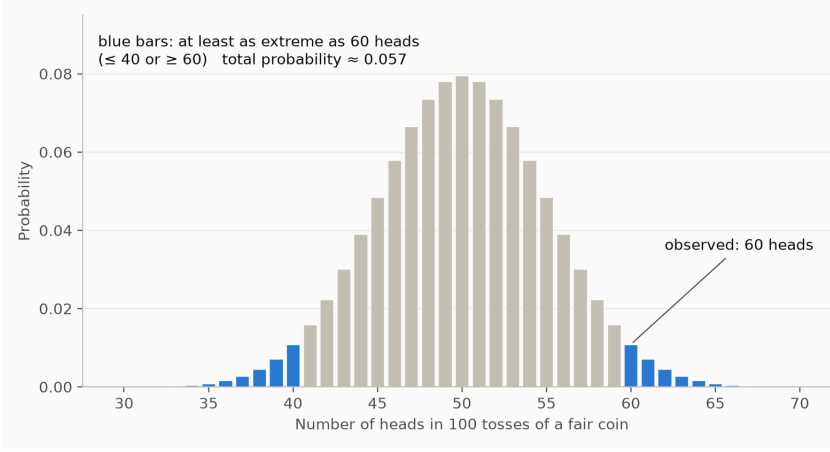
Here's an example without any probability. Imagine that I want to check the hypothesis "there are no students in the classroom". If that were true, what would I expect to see when I open my eyes? Nobody. I open my eyes, I see people sitting there. I reject the hypothesis, and I do it with complete confidence.

That's exactly what a significance test does. The only difference is that statistical models include randomness, so their predictions are not a single outcome but a distribution of outcomes.

Back to the imaginary coin in my hand. Let's assume that it lands heads with probability 0.5. If I tossed it 100 times, what would I expect to see? Most of the time something around 50 heads: 45, 48, 52, 55. I can calculate that distribution precisely. Now I take a real coin, toss it 100 times, and get 100 heads. That would be extremely weird for a fair coin, so I reject the hypothesis that the coin is fair.

That's the p-value: if the hypothesis were true, how probable would it be to observe something as extreme as what I observed, or more? If that probability is very small, the data is very weird under the hypothesis, and I become suspicious of the hypothesis. For 100 heads out of 100, that probability is about 0.000000000000000000000000000002. Weird indeed.

Now let's go back to the less extreme case from the beginning: 60 heads. The figure shows what a fair coin would give us if we repeated the 100 tosses many, many times.



Most of the time we'd get something between 40 and 60 heads. Exactly 50, the most probable result, happens only about 8% of the time. That's why we don't ask "how probable is it to get exactly 60 heads?": any particular result is improbable. What makes 60 heads suspicious is how far it is from what we expected. So we add up the probability of every result that is at least that far from 50, in either direction: 60 or more heads (about 0.028), and 40 or fewer (another 0.028). In total, about 0.057. That's the p-value, and that's what "at least as extreme" means.

Is 0.057 weird enough to reject the hypothesis that the coin is fair? By the usual convention, 0.05, it isn't, though only just. We'll come back to that convention.

Rejecting hypotheses

When using p-values we do something quite peculiar. If I want to know whether there are students in the classroom, I don't propose that hypothesis. I propose the opposite one: that there are no students. If I want to know whether a drug improves cancer treatment, the standard procedure is to assume that it doesn't (the *null hypothesis*) and try to reject that. Why don't I just propose that the drug works and try to accept that? The reason is a philosophical idea: the asymmetry between rejecting and accepting hypotheses. Rejecting can be a clear-cut decision, but accepting is much riskier.

To understand why we test the hypothesis we *don't* believe, we need a short detour through philosophy: swans, flying cows, Newton, and a Scottish philosopher who said things nobody wanted to hear. Bear with me; we'll come back to the p-value with a much better idea of what it's doing.

Black swans and flying cows

The classic example is the swans. For centuries Europeans saw swans, and all of them were white. Thousands of observations, always white. They concluded, with great confidence: "swans are white". Then someone went to Australia and found black swans. Thousands of observations had not been enough to prove the hypothesis, but a single observation was enough to destroy it.

This is the **logical asymmetry**: evidence can count decisively against a hypothesis much more easily than it can establish one.

Another example. Suppose I want to prove that cows fly. That's very hard. I go to the field and look at cows, and they're not flying. But maybe they're resting. Maybe they fly at night. I can never be sure they don't fly just because I haven't seen one doing it. However, if my hypothesis is "cows do not fly", disproving it is easy: one flying cow, and it's done.

Fisher wrote in *The Design of Experiments* (1935) that the null hypothesis "is never proved or established, but is possibly disproved", and that every experiment exists "only in order to give the facts a chance of disproving the null hypothesis". Since you can't accept hypotheses, only reject them, you always state the hypothesis you want to get rid of. If you want to know whether a drug works, the null is that it doesn't. If you're doing a genome-wide association study (GWAS) and want to know whether a locus is associated with a phenotype, the null is that it isn't. And if you want to check that cows don't fly? Then the null hypothesis is that cows do fly. Always the opposite of what you're interested in.

Fisher's idea was: look at whether the data is weird under your hypothesis. If it is, be suspicious of the hypothesis. If it isn't, perhaps the hypothesis is fine, although you can't be sure of that. As we'll see, what happened afterwards is a different story: statistics teaching turned it into a game of broken telephone.

The scientific method that doesn't exist

This idea is connected to what you were taught at school as *the scientific method*, more properly called the hypothetico-deductive method: you form a hypothesis, deduce what you should observe if it's true, go and look, and then see whether the hypothesis is supported or refuted. (There's no such thing as a single scientific method, but let's leave that for another day.)

Can you ever be sure of anything this way? Think about Newton. You were taught Newton's laws, gravity, the apples and the planets. Is Newtonian physics true? For two centuries every experiment fit: the planets, the tides, the cannonballs. And then someone looked carefully at Mercury's orbit, and its perihelion did not move as Newton predicted. Einstein came along with a theory that did fit. It's the swans again: we thought it was true, and observations showed it wasn't. So now we know that Newtonian physics is wrong. But are relativity and quantum mechanics true? So far no observation has contradicted them. Are they true, then?

The philosopher Karl Popper took this asymmetry to its limit. For him, science consists of rejecting hypotheses. A theory can never be proven true, but it can be tested and possibly refuted. And he made that the defining feature of science: **falsifiability**. A hypothesis is scientific only if it could, in principle, be refuted by observation. A scientific idea must put skin in the game: "I predict this, and if it doesn't happen, I'm wrong".

Falsification has a lovely property: in principle it only requires valid deductive logic. If the hypothesis implies X, and X doesn't happen, the hypothesis is false. But it also has serious limitations. In science we want to know which hypotheses are true, not just which ones are false. When I go to the lab, it's good to know I was wrong about something, but I also want to know whether the hypothesis I'm working with now makes sense. Popper would tell you it's impossible to know whether relativity is true. Meanwhile, engineers build GPS satellites with it.

Underneath all of this lies an older problem. David Hume, in the 18th century, pointed out that the sun has risen every day of your life, but nothing logically guarantees that it will rise tomorrow. Later philosophers made the same point with the swans: a thousand white swans don't logically guarantee that all swans are white. This is **the problem of induction**: inductive reasoning is not logically valid. The philosopher C. D. Broad expressed the hope that one day "Inductive Reasoning, which has long been the glory of Science, will have ceased to be the scandal of Philosophy".

Hypothesis testing: falsification with dice

That's the end of the detour, and now we can put the pieces together. **Significance testing is hypothesis rejection applied to probabilistic hypotheses.**

If our hypotheses made exact predictions, we could reject them with pure deductive logic, like the empty classroom. But statistical models include randomness, so our predictions are only probabilistic. The procedure is:

1. Propose a probabilistic hypothesis: "the drug does not work".
2. Calculate the distribution of outcomes you'd expect if the hypothesis were true.
3. Do the experiment and look at the data. If the outcome falls in the expected range, do nothing. If it's too weird, reject the hypothesis.

How weird is "too weird"? The usual convention is 0.05, which is one in twenty. If you use 0.05 you are saying: if I repeated this experiment twenty times and the null hypothesis were true, about once I would see something as extreme as this, or more, and nineteen times I wouldn't.

Let's get back to the cows. Strictly, I could never reject the hypothesis that cows fly just by watching cows sit in a field: maybe they were resting. But the p-value doesn't deal in certainty. It tells me how often I would see something this extreme if the hypothesis were true. If that only happens about one time in twenty, I'm willing to reject it. Let's make the flying-cow hypothesis probabilistic: cows fly, but only one hour a day; the other 23 they rest (poor things, flying is exhausting). If I go to a farm with one cow, I'd most likely see it resting. That proves nothing. What if I see 26 cows, all of them resting? Assuming each cow decides independently whether to fly, the probability of that is $(23/24)^{26} \approx 0.33$. Not weird at all, so I can't reject the flying-cow hypothesis. But if I find a farm with 71 cows and none of them is flying, the probability drops to about 0.05. Now it's weird enough: I reject the hypothesis that cows fly.

A word of warning, though. This is not quite the same as the empty classroom. There, the logic was airtight: if there are no students, I see nobody; I see people; so the hypothesis is false. With probabilities, the argument becomes: if the hypothesis is true, these data would be unlikely; I've seen these data; so the hypothesis is unlikely. It sounds just as reasonable, but it isn't valid. Try it with this example: if a person is an American citizen, it is very unlikely that they are a member of the US Congress. This person is a member of Congress. Therefore, they are probably not an American citizen. Obviously wrong.

So when we "reject" a hypothesis because the p-value is small, we are making a decision: the data are too weird for us to keep taking the hypothesis seriously. We have *not* calculated that the hypothesis is improbable. The p-value is a probability of the data, not of the hypothesis, and to get from one to the other we'll need more ingredients. That's the trap from the philosophy teaser, and we'll come back to it.

Remember: reject, never accept

The rule taught in statistics courses is: *a statistically significant result can lead you to reject the null hypothesis; a non-significant result does not allow you to accept it.*

So the statisticians tell us: you can only reject. You run your clinical trial and reject the hypothesis that your cancer therapy doesn't work. But does the therapy work? Usually we assume it does: "I reject that the drug doesn't work, therefore the drug works". It's as if the statisticians were saying "only reject" while elbowing you in the ribs and winking: *nudge nudge, wink wink, know what I mean?* And then, when it turns out that the drug doesn't work, the statisticians tell you: "I told you to only reject. You accepted the alternative; that's on you". We've reached a rather absurd situation. How did we get here?

Further reading

- Fisher, R. A. (1935). *The Design of Experiments*. Oliver and Boyd.
- Popper, K. (1959). *The Logic of Scientific Discovery*. Hutchinson. (First published in German in 1934.)
- Hume, D. (1748). *An Enquiry Concerning Human Understanding*. (First published as *Philosophical Essays Concerning Human Understanding*; retitled in 1758.)
- Gigerenzer, G. (2004). Mindless statistics. *The Journal of Socio-Economics*, 33, 587–606.

What went wrong with the p-value

We've seen what a p-value is and the logic behind it: assume a hypothesis, calculate what you'd expect to see, and reject the hypothesis if the data are too weird. And we ended up in an absurd situation. The statisticians tell us that we can only reject, and everybody accepts. How did we get here? And does it matter?

The bastard son

The p-value you were taught is a chimera, a mythological monster made from parts of different animals. It's the bastard child of two ideas, and neither of its parents would have recognized it.

Fisher's significance test. You propose one hypothesis, the null. The goal is to see whether that hypothesis is seriously discredited by the data. The p-value is a measure of how weird the data is under the hypothesis. Fisher suggested 0.05 as a convenient convention, but later insisted that no scientist uses a fixed level of significance in every case: for him, a small p-value was a reason for suspicion, not a verdict.

Neyman and Pearson's hypothesis testing. Jerzy Neyman and Egon Pearson (the son of Karl Pearson, the chi-square Pearson) were thinking more like engineers. In their approach there are two well-defined hypotheses, and the goal is to choose between them. They fixed two error rates in advance: α, the probability of rejecting a true null (type I error), and β, the probability of accepting a false null (type II error; $1 - \beta$ is the *power*). Those thresholds are chosen before seeing the data, based on how costly each kind of mistake is, and they describe how often the procedure goes wrong in the long run.

Why would the costs matter? I genotype tomatoes. If I get a genotype wrong, my population structure analysis gets a bit noisier, but nobody dies. A clinical geneticist who wrongly tells a patient that they carry a variant that determines their cancer treatment is in a very different situation. Neyman and Pearson's idea was that you decide in advance how many errors you can tolerate given what's at stake.

Notice the question they're answering. It's not "what should I believe?" but "what should I do?": which decision rule keeps my mistakes acceptable, given what's at stake. Fisher, in contrast, was asking what the data say about the null hypothesis. We'll come back to this distinction, because it's one of the keys to the whole debate.

NHST, null hypothesis significance testing, the hybrid you were taught: there's one null hypothesis, as in Fisher, with a vague alternative that nobody bothers to define. The stated goal is to reject the null, as in Fisher, but in practice you use the result to choose the alternative, as in Neyman–Pearson. You compute the p-value from the data, as in Fisher, but then compare it with a fixed threshold, like α in Neyman–Pearson. And β, the power, is usually forgotten. Neither Fisher, nor Neyman, nor Pearson would have accepted this.

	Fisher	Neyman–Pearson	NHST (what you were taught)
Hypotheses	one, the null	two, both well defined	the null, plus a vague alternative
Threshold	none, use your judgement	α and β, fixed in advance according to the costs of each error	0.05, always
What you get	a degree of suspicion about the null	a decision, with known long-run error rates	"significant", so the effect is real

The psychologist Gerd Gigerenzer called it **the null ritual**:

1. Set up a null hypothesis of "no effect", and don't state what your own hypothesis predicts.
2. Use 5% as the convention for rejecting it. If $p < 0.05$, accept your research hypothesis.

3. Do this always, whatever the problem.

Think about it for a second. You accept a hypothesis that you haven't even stated, about which you've calculated nothing. The only hypothesis you've defined and calculated anything about is the null, and the only thing you were allowed to do with it was reject it. But it's a ritual, and rituals are not meant to be understood. Every journal wants the bar with the asterisks, so you put the bar with the asterisks.

How did this happen? Nobody sat down and proposed it. Nobody ever wrote a paper in a statistics journal arguing that this is how things should be done. It was a sociological process, and the blame lies mostly with the textbook writers and statistics teachers of the 1940s and 50s. They thought that if they explained things properly nobody would understand them. Besides, there's never time for philosophy when there's a syllabus to cover. So they mixed both approaches into a simple recipe: $p < 0.05$, true. And a simple idea spreads well. Journal editors loved it. Compare Fisher's "we have data that make the hypothesis that the drug doesn't work look rather strange" with "$p < 0.05$, the drug works". Much easier to explain. Teachers copied each other, and here we are, seventy years later, still teaching it.

Does this have any practical consequences? Oh yes.

The replication crisis

Let's do a thought experiment. Imagine that someone invents a cancer therapy and runs a clinical trial with 100 patients. The p-value is not significant: the therapy doesn't seem to work. Then someone else tries another therapy. Not significant either. Now imagine that a hundred research groups, each with their own idea, each run their own trial. And suppose that, in reality, none of the therapies works at all.

What happens? About one in twenty of those trials will give $p < 0.05$ just by chance. Maybe the patients in the treated group happened to be a bit younger, or healthier, or exercised more. Nobody did anything wrong. And those five lucky groups will publish that their therapy works.

Now add a second effect. Of those hundred groups, ninety-five write a paper saying "we tried this molecule, and it does nothing". Five write "we tried this molecule, and it cures cancer". Which papers do you think get published? This is **publication bias**: journals select, among all the experiments that were done, precisely the ones that got lucky.

Put the two effects together. Five papers get published, and all five are wrong. The published literature on these therapies is 100% false, and nobody cheated.

Is this a theoretical worry? In 2005 John Ioannidis published a paper with the provocative title *Why Most Published Research Findings Are False*. He did calculations of this kind, somewhat more sophisticated than mine, and it caused a scandal. His key point was that the proportion of published findings that are false depends on how many of the hypotheses being tested were true to begin with. If most of the ideas we test are long shots, as in our thought experiment, most of the significant results will be flukes. That's a question about how probable the hypotheses were *before* the experiment, and it will come back when we talk about inverse probabilities and priors. He also looked at 49 highly cited medical studies; 45 had claimed that a therapy was effective. Of those, 7 were later contradicted by other studies, and another 7 had claimed effects stronger than the ones found later.

And Ioannidis's appraisal of the problem turned out to be too optimistic. In 2012 Glenn Begley reported that his team at a pharmaceutical company had tried to reproduce 53 landmark preclinical cancer studies. They could replicate 6. Eleven percent. The Reproducibility Project: Cancer Biology later completed 50 replication experiments from 23 papers: only 46% of effects replicated successfully, and for the positive effects the median effect size was 85% smaller than in the original papers.

In psychology, in 2011, Brian Nosek organized a collaboration to replicate 100 studies published in prominent journals. You know the kind of result that the popular press loves: holding a warm cup makes you judge people as warmer, a "power pose" makes you feel more confident. 97 of the 100 original papers had reported a significant effect. The replicators agreed on the methods with the original authors beforehand, so nobody could say it was done wrong, and there's no reason to think the original authors were cheating. They managed to replicate 35. And across all the replications, the effects were on average half the size originally reported.

Not everybody took the replication movement well. In the heated debates around it, the psychologist Daniel Gilbert shared a piece describing replicators as "the replication police" and "shameless little bullies". Questioning the p-value and the way science is done can make people angry.

P-hacking: the garden of forking paths

So far I've assumed that everybody does everything by the book. Now let's be realistic.

You run a trial with 100 patients in the control group and 100 in the treatment group. At the end, you find out that one patient in the treated group was hit by a bus. Obviously you remove them from the analysis; being hit by a bus has nothing to do with your drug. And you probably don't even mention it. Or… maybe they were hit by the bus because your drug made them very dizzy?

A very common situation in clinical trials: 20 of your 100 treated patients stop coming to the follow-up visits. What do you do? Do you say you treated 80? That's not what the statistical test assumes: you said you'd treat 100. And maybe those 20 left because the side effects were horrible and they didn't bother to tell you.

Well-run clinical trials have a standard answer to both problems: **intention-to-treat** analysis. You analyse every patient you randomized, in the group they were assigned to, bus or no bus, dropout or no dropout. But not every study is that careful, and outside clinical trials these decisions are often made quietly.

Another one. You test a drug for pancreatic cancer on 200 people and get no significant effect. But wait, there are men and women in the sample. Maybe it works better in women? You test men and women separately. Not significant in women, but significant in men, so you publish that the drug works in men. Is that correct?

It's not that drugs can't work differently in different groups; many do. The problem is statistical. One in twenty is the error rate *for one test*. You've just done three tests (all together, men, women). If you keep going, children from 0 to 10, teenagers, adults from 20 to 40, older people, smokers, non-smokers… each test is another roll of the one-in-twenty die. Keep rolling and eventually it comes up. And that's the one you publish.

Or you don't decide in advance when you'll stop the trial. You check the data every few weeks and stop as soon as the test is significant. That's like a football match in which, instead of playing 90 minutes, we play until my team is ahead. Who do you think is going to win? There are legitimate ways of looking at the data as they come in, called sequential designs, but they

make the threshold stricter at each look to compensate. Peeking without adjusting is the problem.

When it's done in search of a significant result, all of this is called **p-hacking**. But the statistician Andrew Gelman points out that the problem appears even when nobody is fishing on purpose, and he has a beautiful name for it: **the garden of forking paths**. Each decision looks innocent, even scientifically sensible:

- Should we exclude this weird data point? It looks like a measurement error, so dropping it seems like cleaning the data.
- Should we control for age or sex? It's plausible that age affects the outcome.
- Should we transform the variable? Taking the log makes the distribution look better, so it feels like better statistics.
- This test was not significant; should we try a different one?
- When should we stop collecting data?

Each path you take is another test, even if you only ever run "one" in the end. As the American Statistical Association put it in a statement we'll see at the end: proper inference requires full reporting and transparency. If readers don't know how many paths you explored, they can't judge your p-value.

Now add incentives. Imagine you finish your degree and go to work. Will your boss be as happy if your experiment shows that the thing doesn't work as if it shows that it does? If you're a PhD student, a significant result may be the difference between publishing and finishing your thesis or not. If you're a pharmaceutical company whose stock price depends on a new drug passing a clinical trial, what's your incentive to look for "strange" patients in the treated group who are dragging down the results? A company could even run many small trials instead of one big one. One big trial gives you one p-value; many small ones give you many rolls of the dice, and it's very easy for one of them to come out significant. The stock goes up. Later it turns out the drug doesn't work, but in the meantime someone made a lot of money. And people have been treated with drugs that didn't work.

When more data makes things worse

There's one more problem. Suppose your result is $p = 0.06$. So close! The natural idea is to do a bigger study, with more patients. More data means more power to detect things that are true, right?

With p-values, more data also gives you more power to detect things that don't matter. The more data you have, the smaller the difference that becomes significant. With enough data, you can reject any null hypothesis, because in the real world no two groups are ever exactly identical. And if rejecting the null means accepting your hypothesis, then with enough data you can accept any hypothesis you want.

An example. A study with about 19,000 participants asked whether couples who met online were more satisfied with their marriage than couples who met offline. The result was significant at $p < 0.001$. Not 0.05: 0.001! Ultra-significant. And the difference? On the satisfaction scale, 5.64 versus 5.48. Even if that difference is real, it's ridiculous. And it might not even be real, because with so much data any tiny bias in the methodology (say, people who met online were slightly more willing to answer an online questionnaire) is enough to make the result significant.

The result is statistically significant but practically meaningless. With the huge datasets we have today, it's everywhere. The more data you have, the more careful you need to be.

So why do we keep using it?

That's the question students always ask at this point. If the p-value causes so many problems, why is it still used?

There are several reasons. One is historical: textbooks are written this way, and teachers tend to follow the books. Another is the syllabus: explaining all this takes time away from other topics. There's also a philosophical reason, which we'll see later: there are alternatives, but some people, very intelligent people, have principled reasons to reject them.

But I think the biggest reason is that we crave objectivity. A regulator approving a drug needs to say "there was objective evidence that it works, no matter who looked at it". Imagine the company saying "well, we had some evidence that seemed to suggest that it might work; I got a 0.04, and I know I shouldn't take that as the truth, but it's one more piece of evidence". That doesn't fly. So the pressure is not to stop people using the p-value. It's the opposite, because the p-value looks objective. When you send a paper to a journal, the reviewers ask: "what objective statistical analysis supports this conclusion?" But in science that objectivity is not real: we weigh different pieces of evidence. When an effect is huge, when one bar is much taller than

the other, everybody agrees and you don't even need a test. When the bars are almost equal and you need a sophisticated test to tell them apart, the honest conclusion is: whatever the p-value says, be very suspicious.

Some fixes are being tried. In clinical trials there is now pressure to **preregister**: before starting the trial you declare how many patients you'll recruit, which groups you'll compare, which analyses you'll run, even whether you'll look at age groups. Then you run the experiment and fill in the numbers. That removes the forking paths, and for confirmatory studies, like a clinical trial that will decide whether a drug is approved, it's a great idea. But preregistration is not the answer to every problem. A lot of science is exploratory: "this came out weird, let me add a methyl group here and see what happens". Preregistration doesn't forbid that kind of work; it separates it from the confirmatory tests that should come after. Exploration still needs judgement, and no registration form can provide that. There's no objective set of rules that guarantees good science. That would be a scientific method, and as we said, it doesn't exist.

What the statisticians say

In 2016 the American Statistical Association was worried enough about how p-values were being misused that it issued an official statement with six principles. By now, you should recognize all of them:

- **1.** p-values can indicate how incompatible the data are with a specified statistical model.
- **2.** p-values do not measure the probability that the studied hypothesis is true, or the probability that the data were produced by random chance alone.
- **3.** Scientific conclusions and business or policy decisions should not be based only on whether a p-value passes a specific threshold.
- **4.** Proper inference requires full reporting and transparency.
- **5.** A p-value, or statistical significance, does not measure the size of an effect or the importance of a result.
- **6.** By itself, a p-value does not provide a good measure of evidence regarding a model or hypothesis.

Let me be clear about one thing: the p-value is not evil. Used the way Fisher used it, as one piece of evidence that makes you more or less suspicious of a hypothesis, it's perfectly reasonable. The problem is believing that $p < 0.05$

means the drug works. If you understand what you're doing, the conclusion is: "here is one more piece of evidence that seems to suggest that the drug works".

But notice what we've been doing all along. Every calculation so far has been of the form "if the hypothesis were true, how probable is what I've seen?". That probability has a name, the likelihood, and next we'll see that it can be used in a different way: instead of asking whether the data is too weird for one hypothesis, we can ask which of all possible hypotheses explains the data best.

Further reading

- Wasserstein, R. L. & Lazar, N. A. (2016). The ASA statement on p-values: context, process, and purpose. *The American Statistician*, 70(2), 129–133.
- Gigerenzer, G. (2004). Mindless statistics. *The Journal of Socio-Economics*, 33, 587–606. The source of "the null ritual".
- Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLoS Medicine*, 2(8), e124.
- Ioannidis, J. P. A. (2005). Contradicted and initially stronger effects in highly cited clinical research. *JAMA*, 294(2), 218–228.
- Begley, C. G. & Ellis, L. M. (2012). Raise standards for preclinical cancer research. *Nature*, 483, 531–533.
- Open Science Collaboration (2015). Estimating the reproducibility of psychological science. *Science*, 349, aac4716.
- Errington, T. M. et al. (2021). Investigating the replicability of preclinical cancer biology. *eLife*, 10, e71601.
- Gelman, A. & Loken, E. (2014). The statistical crisis in science. *American Scientist*, 102, 460–465.
- Cacioppo, J. T. et al. (2013). Marital satisfaction and break-ups differ across on-line and off-line meeting venues. *PNAS*, 110(25), 10135–10140.

Maximum likelihood

We have spent a lot of time on the p-value, and on the problems that come from misunderstanding it. But underneath the p-value there was a perfectly sound idea, and we're going to keep using it: if I have a well-defined hypothesis, I can always calculate what I expect to see. We can also use that idea to look for the hypothesis that best explains the data.

Let's go back to my imaginary coin. It's in my hand, and since I'm imagining it, I can decide that it lands heads with probability 0.5. If I tossed this imaginary coin 10 times, how often would I get 0 heads? 1 head? 2? … 10? I can calculate all of that. If I tossed it 100 times, how probable would it be to get exactly 50 heads and 50 tails? I can calculate that too. If my hypothesis is precise enough, I can always predict the probability of what I'd observe in the world I've imagined.

We write that probability like this:

p(data | hypothesis)

The vertical bar reads "given" or "assuming": the probability of the data, assuming the hypothesis is true. Every calculation we did with p-values was of this kind: the probability of observing some data, assuming a particular hypothesis.

A different question

With the p-value we asked: *is what I'm seeing so weird that I'd better reject the hypothesis?* I toss a real coin 100 times, get 100 heads, and ask how weird that would be for a fair coin.

Maximum likelihood asks a different question: *of all the hypotheses I'm considering, which one explains the data best?*

What does "all the hypotheses" mean? Think of all the imaginary coins I could hold in my hand. I could imagine a coin that always lands heads. Or one that always lands tails. Or one that lands heads with probability 0.5, or 0.9, or 0.8, or 0.1234. Now suppose I take a real coin, toss it 100 times, and get 100 heads. Of all those imaginary coins, all those different models, which one makes 100 heads out of 100 most probable?

A coin with a probability of 0.1 of landing heads could produce 100 heads; that would be possible, just extraordinarily unlikely. If the coin had a probability of 0.9, 100 heads would be much more plausible. And with a coin that always lands heads, probability 1, getting 100 heads is not only plausible: it's certain. So the coin that best explains my data is the one that always lands heads.

That's maximum likelihood. Of all the hypotheses we're considering, choose the one that makes the data you observed most probable. Notice the "we're considering". Maximum likelihood doesn't invent hypotheses: it chooses among the ones we propose. As we explained before, models don't come from data alone. We decide beforehand which kind of model makes sense, and the data then pick the best version of it.

Likelihood

The probability of the data assuming a hypothesis is called the **likelihood**. (Strictly speaking, it's not always a probability: for continuous data we use probability densities instead. But the idea is the same.)

If I assume cows fly, what's the likelihood of going out to the field and seeing a cow flying? If I assume they don't fly, what's the likelihood of seeing a cow flying? None. If I assume a fair coin, what's the likelihood of getting two heads in two tosses? That one is also easy: 0.5 for the first toss times 0.5 for the second, $0.5^2 = 0.25$.

Now notice something. That 0.25 depends on the 0.5, the probability of landing heads. If my imaginary coin had a probability of 0.4, the likelihood of two heads would be $0.4^2 = 0.16$. If it were 0.9, it would be 0.81. If the coin never landed heads, the probability of getting two heads would be 0. And if it always landed heads, it would be 1.

So let's stop writing 0.5 and call the probability of heads h. Now the model has a *parameter*, and the likelihood is a function of that parameter. For two observed heads, it's h^2. That's called the *likelihood function*:

$L(\theta) = f(\text{data} \mid \theta)$

where θ stands for the parameters of the model (here, θ is just h), and the data are held fixed. Once we've done the experiment, the data are simply what we observed; what we vary is the hypothesis.

Finding the maximum

Suppose I toss a coin three times and get three heads. The likelihood function is h^3. If $h = 0.5$, it's 0.125. If $h = 0.1$, it's 0.001: very unlikely. If $h = 0.9$, it's 0.729. If I plot h^3 for every value of h between 0 and 1 and look for the highest point, I find it at $h = 1$. Given that data, three heads in three tosses, the maximum likelihood estimate is that the coin always lands heads.

Would you bet that this coin always lands heads? Probably not. You know how coins usually behave, and three tosses are not much evidence. But maximum likelihood only looks at the data, not at what we already know, so with little data it can be overconfident. Keep that discomfort in mind; we'll come back to it later.

Now suppose I toss it a fourth time and get tails: three heads and one tail. The likelihood function is now $h^3(1 - h)$. If $h = 1$, the likelihood is now zero: a coin that always lands heads could never produce that tail. If I plot the function and look for the peak, it's at 0.75.

That's what statistical programs do when they compute maximum likelihood estimates: they go through the possible values of the parameters and pick the one that makes the observations most probable.

Sometimes we can do this analytically. Finding the maximum of a function is a calculus problem: take the derivative, set it to zero and solve. For the coin, the probability of getting x heads in n tosses is the binomial:

$$p(x \mid h, n) = n! / (x! (n - x)!) \cdot h^x (1 - h)^{n-x}$$

Take the derivative with respect to h, set it to zero, and you get:

$$\hat{h} = x / n$$

The little hat over the h is how statisticians mark a maximum likelihood estimate. When you see a hat in a paper or a textbook, that's what it means. And the result is: count the heads and divide by the number of tosses. Exactly what your intuition told you. In this case, all this statistical machinery looks like overkill.

For many complicated models, the derivative can't be calculated, so programs search numerically: they try values, see whether the likelihood goes up or down, and keep climbing until they reach the top.

When intuition fails

For the coin, maximum likelihood gives the obvious answer. Is it always that boring? No.

Imagine we want to know how long patients live, on average, after being diagnosed with a particular kind of cancer. This matters a lot to patients and their families, to doctors choosing a treatment, and to anyone comparing a new therapy with the old one. So we follow 10 patients from the moment of diagnosis. We can't wait forever, so the study ends after five years. By then:

- 6 patients have died, after 0.5, 1, 1.5, 2.5, 3 and 4.5 years,
- 4 patients are still alive.

What is the average survival time after diagnosis?

The intuitive answer is to average the times we have: (0.5 + 1 + 1.5 + 2.5 + 3 + 4.5) / 6 ≈ 2.2 years. But that's wrong, and it's far too pessimistic. What do we do with the four patients who are still alive? We can't just ignore them, because they're telling us something important: some patients live more than five years. And we can't assign them the value five, because we know they live longer.

This is called *censored* data, and it's everywhere: patients still alive at the end of a study, machines that haven't failed yet, plants that haven't flowered when the experiment stops.

Maximum likelihood handles this naturally. We propose a model of how survival times are distributed, and we ask: what mean survival time makes our observations most probable? And "our observations" include both facts: six patients died at those particular times, and four were still alive after five years.

A simple model is the exponential distribution. It has one parameter, the rate λ, and the mean survival time is 1/λ. It assumes that the risk of dying stays the same year after year. That's a simplification, but it's enough to show the idea. The likelihood has two parts:

- for each patient who died at time t_i, the probability density of dying exactly then: λe^(−λt_i);

- for each patient still alive at $T = 5$ years, the probability of surviving at least that long: $e^{(-\lambda T)}$.

Multiply all of them, take the logarithm, take the derivative, set it to zero, and you get:

mean survival time = (sum of the observed survival times + years followed by the patients still alive) / number of patients who died = $(13 + 4 \times 5) / 6 = 33 / 6 =$ **5.5 years**

The maximum likelihood estimate of the mean is 5.5 years: more than twice the naive average, and beyond the end of our study. That makes sense, since almost half the patients were still alive after five years.

Comparing models

There's one more thing we can do with likelihoods. Once we have the maximum likelihood estimates of a model's parameters, we can calculate the likelihood of the observations under that model. So if we have several alternative models, we can repeat the process for each one: fit its parameters by maximum likelihood, then calculate how probable the data are under the best version of that model. Then we choose the model with the highest likelihood.

There's a catch, though. A model with more parameters will always fit the data at least as well as a simpler one, so if we only looked at the likelihood, the most complex model would always win, even when its extra parameters are just fitting noise. This is called *overfitting*. In practice we use criteria such as AIC or BIC, which take the likelihood into account but penalize each extra parameter.

Likelihood as evidence

Likelihoods can also be used to compare two specific hypotheses directly. Go back to the three heads in three tosses. A coin that always lands heads gives that result with probability 1; a fair coin, with probability 0.125. The ratio is 8: the data favor the always-heads coin over the fair one by a factor of 8.

Some statisticians, the *likelihoodists*, think this ratio is the best answer to the question "what do the data say?". It measures the evidence, and it's the same for everybody, whatever they believed before. Remember the bet: you probably still wouldn't bet that the coin always lands heads. For a

likelihoodist that's no contradiction, because what you believe, or what you decide to do, are different questions from what the data say. We'll come back to these three questions. The approach works best when there are just two precise hypotheses to compare; with more, things get more complicated.

What did we actually calculate?

Let's stop for a second and think about what we've done. We wanted to know the probability that the coin lands heads, given what we observed. But what did we calculate? The probability of what we observed, given each possible coin. And then we picked the coin that made our observation most probable.

Those are not the same thing. "Which hypothesis makes the data most probable?" is a different question from "which hypothesis is most probable, given the data?" You may think this is just a play on words. It isn't. As we'll see, confusing those two probabilities can lead you to believe you have HIV when you probably don't, and can send an innocent person to prison.

Inverse probability: the probability we actually want

What we really want to know

Imagine that we want to know **whether a drug works**, given the experimental data. Which probability would we calculate with the p-value? The probability of observing a certain number of people getting better *if we assume that the drug does not work.*

When we used maximum likelihood to estimate the probability that a coin lands heads, given what we'd seen it do, what did we calculate? The probability of observing the heads and tails we observed *if we assume the coin has a given probability of heads.*

When a doctor wants to know whether you have a disease, given your test results, what can be calculated easily? The probability of getting those test results *if you have the disease.* If I have a cold, what's the probability that I'm coughing and have a runny nose? High. Now, if I'm coughing and have a runny nose, what's the probability that I have a cold? That's a different question. Maybe it's an allergy. Maybe it's the flu. Maybe it's something else.

Do you see the pattern? In classical statistics we've always worked with

p(evidence | hypothesis)

but what we really want is

p(hypothesis | evidence)

They look like the same thing written backwards, and our mind tends to treat them as more or less equivalent. But they're not, and they can be wildly different. When I think "I'm coughing and sneezing, so I have a cold", I'm

reasoning about p(cold | symptoms). But all the statistics we've seen so far only gives me p(symptoms | cold).

You've tested positive for HIV

You're walking down Main Street, there's an awareness campaign, and you agree to take a test as part of a study on HIV prevalence in your community. They do the blood work. A few days later they call you: the test came back positive.

They also tell you that the test is quite good. Its false positive rate is very small. That's the probability of testing positive if you're not infected: p(positive | not infected).

So what's the probability that you're infected?

When people are asked this question, most of them, including many doctors, answer something like "very high, maybe 98% or 99%". After all, the test hardly ever gives a positive result to someone who isn't infected.

However, the probability of being infected can't be calculated from the information given. We're missing something essential. And, as we'll see, you're probably not infected, even though you tested positive.

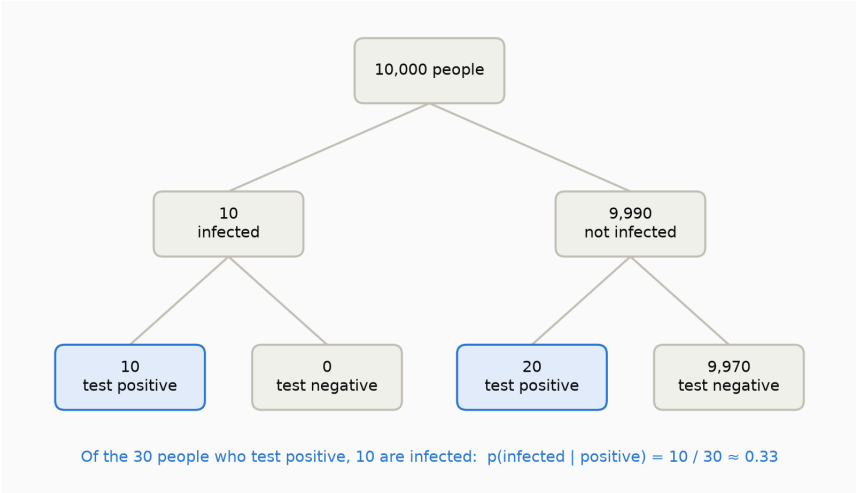
Think of everybody who took part in the campaign. If nobody had been infected, a few of them would still have tested positive: the false positives. If everybody had been infected, almost all of them would have. What you want to know is: *of all the people who tested positive, how many are infected?* And that depends on how many infected people there were to begin with.

Let's put in some realistic numbers. Suppose:

- the prevalence of HIV in the population is about 0.1%: one person in a thousand,
- the probability of a false positive is 0.2%,
- the probability of a false negative is very small, so we'll ignore it.

Take 10,000 people at random and test them all.

- 10 of them are infected (one in a thousand). They all test positive.
- 9,990 are not infected. Of those, 0.2% test positive anyway: about 20 false positives.



So there are about 30 positive results, and only 10 of them are infected. When they tell you that you've tested positive, you are one of those 30. The vertical bar in p(infected | positive) means exactly that: we're now restricted to the group of 30. And the probability of being infected is 10 out of 30: about 0.3. A test with a 0.2% false positive rate, and yet if you test positive, it's twice as likely that you're *not* infected.

Now change only one thing: the prevalence. Imagine a population where 1% of people are infected. Out of 10,000, 100 are infected and test positive. Of the 9,900 who aren't, about 20 test positive anyway. Now, if you test positive, the probability of being infected is 100 out of 120: about 0.83. Same test, same false positive rate, radically different conclusion.

That's also why it matters that you were tested at random, walking down the street. If you had taken the test because you'd been exposed to the virus, you would belong to a group where the infection is much more common, and a positive result would be much more worrying. And it's why, in practice, a positive result in an HIV screening test is always checked with a second, different test.

It's so easy to get this wrong that the mistake has a name: the base rate fallacy. In this case the base rate is the prevalence: how common the thing was before you looked at any evidence. Ignore it, and your conclusions can be nonsense.

The prosecutor's fallacy: the Sally Clark case

The same mistake has another name when it happens in a courtroom: the prosecutor's fallacy. In a trial, evidence is presented, and the jury must decide: given this evidence, is the accused guilty? The fallacy is to argue that, because the evidence would be very unlikely if the accused were innocent, the accused is very unlikely to be innocent. Once again: p(innocent | evidence) is not the same as p(evidence | innocent). Here's a real case.

In 1996, Sally Clark, an English solicitor, gave birth to an apparently healthy baby boy. Less than three months later, he died. She was alone with him; she said that he stopped breathing after she put him to bed. Sadly, sometimes babies die like this, suddenly and without any apparent cause. It's called sudden infant death syndrome, SIDS.

In 1997 she had another baby. He died within eight weeks, in similar circumstances. This time the baby showed signs of trauma. Clark said that the trauma was due to her and the paramedics trying to resuscitate him.

She was charged with murdering both children. And there wasn't much more evidence than this. So a paediatrician, Professor Roy Meadow, was called to testify about how likely it was that this had happened by chance. His reasoning went like this:

- The incidence of SIDS in a healthy, non-smoking family like the Clarks is 1 in 8,543.
- So the probability of two SIDS deaths in the same family is 1 in 8,543 × 8,543: **1 in 72,982,849**. One in 73 million.

Before reading on, put yourself in the jury's place. Is the calculation right? And given this evidence, is she guilty?

Flaw 1: the number itself

The figure of 1 in 8,543 had been obtained by picking the lowest-risk subgroup. The estimated rate of SIDS at the time was around 1 in 1,300. With that number, the same reasoning gives about 1 in 1.7 million. Still small, but 40 times larger.

Flaw 2: rare things happen all the time

Suppose something happens once in a million times. If you play a million times, will it happen? Very likely. If you play two million times, even more likely. Now, how many families with two children are there in the United Kingdom? Millions. So something with a probability of one in a million will

happen to some family, somewhere, every so often. The fact that it happened to *this* family is much less surprising than it sounds.

Flaw 3: independence

When can we multiply probabilities? If I roll a die, the probability of a six is 1/6. If I roll it twice, the probability of two sixes is 1/6 × 1/6 = 1/36. That works because the two rolls are independent: the first doesn't affect the second.

p(A and B) = p(A) × p(B), only if A and B are independent.

Were the two deaths independent? If there's a genetic factor that predisposes to SIDS, or an environmental one shared by the household, the first death makes the second more likely. A family that has already lost a child to SIDS is not a random family. If the deaths aren't independent, the probability of a second death might be 1 in 100, or 1 in 1,000. We don't know, because we don't understand the causes of SIDS well enough. But we certainly can't assume independence, and Meadow's calculation did. As we saw before, every model has assumptions, and independence is a common one. Here, ignoring one helped send an innocent woman to prison.

Flaw 4: the inverse probability strikes again

But the deepest problem remains even if we fix all of the above. Let's say that everything was calculated correctly. What has been calculated?

p(two deaths | SIDS)

The probability of two babies dying, assuming they died of natural causes. Is that what the jury needs? No. The jury needs

p(murder | two deaths) = 1 − p(SIDS | two deaths)

Given that two babies have died, what's the probability that their mother killed them? And p(SIDS | two deaths) is not p(two deaths | SIDS). It's the same kind of problem that we had with the HIV test.

Think about what would be needed to answer the real question. There are two hypotheses on the table: the babies died of natural causes, or the mother killed them. Double SIDS is extremely rare. But a mother murdering two of her babies is also extremely rare. The evidence (two dead babies) is very improbable under both hypotheses, so the question is which of the two very rare explanations is less rare. Showing that one explanation is very unlikely tells you nothing about guilt unless you also consider how unlikely the other one is.

We can count, as we did with the HIV test. In 2004 the mathematician Ray Hill made this comparison using the rates for families like the Clarks. He estimated that double SIDS is between 4.5 and 9 times more common than double murder. So think of all the families like the Clarks that lose two babies. For every 10 of them, roughly 8 or 9 lost them to SIDS, and only 1 or 2 to murder. The two deaths alone made Sally Clark's guilt possible, but not probable: somewhere between 1 in 10 and 1 in 5. Very far from one in 73 million.

Sally Clark was convicted in 1999. The Royal Statistical Society issued a statement criticizing the misuse of statistics in her trial. In 2003 her conviction was overturned on appeal. It had emerged that evidence of an infection in her second son had been withheld. She never recovered from the experience, and died in 2007.

What we really want

Everything we've seen in classical statistics, from the p-value to maximum likelihood, is built on p(data | model). We can calculate it whenever we have a well-defined model: we imagine how the world works, and compute what we'd expect to see.

But what we really want to know is whether we should trust the model, given the data we've observed: p(model | data). That is, the probability that the drug works, that the coin is fair, that the patient is sick or that the accused is guilty.

Statisticians used to call this *inverse probability*; today it's usually called the *posterior probability*. It's kind of funny, if you think about it: they called "inverse" precisely the probability we're interested in. Inverse to what? Inverse to the one that's easy to calculate from the model.

So, how can we calculate the probability of a hypothesis given the evidence, when all we know how to calculate is the probability of the evidence given the hypothesis? We've already seen a hint in the HIV example: we needed one more ingredient, the prevalence: how probable the hypothesis was *before* we saw the evidence. There's a theorem that tells us exactly how to combine those ingredients. As we'll see, using it for this purpose started one of the most bitter fights in the history of statistics.

Further reading

- Gigerenzer, G. (2002). *Calculated Risks: How to Know When Numbers Deceive You*. Simon & Schuster. (Published in the UK as *Reckoning with Risk*.)
- Gigerenzer, G., Gaissmaier, W., Kurz-Milcke, E., Schwartz, L. M., & Woloshin, S. (2007). Helping doctors and patients make sense of health statistics. *Psychological Science in the Public Interest*, 8(2), 53–96.
- Royal Statistical Society (2001). Royal Statistical Society concerned by issues raised in Sally Clark case. News release, 23 October 2001.
- Hill, R. (2004). Multiple sudden infant deaths: coincidence or beyond coincidence? *Paediatric and Perinatal Epidemiology*, 18(5), 320–326.

Bayesian inference

We have reached an uncomfortable conclusion. All the classical statistics we've seen calculates p(data | hypothesis), but usually we are interested in p(hypothesis | data): the probability that the drug works, that the patient is sick, that the accused is guilty, given what we've observed. How do we get from one to the other? The answer lies in one simple theorem.

A priest and his friend

In 1763 a theorem was published that lets us calculate the probability we actually want. Thomas Bayes (1701–1761) was a Presbyterian minister with an interest in mathematics. After Bayes died, his friend Richard Price, another minister, found among his papers an essay on how to reason from observed events back to their causes. Price edited it and published it in 1763. That's the origin of what we now call Bayes' theorem, although the essay doesn't look like the formula we write today. It's written in words and geometry.

Why were two ministers interested in this? Price was interested in thinking rationally about the existence of God, and the essay offered a rigorous method to reason from evidence back to causes: from what we observe in the world, such as the order and design we see in nature, to what produced it. Think of a cat: claws, fangs, fast reflexes, perfectly built for hunting. What's the probability that it was designed by someone, given what we see? That's the kind of question this new mathematics seemed to answer. So the theorem at the heart of Bayesian inference was born, in part, out of theology.

The person who really turned this into a working tool was Pierre-Simon Laplace, the great French mathematician. He arrived at the theorem independently in 1774, wrote the formulas in the form we use today, and applied them to statistical problems. One of the first Bayesian analyses in history

was his: given the observed numbers of boys and girls born, what's the probability that the underlying chance of a boy is greater than one half? He found it was so close to 1 that, in practice, it was a certainty. Laplace was a pragmatist, and he also used what we'd now call frequentist methods when they were convenient.

Bayes' theorem

$p(A \mid B) = p(A) \cdot p(B \mid A) / p(B)$

That's Bayes' theorem. It's very easy to prove; it follows almost directly from the definition of conditional probability. And notice what it does: it has $p(A \mid B)$ on one side and $p(B \mid A)$ on the other. It's precisely the bridge between a probability and its inverse.

Now replace A and B with hypothesis and data:

$p(\text{hypothesis} \mid \text{data}) = p(\text{hypothesis}) \cdot p(\text{data} \mid \text{hypothesis}) / p(\text{data})$

That's *Bayesian inference*. Given the evidence, what probability should I give to my hypothesis? Given that the X-ray shows a shadow, what's the probability that my patient has cancer? Given that two babies have died, what's the probability that their mother killed them?

There's an important subtlety here. Nobody disputes the first formula. It's a theorem. But the second one, replacing A and B with hypotheses and data, has been fiercely disputed throughout the 20th century, and some people are still unhappy with it today.

The pieces of the formula

We have already seen *the likelihood*, $p(\text{data} \mid \text{hypothesis})$. If I assume the hypothesis is true, if I assume my patient has cancer, how likely is it that I'd observe the data I've observed?

The probability of the data, $p(\text{data})$, is the probability of seeing what we've seen, taking into account all the hypotheses. If the mother is a murderer, how probable is it that the babies die? And if she isn't? If my patient is sick, how probable is it that I see a shadow on the X-ray? And if they're healthy? We weigh each hypothesis by how much we believed in it and add everything up. (In Bayesian inference we always consider all the hypotheses, not just one, as we did with the null hypothesis.)

The ratio between the two, p(data | hypothesis) / p(data), tells us how strongly the data favor this hypothesis over all the others. Let's call it the *strength of the evidence*. Here's an example of evidence that has no strength at all. Suppose I bring a bottle of water from my house and tell you: this is the water from my tap, and if you drink it, you'll live for many years. You drink it, we wait, and indeed, most people who drink it live for many years. Should we conclude that the water from my tap is miraculous?

No. Why not? Because those people would have lived for many years without it too. Both hypotheses, "the water is miraculous" and "the water is just water", predict the same thing. When we compute the ratio, it comes out at about 1: these data don't support one hypothesis more than the other. So I'm left with what I believed before starting. And before starting, how probable did I think it was that my tap water was miraculous? I don't know many miraculous waters. That's exactly how homeopathy works. "Take this for your cold and you'll get better." And you do get better. But you'd also have got better without it.

The prior

And so we reach *the prior*: p(hypothesis). It's the probability of the hypothesis before considering the data.

This is not optional. If I want to know how probable it is that my patient is sick after seeing the test results, then, before the patient walks into the consultation room, I already need a probability that they're sick. Before going to Sally Clark's trial, I need a probability that she's a murderer. That's what was missing in the HIV example: the prevalence was the prior.

To see why likelihood alone is not enough, here's an example. You arrive at a friend's house and find a broken window and a stone on the floor. What happened? Let's propose two hypotheses:

1. An elephant was playing in the garden and broke the window. Then, to cover up for the elephant, his friend, a Martian, put the stone on the floor and they both ran away.
2. The kid who lives in the house was playing, throwing stones in the garden, and one broke the window by accident.

What's the likelihood of hypothesis 1? If we assume that the elephant and his Martian friend did it, how probable is it that we'd find a broken window and a stone on the floor? Very high: the hypothesis predicts exactly that.

And the likelihood of hypothesis 2? Also high. The likelihoods are similar. The data don't distinguish between them.

And yet nobody hesitates for a second. Why? What makes the difference is not the likelihood but the prior: how probable is it that an elephant with a Martian friend is playing in your friend's garden, compared with a kid throwing stones? The evidence alone can't settle this. What settles it is what we knew before. And until we met inverse probabilities, everything we were doing in statistics was based on likelihood, on the probability of the data given a hypothesis. If what you want is the probability of a hypothesis, that's a fundamental gap in most of the statistics you were taught: it ignores how probable the hypotheses were to begin with.

So Bayesian inference has three parts:

- *Prior credence*, p(hypothesis): what I believed before.
- *Strength of the evidence*, p(data | hypothesis) / p(data): how much the data favor each hypothesis.
- *Posterior credence*, p(hypothesis | data): what I should believe now.

A doctor does this constantly: disease prevalence (prior) + test results (evidence) ⟶ diagnosis (posterior).

There's a price to pay, though. To do all this, we have to accept that our credence in a hypothesis can be represented by a probability. Nobody disputes the likelihood: if I imagine a coin with a probability of heads of 0.5, I can calculate the probability of getting ten heads in ten tosses. But saying that a *hypothesis* has a probability is something people do dispute. The statisticians known as frequentists rejected it. That's a legitimate position, and later we'll see what its defenders propose instead.

What do we get in exchange for paying that price? Correct reasoning with uncertain premises. Bayes' theorem doesn't give you good beliefs; it carries forward the ones you have. If your starting credences follow the rules of probability, and you update them with Bayes' theorem, your new credences will follow those rules too. Your beliefs will keep fitting together, or, as philosophers say, they will stay coherent. You won't end up holding beliefs that contradict each other.

And the new credences won't just be coherent: they will have taken the evidence into account in the right way, each piece with exactly the weight it deserves. It doesn't even matter in which order the evidence arrives;

you end up in the same place. In fact, if your credences are probabilities, Bayes' theorem is the only way of updating them that stays consistent. As in logic, there's no guarantee that your conclusions are true, but you will have reached them through valid inferences. We'll come back to that too.

All of this depends on accepting that credences should be probabilities at all. Some statisticians don't, and we'll come back to that.

The coin bet

Let's do a complete example. A friend proposes a game of chess, and to decide who plays white (white moves first, so it has an advantage), he tosses a coin: heads, he plays white. You're a bit suspicious, so you ask him to toss it three more times. Heads, heads, heads. Four heads in a row.

What's the probability that the coin is not fair?

A tempting reasoning: if the coin were fair, the probability of four heads would be $0.5^4 = 0.0625$, so the odds are 1 in 16 that the coin is fair. What do you think about this reasoning?

It's the inverse probability mistake again. 0.0625 is p(4 heads | fair coin), not p(fair coin | 4 heads).

Let's look at this with each of the approaches we've seen.

With a p-value: assume the coin is fair. Four heads in four tosses is not that weird; it happens more than once in twenty. We can't reject the hypothesis that the coin is fair. With five heads it's still not weird enough (two-sided $p \approx 0.06$); with six, it is ($p \approx 0.03$). But even then, all we can say is that we reject the fair coin. We never get a probability for anything.

With maximum likelihood: what probability of heads makes four heads out of four most probable? A coin that always lands heads. The maximum likelihood estimate of the probability of heads is 1.

With Bayes: first, we need to state the hypotheses and their priors. "Fair coin" is not enough: what are the alternatives? Let's try several options.

Option a: all coins are equally probable. Before going to my friend's house I have no idea, so I assume that any probability of heads between 0 and 1 is equally likely. That's a flat prior: it represents no prior knowledge. If we compute the posterior with this prior… it has exactly the same shape as the likelihood. Why? Because the prior is the same for every hypothesis, so it

doesn't change anything: the posterior is just the likelihood, normalized. The conclusion is the same as with maximum likelihood.

Option b: most coins are fair. Before going to your friend's house, do you really believe that a coin that always lands heads is as likely as a normal coin? If you had to bet 100 euros, would you bet on that? Of course not. Most coins in the world are pretty fair. Maybe not exactly 0.5, but close. So let's use a prior concentrated around 0.5. Now the likelihood is still the same (the data haven't changed), but the posterior has moved only slightly towards 1. With four heads, I start getting a bit suspicious: if I had to bet on the next toss, heads is now a bit more likely than tails. And if I keep getting heads, the posterior keeps moving towards "this is not a normal coin".

Option c: my friend loves Batman. Do you remember the Batman film with Two-Face? He tosses a coin to decide people's fates, and it always lands on the same side. Why? Because his coin has two heads. Now suppose my friend is a joker and a Batman fan. Then my hypotheses are different: either the coin is normal (probability of heads 0.5), or it's a two-headed joke coin (probability 1).

The evidence is the same in every case: four heads. Look at what happens with different priors:

Prior for "fair coin"	Posterior for "fair coin" after 4 heads
0.999 (most coins are fair)	≈ 0.98
0.5 (my friend is a joker)	≈ 0.06
0.2 (my friend is a huge Batman fan)	≈ 0.015

Same data, three very different conclusions. In one, the coin is almost certainly fair. In another, it's almost certain my friend is pulling my leg.

So which conclusion is right? What should a reasonable person think?

Are we allowed to have different truths?

Someone always says at this point: "Get more evidence! Look at the coin!" And that's a good idea, but often the evidence you have is the evidence you have. You've done an X-ray and seen a shadow. You can order a biopsy, but there's a waiting list. Do you start treatment or not? I've analyzed samples that were so valuable that there was no way to get more material. The conclusion depends on the data, but also on what you believed at the

beginning. And if two people believed different things at the beginning, they reach different conclusions.

This is one of the things people dislike most about Bayesian statistics. And it's not just theoretical. Imagine two reasonable people, Elena and Pablo, who look at the same patient. Pablo thinks: this person smokes a lot, so a priori it's quite probable that they have cancer. Elena thinks: given the prevalence for someone of that age, it's not very probable. They see the same X-ray, both reason correctly with Bayes' theorem, and they reach different conclusions. Does this mean that each of them is entitled to their own truth?

When I ask this in class, most people say: yes, I suppose each of them has their own truth. But here's the thing. The patient either has cancer or doesn't. That's one question. A different question is what Elena, a reasonable person, *should think*, given her evidence and what she believed before. Those are two different things, and we've seen this before. Remember: *the map is not the territory*. In the world, the territory, the patient has cancer or doesn't. In Elena's head, the map, there's a degree of confidence. Bayesian inference doesn't tell Elena what's true about the world. It tells her how her map should be drawn and how it should be updated. Whether the map is right is a different matter.

Some practical consequences

The posterior is the prior multiplied by the strength of the evidence. That simple fact has some very practical consequences.

Look for the evidence that separates your hypotheses. Not all data are equally useful. Data that are just as probable under every hypothesis can't change anything: if both a cold and the flu make you cough, coughing doesn't help you choose between them. What moves the posterior is evidence that is much more probable under one hypothesis than under the others. Go back to the coin: each head is only twice as probable with the two-headed coin as with a fair one, so the heads nudge us a little at a time. But a single tail is impossible with the two-headed coin, and it settles the question at once. So when you design an experiment or order a test, ask yourself: which result would look very different depending on which hypothesis is true?

If your priors are strong, you'll need strong evidence to change them. Before any experiment, what probability do you give to the hypothesis that my tap

water cures any disease? Very low. So you shouldn't accept the usual level of evidence. Drinking the water and not dying is not enough; you should require a huge amount of evidence. As Carl Sagan used to say: *extraordinary claims require extraordinary evidence.* A rational person does exactly that.

If your priors are very strong, the evidence may not change your mind at all. If you start out very convinced of something, it will be very hard for any evidence to move you. That's not irrational in itself, but it's something you should know about yourself.

And in the extreme, it becomes impossible. If you give a hypothesis a prior probability of exactly zero, no amount of evidence will ever change your mind: the posterior is the prior multiplied by the strength of the evidence, and zero times anything is zero. The same happens with a prior of one, because being completely sure that a hypothesis is true means giving zero to all the others. The statistician Dennis Lindley called this Cromwell's rule, after Oliver Cromwell's plea to the Church of Scotland: "I beseech you, in the bowels of Christ, think it possible you may be mistaken." Never give a prior of exactly 0 or 1 to anything that isn't logically impossible or logically certain. You can say that a hypothesis is very improbable a priori. But how improbable? Depending on how small, you might send someone to prison or not.

The frequentist alternative

Not everybody likes Bayesian inference

If Bayesian statistics is so nice, why are students taught mainly methods based on p(data | model), like those built on the p-value? Everybody accepts Bayes' theorem:

$$p(A \mid B) = p(A) \cdot p(B \mid A) / p(B)$$

That's mathematics, and nobody disputes it. Bayesian inference is another matter:

$$p(\text{hypothesis} \mid \text{data}) = p(\text{hypothesis}) \cdot p(\text{data} \mid \text{hypothesis}) / p(\text{data})$$

Why was Bayesian inference rejected for so long by many principled statisticians? There were two main philosophical problems:

1. *What is probability?* Does it make sense to talk about the probability of a hypothesis, or of rain tomorrow? Is probability a property of the world, or of our knowledge about the world?
2. *Priors are subjective.* If my conclusions depend on my priors, and my priors are mine, where's the objectivity of science?

There was also a practical problem: Bayesian inference usually needs a lot of computation, and until the 1980s computers were not up to the task.

Probabilities are frequencies

In the 19th century, some mathematicians and philosophers, like Robert Leslie Ellis, George Boole and John Venn, criticized the way Laplace assigned probabilities to hypotheses. For them, a probability should be a frequency. Roughly from the 1840s until the 1980s, statistics was dominated by what we now call the frequentist approach.

What does it mean that the probability of a coin landing heads is 0.5? It sounds like an easy question. In 1866 John Venn gave the classic answer: a probability is the proportion of times something happens in the long run, if we could repeat the experiment indefinitely. Toss the coin many, many times, an infinite number of times if you could, and the probability of heads is the proportion of heads you'd get. This is the *frequentist interpretation*. In it, probability is an objective property of the world: of the coin, of the die, of the population.

It has a clear consequence. What's the probability of a hypothesis being true, for instance the null hypothesis? For a frequentist, the question makes no sense. The hypothesis is either true or false; there's no long run of repetitions in which it's sometimes true and sometimes not. That's why, after a significance test, a frequentist won't tell you how much your confidence in the null hypothesis, or in the alternative, should change.

Now you can see why frequentists are happy to reject hypotheses, but not to increase or decrease our degree of belief in them. For a frequentist, it makes no sense to ask how probable a hypothesis is. Tomorrow it will rain or it won't; your patient has the condition or doesn't; the hypothesis is true or it isn't. Probability is not about our hypotheses or about what we believe.

The biometricians

At the end of the 19th century, a group of researchers set out to study heredity and evolution by measuring variation in whole populations: Francis Galton (Darwin's cousin), Karl Pearson and Walter Weldon, among others. They invented many tools that are still widely used: correlation, regression and, in 1900, Pearson's chi-square test.

Their aims were mostly descriptive, and their samples were large: thousands of individuals. With so much data, the prior hardly mattered to them in practice. In fact, Karl Pearson wasn't against Bayesian reasoning at all. He accepted inverse probability and used it himself, for instance in a 1920 paper on what he called "the fundamental problem of practical statistics". The real attack on priors would come later, from Fisher.

Priors are not objective

The second problem was the priors. If you want to say how probable a hypothesis is, you need a prior; Bayes' theorem leaves no way around that.

What's the prior probability that Barça will beat Madrid in their next match? Or that a new drug, before any experiment, will cure a disease? What's the probability that somebody is a killer, before we have any evidence? Ask experts with the same information, and they'll give you different numbers. Nobody has found a rule that settles which of them is right. If my conclusions depend on my priors, and my priors are mine, where's the objectivity of science?

Priors can also be dangerous, because base rates can carry prejudice. Imagine using the proportion of a minority in a country's prisons as the "prior" probability that a young man from that minority is a criminal. Those base rates are themselves the product of a history of discrimination, and using them as priors can disguise our biases as rational inputs. A prior is a premise, and you have to be able to defend it.

Fisher

Ronald Fisher (1890–1962) was probably the most influential statistician of the 20th century. He developed likelihood and maximum likelihood, analysis of variance (ANOVA), the principles of experimental design, such as randomization, and he popularized the p-value.

Fisher also made key contributions to biology. Until 1918, many people thought that Mendel's genetics and Darwin's evolution were incompatible. Mendelian genetics says that inheritance is discrete: Mendel saw yellow peas and green peas, nothing in between. But Darwinian evolution needs continuous variation, all the intermediate shades, to move gradually from one form to another. Fisher's 1918 paper showed that if a trait is controlled by many Mendelian genes, each with a small effect, you get continuous variation. That paper reconciled Mendel and Darwin, and from it came both population genetics and quantitative genetics.

He did something just as important. He took these tools, his own and other people's, and spread them through the whole scientific community: William Gosset's t-test (published under the pseudonym "Student"), Pearson's chi-square test, Galton's regression. Fisher worked at an agricultural research station (Rothamsted), among scientists who needed practical methods rather than mathematical theory. He wrote two books for them: *Statistical Methods for Research Workers* (1925) and *The Design of Experiments* (1935). They're full of worked examples and contain very little mathematics. They were enormously successful, and they carried his practical approach, and

his frequentist perspective, to biologists, agronomists, psychologists and doctors all over the world. Then statisticians and textbook writers took those books, often stripped away Fisher's own nuances, and turned them into a dogma: the frequentist orthodoxy that's still taught in many classes.

And Fisher was firmly against Bayesian inference. In *Statistical Methods for Research Workers* he wrote that "the theory of inverse probability is founded upon an error, and must be wholly rejected". Fisher's authority, together with Neyman and Pearson's framework, which had no place at all for probabilities of hypotheses, made Bayesian inference a minority position in academic statistics for half a century.

Neyman and Pearson needed no Bayes

Does science need probabilities of hypotheses at all? Frequentists say no. What science needs, in their view, is procedures that rarely lead us astray, and we can judge a procedure by frequencies: if I follow it, how often will I be right, and how often wrong? Jerzy Neyman and Egon Pearson formalized this way of thinking. Instead of asking how sure you are about this particular hypothesis, you set up a procedure, with a threshold for choosing to do one thing or another.

For instance, you create a set of rules for treating patients according to their symptoms and the results of their tests, and then you check in how many cases the rules gave the right answer. We can apply the same way of thinking even to the justice system. A court can't know for sure whether this particular defendant is guilty, and it doesn't need to put a probability on it. What it does is follow a rule, "beyond reasonable doubt", designed so that, over thousands of trials, few innocent people end up in prison. Neyman called this *inductive behaviour*. The question is not "what should I believe?" but "what rule should I follow, knowing that I'll apply it again and again?"

In Neyman and Pearson's approach you are allowed to accept a hypothesis: not as a belief, but as a decision to act as if it were true, knowing how often that kind of decision goes wrong. The doctor doesn't need to know the probability that the patient has cancer to follow a rule that, across all patients, keeps both kinds of mistakes under control.

Doctors already use rules designed from the point of view of the system. Screening thresholds, reference values for a blood test, clinical guidelines: all of them were chosen for how they perform across many patients, how

many sick people they catch and how many healthy people they send for unnecessary tests. When a doctor follows a guideline, they're applying a Neyman-style rule, even if they think about their patient in a Bayesian way.

Courts sometimes choose the point of view of the system on purpose. The presumption of innocence asks the judge to set aside any prior about the kind of people who end up in the dock. That may look like throwing information away, but there's a good reason for it. As we saw, base rates can carry prejudice, and a prior can be a thin disguise for it: "people like him are usually guilty" sounds like a probability, but it may be nothing more than a bias with a number attached. A legal system may prefer a procedure that protects the innocent, whatever the judge's priors are.

The philosopher Deborah Mayo has developed Neyman's ideas further, and connected them with Popper. For her, a claim is well supported when it has passed a *severe test*, one that would very probably have failed if the claim were false. Passing an easy test means nothing; passing a hard one means a lot.

This view has a real strength: it explains some of the problems we saw with the p-value. Why is p-hacking wrong? If you try many analyses, or keep collecting data until $p < 0.05$, the result you finally report may look exactly like an honest one. What has changed is the procedure: it would have "found" an effect even if there was none. The test was not severe, and the error rates you claim are not the real ones. Error statistics condemns this directly. A strict Bayesian, who only looks at the data actually observed, has to work harder to explain why the garden of forking paths is a problem.

The Bayesian defense

Academic Bayesians

Bayesian reasoning was a minority position in academia, but it never really disappeared. A few statisticians and philosophers kept discussing and developing it. Harold Jeffreys, a geophysicist, argued with Fisher in print and published his *Theory of Probability* in 1939. Frank Ramsey and Bruno de Finetti worked out what it could mean for a probability to be a degree of belief. And in the 1950s and 60s, Leonard Savage, Dennis Lindley and I. J. Good, among others, laid the foundations of modern Bayesian statistics.

Jeffreys tried to answer the subjectivity objection directly. He believed in objective priors: rational people with the same knowledge should agree on the prior, and the task was to find the rules that give it. Few people defend that ambitious program today, for the reason we saw: experts with the same information give different numbers. But priors chosen by conventional rules are still widely used, and they're often called "objective" priors.

Bayesians also proposed another interpretation of probability. In 1926 Frank Ramsey, a young Cambridge philosopher, defined probabilities as *degrees of belief*. When I say that the probability of my patient having the condition, or of rain tomorrow, is 0.9, I'm describing the state of my knowledge, not my patient or the clouds. But how can you measure a degree of belief? Ramsey's answer: by the bets you'd accept. If you'd pay up to 90 cents, and no more, for a ticket that pays one euro if it rains, your degree of belief is 0.9.

The Italian mathematician Bruno de Finetti, who worked for years as an actuary, took this idea all the way. "Probability does not exist", he wrote. It isn't out there in the world; it's only in our heads. Probability is about the map, not about the territory. He also showed why degrees of belief should

follow the rules of probability. If they don't, someone could offer you a set of bets that each look fair to you, but that together make you lose whatever happens. In the betting world this is called a *Dutch book*. Here's a simple case. Take a cup final, where someone has to win, and suppose you're 90% sure that Madrid will win, and also 90% sure that Barça will. Then you should be happy to pay 90 cents for a ticket that pays one euro if Madrid wins, and another 90 cents for a ticket that pays one euro if Barça wins. You've paid 1.80 euros, and whatever happens, exactly one of the tickets pays: you get one euro back. You lose 80 cents for sure. If you're 90% sure that Madrid will win, you have to be 10% sure that Barça will. That's the coherence we met with Bayesian inference. Notice that coherence doesn't single out one set of beliefs: many different sets can be coherent. It's coherent to think that Madrid will win with a probability of 0.9 and Barça with a probability of 0.1, or the other way around. What you can't do is believe that both will win with a probability of 0.9.

So for the frequentists, probability lives in the territory: it's a property of coins and populations. For the Bayesians, it lives in the map: in the reasoner, describing what they know. This map-versus-territory business, which looked like a rather peculiar distinction, is at the heart of the fight between statisticians. Some say statistics is about the world. Others say statistics is about what we should think about the world. That's pure philosophy of statistics, and it has fuelled many arguments between professionals.

Frequentists are not objective after all

The big objection to Bayesian inference is that priors are subjective. Frequentists tried to escape from that explicit subjectivity, but one could say that they just swept it under the rug. For instance, frequentists like to see their error rates as objective. But error rates also depend on choices: the model and its assumptions, the significance threshold (a convention), and even the analysis plan, that is, what you would have done if the data had been different.

Both approaches rest on choices. Frequentists make theirs about the model and the procedure; Bayesians also make them about how credible the hypotheses are. The statisticians Andrew Gelman and Christian Hennig have proposed dropping the words "subjective" and "objective" altogether. Instead, we should ask for the virtues we actually want, such as transparency (state your assumptions openly), consensus (could other people

agree with them?) and awareness of how much your conclusions depend on those choices.

Hypotheses aren't objective

In statistical inference, even in the frequentist approach, we're often supposed to start with a hypothesis, or a model, and then look at the data. But where did the hypothesis come from? We tend to assume that hypotheses are built from the data, but in science that's often not what happens.

For centuries, astronomers could predict the positions of the planets with great precision without any physical hypothesis about what the planets were doing. Astronomy was considered a branch of mathematics, not of physics: physics was about what the world is really like, while mathematics only described it. Astronomers had mathematical models, circles on top of circles, and they weren't very interested in anything more. When Copernicus's book was published in 1543, it came with a preface written not by Copernicus but by the theologian Andreas Osiander. The preface presented the model in exactly this way: as a tool for calculation, not as a description of reality. The Earth didn't have to really move; the formulas were just easier to use. Then Galileo caused an uproar by insisting that the Earth really does move. For him, it was a physical hypothesis.

Kepler's story shows how far a hypothesis can be from the data. The data were Tycho Brahe's, the most precise astronomical observations of the time. Tycho was a colourful character. He had lost part of his nose in a duel and wore a metal replacement. He also kept a tame elk that, the story goes, drank too much beer at a dinner, fell down the stairs and died. He hired Kepler to analyse his observations, but he guarded them jealously, perhaps afraid that Kepler would draw the great conclusions and get the credit. When Tycho died in 1601, Kepler made sure the data didn't escape him. In his own words: "I confess that when Tycho died, I quickly took advantage of the absence, or lack of circumspection, of the heirs, by taking the observations under my care, or perhaps usurping them."

But Kepler didn't read the ellipse off the data. He was a Platonist, and for a Platonist the circle is the perfect motion, so the planets had to move in circles. He spent years trying one combination of circles after another, and none of them fitted Tycho's observations of Mars. Only then did he try other shapes, ovals first, until he found the ellipse. In the end the evidence did

win. But the winning hypothesis didn't come out of the data; it came from Kepler, once his favourite one had failed.

Archimedes described his own method like this: get your idea any way you can, even with rough mechanical reasoning, and then prove it rigorously; if the proof works, it doesn't matter where the idea came from. (The text where he explains this, *The Method*, was lost for centuries. It turned up in 1906 in a palimpsest: a medieval prayer book written over an older copy of his work.) And the chemist August Kekulé claimed that the ring structure of benzene came to him in a dream of a snake biting its own tail.

Philosophers call this the difference between the *context of discovery* and the *context of justification*. Discovery is how we come up with a hypothesis: a hunch, a dream, a Platonist prejudice, a failed attempt. Justification is how we test it against the data. As long as a hypothesis passes the test, we don't worry much about where it came from.

A frequentist could reply: "Exactly. I don't mind where your hypothesis came from, because my test judges it. But your prior doesn't stay in the context of discovery: it enters the verdict itself." That's a fair point. An objective justification would be great, if it were possible. But, as we've just seen, frequentist justification also rests on choices. And even a perfectly objective test wouldn't solve the main problem, because a test can only judge the hypotheses we chose to test. By choosing some hypotheses we ignore all the others, and that's often where science goes wrong. A trial may conclude that a drug doesn't work when it was only tested in the wrong population: maybe it works in elderly women, and nobody looked there. The test was impeccable; the choice of hypothesis wasn't.

And frequentists choose their hypotheses, their models and their assumptions just like Bayesians do; as we saw, models don't come from data alone. So the objectivity of frequentist statistics is not as complete as it sounds, and Bayesian subjectivity is not as bad as it sounds.

Enough evidence wins

Fortunately, there's a way out, at least sometimes. Given enough evidence and the same hypotheses, two people reasoning with Bayes' theorem will end up with the same conclusions, whatever their initial priors.

Go back to the coin. With four heads, the person who started at 0.999 for the fair coin still thought it was fair. But if the friend keeps tossing and gets

fifteen heads in a row, even that person's posterior for the fair coin drops to about 0.03. Everybody now agrees that something is going on. With enough data, the priors are washed out.

This is a deeper property than it seems. Even if we start with a wrong map, with priors that are far from reality, as long as they're coherent and don't rule out the truth with a prior of zero, updating with the data lets the territory reshape the map. Statistics done well, Bayesian or frequentist, lets the evidence correct our starting point. That's one of the great strengths of these tools.

But reshaping the map has a price: data. The further our starting map is from the territory, the more evidence we need to reach the same conclusion. In the coin example, the person who started at 0.999 for the fair coin needed fifteen heads in a row to get down to about 0.03. Someone who had started at 0.5 would have got there with only five. What costs the most data is a prior that is confidently wrong: a vague prior, like 0.5, costs little even when it's off, while a firm one, like 0.999, costs a lot. So good priors are not just a matter of principle. A prior that represents the territory well gets us to good conclusions with less data, and in science data cost time and money.

The flip side is that priors matter a lot when we don't have much data. And that's useful to know. Being aware that our data are weak, that they don't favor one hypothesis over another, is valuable information in itself.

Will it rain tomorrow?

For an engineer designing a production process, or for someone managing a health system or a justice system, Neyman's question is the natural one: across thousands of cases, how many mistakes will my rule make? But think of patients and doctors. For the doctor, the natural question is "how probable is it that this person is ill?". It would be very strange for a doctor to treat a patient without thinking about how likely it is that this particular patient has the condition, and to think only "I'm following a procedure that, across all my patients, has a type I error rate of 5% and a type II error rate of 10%". The same goes for a judge facing one particular defendant: what the judge has to decide is whether this person is guilty.

We could say that the Bayesian approach takes the point of view of the case, and the frequentist approach the point of view of the system. Thinking about the justice system as a whole in Neyman's way makes perfect sense;

asking a particular judge to think that way doesn't. And if you want to reach a conclusion about the probability of a hypothesis, there's no way around it: Bayesian inference is the only consistent way. We just have to accept that credences can be represented by probabilities, and then Bayesian inference gives us consistent probabilities for hypotheses.

Bayes works

Outside academia, some people were using Bayesian ideas for practical purposes, even when those ideas were unpopular in universities. Two examples: money and war.

Insurance actuaries needed to put prices on risks, and they didn't much care about the academic debates. Early in the 20th century, American actuaries developed what they called *credibility theory*. To price a policy for a particular company, they combined the company's own claims history with the experience of the whole industry, giving more weight to the company's own data the more of it there was. That's Bayesian reasoning in all but name: a prior from the industry, updated with the company's data. Later, actuaries showed that it was indeed Bayesian.

During the Second World War, the German armed forces encrypted their messages with the Enigma machine. Polish mathematicians had already made remarkable progress breaking earlier versions of Enigma, and shortly before the war they shared what they knew with the British and the French. But the settings changed every day, and without them the messages were gibberish. At Bletchley Park, north of London, the British gathered a team of mathematicians, linguists and crossword enthusiasts to break the code.

One of them was Alan Turing. In mathematics, Turing is a giant: in 1936 he had defined, in the abstract, what a computing machine is, laying the theoretical foundations of computer science. At Bletchley he designed the electromechanical machines (the *bombes*) used to find the Enigma settings, and he developed statistical methods to decide which hypotheses about the settings were worth testing. Those methods were Bayesian: he weighed the evidence for each hypothesis, and even invented a unit to measure the strength of evidence, the *ban*. They worked, and the intelligence obtained was crucial, for instance, in the battle against the U-boats in the Atlantic. (Bletchley Park has been restored and is now a museum.) But all of this was a state secret, and remained so for decades, until the 1970s.

The end of Turing's story is a sad one. Turing was gay, and he didn't hide it. In 1952 he was prosecuted for "gross indecency" and, to avoid prison, accepted chemical castration with hormone injections. Nobody knew then what he had done for his country. In 1954, two weeks before his 42nd birthday, he was found dead of cyanide poisoning, with a half-eaten apple by his bed. The apple was never tested, and it's still unclear whether it was suicide or an accident (he did chemistry experiments at home). He was fascinated by the scene in Disney's *Snow White* where the witch poisons the apple. One of the most important mathematicians of the 20th century, a key figure in winning the war, destroyed by the people he had served. The British government officially apologized in 2009, and he received a royal pardon in 2013.

You've probably heard that the bitten-apple logo of a certain computer company is a homage to Turing. It's a lovely story. Unfortunately, the designer of the logo has said that it isn't true, and the company never claimed it. The writer and actor Stephen Fry once asked Steve Jobs about it, and, as Fry tells it, Jobs answered: "It isn't true, but God, we wish it were." It's the kind of legend that deserves to exist.

A pragmatic truce

Different questions

So what should we do? Which approach should we accept? Should we reject Bayesianism or frequentism?

In the last decades, more and more statisticians have taken a pragmatic position: use the tool that fits the question, and understand what each one does. We could accept that both interpretations of probability are legitimate, and think of the two schools as approaches to different questions. The statistician Richard Royall pointed out that when we look at data, we may be asking three different questions:

1. *What do the data say?* How much do they favor one hypothesis over another? You need only the data and the hypotheses.
2. *What should I believe?* That's Bayes: you also need a prior.
3. *What should I do?* That's Neyman and Pearson: you also need error rates, or the costs of each kind of mistake.

The first question is the one asked by the likelihoodists, Royall among them. They measure the evidence with the likelihood ratio. Go back to the chess coin and its four heads. A two-headed coin gives that result with probability 1; a fair coin, with probability 0.0625. The ratio is 16: the data favor the two-headed coin over the fair one by a factor of 16. That number is the same for everybody, whatever they believed before.

Bayesians agree with this way of measuring the evidence. With two hypotheses, Bayes' theorem says exactly that: the odds after seeing the data are the odds before, multiplied by this ratio. So the answer to the first question is part of the answer to the second: the Bayesian takes the evidence

and multiplies it by a prior. And if the question is "what should I believe?", the most direct way to answer it is a Bayesian analysis.

Each school answers its own question well, and many of the fights between statisticians come from answering one question with another school's tool. The null ritual is the extreme case: it runs a decision procedure, built for the third question, and then reads the result as evidence and as belief: "$p < 0.05$, the drug works".

Moreover, classical and Bayesian methods often lead to the same conclusions in practice. That tends to happen when:

- you have plenty of data,
- the result is clear, not borderline,
- your hypotheses are well defined (for instance, a null and an alternative), and ideally the alternative states the size of the effect you care about,
- the hypotheses are roughly equally probable a priori,
- you do one test, or correct properly for multiple testing and avoid p-hacking,
- and you understand what you're doing: the p-value is not the probability of any hypothesis, and even a tiny p-value says nothing about how large the effect is.

The "not borderline" matters. You might think that with lots of data the two approaches always agree, but there's a well-known exception. With a huge sample, a p-value just under 0.05 corresponds to a tiny effect. The p-value says "reject the null", while a Bayesian analysis of the same data says that the null has gained support. This is known as Lindley's paradox. With lots of data, look at how large the effect is, not only at whether p crossed a threshold.

Frequentists also use priors

Frequentists also use priors and Bayes' theorem, in some cases. When a base rate exists, there's no disagreement at all. Think of the prevalence in the HIV test. Both schools put the same number into the same formula, Bayes' theorem. The Bayesian calls it a prior; the frequentist calls it a base rate, and insists that it must be a measured frequency. But they're using the same mathematics, the same theory and the same numbers.

We can see this in the Neyman–Pearson approach. Its two error rates are each calculated assuming a state of the world: α is the probability of a

positive result *if* the patient is healthy, and β the probability of a negative result *if* the patient is sick. Neither needs the prevalence. These error rates describe the procedure, not a particular result. So Neyman can guarantee that only a small proportion of the healthy people get treated, and only a small proportion of the sick people get missed. But to know how many mistakes a hospital will make in total, you need to know how many sick people there are: α times the number of healthy patients plus β times the number of sick ones. You need the prevalence, which is a prior by another name.

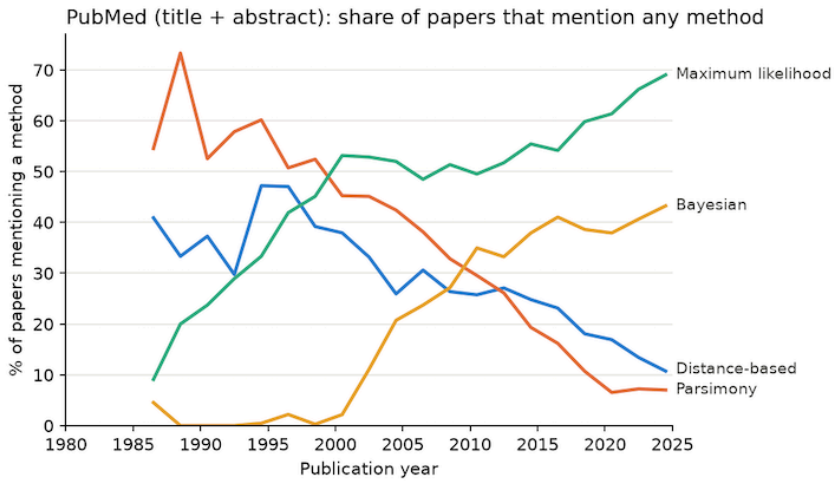
The real disagreement between frequentists and Bayesians is limited to hypotheses with no frequency behind them, like Sally Clark's guilt or the truth of a scientific theory, such as general relativity. There, the frequentist refuses to give a number, and the Bayesian has to give one. For a Bayesian, the problem with leaving the prior out is this: a field can use $\alpha = 0.05$ faithfully and still have most of its significant findings be false, if most of the hypotheses it tests were unlikely to begin with. That's part of what is happening in the replication crisis.

Bayes had to wait for computers

Even if both approaches are useful, for a long time only one of them could be used in practice. For most real problems, the posterior can't be calculated by hand. The denominator, p(data), requires adding up over all the possible hypotheses, and with models that have many parameters, that's a calculation nobody can do on paper. For most of the 20th century, the debate between Bayesians and frequentists was largely a discussion among statisticians and philosophers, with few practical consequences: Bayesian analyses of realistic problems were simply out of reach. Even maximum likelihood, except in a few simple cases, was too much without computers.

The key algorithms existed long before they were used. The Metropolis algorithm, a way of exploring a probability distribution with a kind of guided random walk, was published in 1953 by physicists at Los Alamos, and W. K. Hastings generalized it in 1970. But it was only in 1990, when Alan Gelfand and Adrian Smith showed what these simulation methods could do for statistics, at a time when computers were becoming cheap, that Bayesian statistics became practical. These methods, known as Markov chain Monte Carlo (MCMC), are what phylogenetic programs like MrBayes run when they give you the posterior probability of a tree.

You can see this history in my own field, in how phylogenetic trees have been built. In the 1980s, almost everybody used distance methods and parsimony, which are fast. Maximum likelihood trees only became common in the 1990s, when computers could handle them. Bayesian trees barely existed until around 2000. And universities change slowly: methods that only became practical a generation ago are still making their way into statistics courses. That's part of the answer to the question we asked earlier: why you were taught the p-value instead.



Pluralism and craft

I was a Bayesian

For years I considered myself a Bayesian, and my main reason was that I rejected the way p-values are normally used. I still reject that way of using p-values, but I now realize that, as a criticism of frequentist statistics, it was unfair: a strawman. As we saw, neither Fisher nor Neyman would have accepted the null ritual, and principled frequentists today, like Deborah Mayo, reject it too. Criticizing frequentist statistics because of the ritual is like criticizing Bayesian statistics because someone could choose any prior they liked and get the answer they wanted. Principled Bayesians don't defend arbitrary priors, and principled frequentists don't defend the ritual.

What about the frequentist case against the Bayesians? It's really two objections. The first says that degrees of belief can't be represented by probabilities. Let's see why, as a formal argument, that doesn't hold.

Kolmogorov: a formal theory of probability

In 1933 the Russian mathematician Andrey Kolmogorov did something that, in hindsight, makes much of the fight between the two schools look less necessary. In his formulation, a probability is any function that assigns numbers to events and follows three rules:

1. Probabilities are never negative.
2. The probability that something happens is 1.
3. If two events can't happen at the same time, the probability of "one or the other" is the sum of their probabilities. For instance, heads and tails in a single toss: p(heads or tails) = p(heads) + p(tails).

Everything else, Bayes' theorem included, follows from these rules. And here's the important part: Kolmogorov didn't say what probability *is*. He built a mathematical structure, and anything that follows those rules can be studied with it. The theory didn't deal with what probability means or how to use it in science. Like logic, as we'll see next, it's a *formal* theory: it tells us how probabilities must relate to each other, not what they mean.

Long-run frequencies follow Kolmogorov's three rules. Coherent degrees of belief follow them too; that's exactly what de Finetti's betting argument shows. So both are legitimate uses of the same mathematics. And that's my position: whatever behaves according to Kolmogorov's rules can be represented with probabilities. Frequencies can, and so can coherent degrees of belief. This position is a kind of pluralism about the interpretation of probability.

The second objection says that degrees of belief can't be measured, or at least not objectively. That one survives, not as a knockdown argument but as a practical caution. Ramsey's bets measure beliefs only roughly, and a prior can be perfectly coherent and badly wrong, or a thin disguise for prejudice. A prior is already a map, and we should try to make it a good one, by measuring it when we can and by justifying it openly when we can't. But there's no guarantee that two reasonable people will start from the same map. In general, priors can't be objective.

So we can't have completely objective priors. But that doesn't mean that any prior will do. If we want our conclusions to resemble the territory, our priors have to represent the territory well enough: a prevalence measured in the right population, the results of previous experiments, what we know about how the world works.

Probability as logic

Bayesian inference can also be seen as a generalized form of logical inference. Logic is the discipline that studies the rules of valid reasoning. Aristotle invented it, more or less on his own. He wanted a way to tell sound reasoning from the clever but empty arguments of the sophists, the professional debaters of his time. An argument is *valid* if, once we accept the premises as true, the conclusion necessarily follows:

1. All people are mortal.
2. Socrates is a person.

3. So Socrates is mortal.

Validity is a property of the argument's form, not of its content. "If P, then Q; P; therefore Q" is valid whatever P and Q are. Truth, on the other hand, is a property of each individual statement. The central question of logic is "does the conclusion follow from the premises?", not "are the premises true?".

That's why validity doesn't guarantee truth. This argument is perfectly valid:

1. All cats can fly.
2. Luna is a cat.
3. So Luna can fly.

Validity preserves truth, but only if the truth was already there in the premises. And the opposite also happens: an invalid argument can have a true conclusion.

1. Socrates is mortal.
2. Plato is mortal.
3. Aristotle is mortal.
4. So all people are mortal.

The conclusion is true, but this argument doesn't justify it. And notice what kind of argument this is: an induction, from some cases to all of them. As we saw with Hume and the sunrise, inductions are not logically valid. But science can't live without them, because we need conclusions that go beyond the data we have.

So Bayesian inference can be seen as a logic for reasoning with uncertain premises. The premises are the hypotheses, their prior credences and the evidence; the conclusions are the posterior credences. In 1946 the physicist Richard Cox showed that reasoning with degrees of plausibility has to follow the rules of probability, if it meets a few basic requirements of consistency. The physicist Edwin Jaynes later built a whole approach to statistics on that result, and called it *probability theory: the logic of science*. If your degrees of belief don't follow these rules, your reasoning is inconsistent, in the same way that concluding "P" from "if P, then Q" and "Q" is invalid.

Coherence and truth

But following Kolmogorov's rules makes a set of numbers coherent, not necessarily true. A degree of belief in "Sally Clark is guilty" can follow them

perfectly and still be a terrible description of the world. This is, in fact, the strongest version of the frequentist objection: satisfying the axioms is not enough; there has to be something in the world that the numbers answer to.

Remember Elena and Pablo. They looked at the same patient's X-ray and both reasoned correctly with Bayes' theorem, but they reached different conclusions because they started from different priors. Is each of them entitled to their own truth? To answer that, we need to say what truth is.

Aristotle gave the classic definition of truth: an idea is true when it corresponds to what is in the world. "The patient has cancer" is true if the patient has cancer. It sounds trivial, but it's not; it's called the *correspondence theory of truth*. The world itself is neither true nor false; the world just *is*. What's true or false are our beliefs, the things in our heads: they're true when they match the world, and false when they don't. Magritte's pipe again: the painting is a good or bad painting of a pipe, but it isn't a pipe.

So *valid conclusions and truth are not the same thing*. Bayesian inference, like deductive logic, is a formal machine: it guarantees that your conclusions follow correctly from your premises (your hypotheses, your priors, your evidence). It says nothing about whether your premises are right. We're allowed to have different valid conclusions; we're not allowed to have different truths. Truth depends on reality, not on my priors.

There is an alternative to Aristotle's theory: the *coherence theory of truth*, in which a belief is true when it fits well with the rest of our beliefs. The problem with this theory is that it isn't grounded in reality. A good novel can be perfectly coherent and completely false. A map can be perfectly consistent, every road connecting nicely with every other, and still not match the territory. Bayesian inference guarantees coherence, and that's valuable, but it's not enough. It's the same problem we'd have if we tried to study the world with deductive logic alone. Logic guarantees that true premises lead to true conclusions. But if you start from premises that don't describe the real world, a perfectly valid argument can take you to a perfectly wrong conclusion.

If we want our conclusions to correspond to the world, coherence is not enough. If we want our maps to resemble the territory, we need premises that represent the world well enough: good models, priors that are not too far from reality, and good evidence. And it's even better when the evidence

is strong enough to wash away our priors: then even unrealistic priors lead to approximately true conclusions.

Be careful: you can't be objective

Eugenics and genetics

The biometricians we met earlier also had a far less presentable side. Galton and Pearson were openly racist. Much of their measuring aimed to demonstrate the superiority of some groups over others: measuring skulls, for example, to show that well-off white Englishmen had better heads than everybody else. Galton coined the word *eugenics* in 1883. And the link between the new statistics and eugenics was institutional, not just personal. In 1911 Pearson became the first Galton Professor of Eugenics at University College London. There, his department of applied statistics grew up alongside Galton's eugenics laboratory. Modern statistics was born, in part, in the same institutions as eugenics, and was led by the same people.

I'm a geneticist, and we geneticists are often reminded of eugenics, with reason. Many prominent early geneticists supported it: Charles Davenport ran the Eugenics Record Office in the United States. And eugenics was not only a right-wing idea: the Nobel laureate Hermann Muller, a geneticist and committed socialist, defended his own version of it. Others, like Thomas Hunt Morgan, J. B. S. Haldane and Lancelot Hogben, were among its early critics. But geneticists were not alone. As we've seen, the founders of modern statistics were at the centre of the movement. And Fisher, a eugenicist, a geneticist and a statistician all at once, shows how both disciplines contributed to these awful ideas.

Why tell this story in a book about statistics? Not because it makes their methods wrong. Correlation, regression and analysis of variance work just as well whoever invented them, and Fisher was one of the most brilliant scientists of the 20th century. The lesson is about people, and also about methods. These were among the most rigorous scientists of their time. Objectivity was their ideal; in fact, it was one of the main reasons why Fisher rejected Bayesian priors. And they were convinced that they were being objective. Yet they were deeply biased, especially when they reasoned about matters outside their own speciality, like race and class. There they shared the prejudices of their time and of their social circle.

Fisher gives us another example, from later in his life. In the 1950s, when the first large studies linked smoking to lung cancer, Fisher argued publicly that the link hadn't been shown to be causal: perhaps the same genes that made people want to smoke also made them prone to cancer. It was a clever argument, and in principle a legitimate question. But Fisher was a keen pipe smoker, and he worked as a consultant to the tobacco industry. Here was the greatest statistician of his time, convinced that he was being rigorous, defending the wrong side of one of the most important public health questions of the century.

Fisher was extraordinarily smart. But he was a man of his time, with the racism and eugenic ideas common among people like him, and with his own personal biases, like all of us. That's the uncomfortable lesson: intelligence and rigour don't protect us from bias. It's the same warning we met with priors, where a premise that looks like a neutral number can be a thin disguise for prejudice. And it applies to all of us: we are all biased.

No methodology will save us from this. These people had the best statistical tools of their time, many of which they had invented themselves, and the tools didn't protect them. You can have a magnificent toolshed, but tools only produce excellent work in the hands of a good artisan. You need the craft. No tool creates great work on its own.

So what does help? In my view, the best protection we have is dialogue with people who don't think like us. A real dialogue, in which we're willing to listen to people we disagree with, and to learn from them. Our own biases are very hard to see from the inside, and they're often obvious to someone who doesn't share them.

Listening is not easy

Two of the most important statisticians of the 20th century, Fisher and Karl Pearson, hated each other. Why? When Fisher submitted his great 1918 genetics paper to the Royal Society, it wasn't accepted, allegedly because of bad reviews from Pearson. Reviews are anonymous, but these things always come out, and Fisher never forgave him.

Fisher's revenge was exquisite. Pearson's greatest contribution to statistics was the chi-square test. So Fisher studied it carefully and found that Pearson had got it wrong: he hadn't correctly accounted for the degrees of freedom. Imagine: someone comes and tells the world that the inventor of the chi-

square doesn't know how to calculate the chi-square. After that, they never spoke again.

They were both racists and both prominent in the eugenics movement. Galton founded eugenics, the idea that the "unfit" (the poor, for instance) should be discouraged or prevented from having children, so that society would improve. Pearson founded the journal *Annals of Eugenics*. As a Cambridge student, Fisher helped found the university's eugenics society, and in 1933 he succeeded Pearson both as editor of the *Annals* and as Galton Professor of Eugenics. A beautiful friendship between racists that could have been, ruined by a peer review.

The hatred lasted for generations. Pearson's son, Egon, developed hypothesis testing, with its α and β, together with Jerzy Neyman. He ended up heading a department in the same building as Fisher. They hardly ever spoke. And Fisher fought Neyman and Egon Pearson's ideas for the rest of his life. Here's the irony: the p-value you were taught is the bastard child of Fisher's significance test and Neyman and Pearson's hypothesis testing. Their authors couldn't stand each other, and textbook writers merged the two approaches into something none of them would have accepted.

Become a data artisan

After all this philosophy, here's a piece of practical advice, and it doesn't depend on which school you belong to. As we said at the very beginning: describe your data, plot it, and never skip that step.

Plots show what numbers hide. In 1973 the statistician Francis Anscombe published four small datasets with the same mean, the same variance, the same correlation and the same regression line. Plotted, they look completely different: one is a reasonable cloud of points around a line, one is a curve, one is a perfect line with a single outlier, and in the last one all the points but one sit on the same vertical line. Any calculation that only looked at those summaries would treat the four datasets as identical. (A modern and funnier version, the *Datasaurus*, hides a dinosaur among datasets with identical statistics.) Plots are also how we check some of the assumptions that every analysis, Bayesian or frequentist, depends on. Is the relationship really a line? Are there outliers doing all the work? Are the points really independent?

Statisticians have an old joke, attributed to Joseph Berkson: the best test is the *interocular trauma test*, the one where the result hits you right between the eyes. If an effect is obvious in a plot, the statistics will mostly confirm what you already see. If you need a sophisticated calculation to find it, and the data don't speak clearly in the plots, be extra careful, whatever your p-value or your posterior says.

One warning, though: our eyes can fool us too. We see faces in clouds and patterns in noise, and looking at dozens of plots until something shows up is just another garden of forking paths. There's a nice way to check yourself, called a *lineup*: hide the plot of your real data among several plots made from data where nothing is going on, and see whether someone can pick out the real one. If they can't, the pattern you saw may be noise. It's a significance test done by eye. So looking and calculating are not rivals. Plots tell you what to suspect and what to check; calculations tell you how much to trust what you saw.

Statistics is one of the most powerful tools we have for building maps of the world. But it's a tool for reasoning well, not a machine that produces truth. The map is not the territory, and no method, however elegant, can make it so. What good statistics can do is help us draw better maps, know how much to trust them, and change them when the territory tells us we were wrong.

And, as we saw with Fisher and Pearson, no tool will do that on its own. Good tools need a good artisan. Becoming one takes practice and understanding, and it isn't easy. And even good artisans have blind spots. That's why I think the best thing we can do, as scientists and as people, is to keep talking with those who don't think like us, willing to listen and to learn from them.