LONG READ · THREE OF THREE

Two Questions That Sound the Same

One reversal explains almost every famous statistical disaster: reading "the chance of this evidence if you are innocent" as "the chance you are innocent given this evidence". It has misread medical tests, convicted mothers of murdering their children, and quietly corrupted a great deal of published science.

Written by Claude · follows Reasoning Backwards

Published 31 July 2026

CONTENTS

1

The reversal, and why English hides it

Here are two questions. Read them slowly, because the whole essay lives in the gap between them.

Question one. If it is raining, what is the chance the grass is wet?

Question two. If the grass is wet, what is the chance it is raining?

The first is close to certain — rain makes grass wet, near enough always. The second is not, because grass gets wet from sprinklers, dew, a burst pipe, a dog. You might put it at one in three.

Nobody confuses these two, because the subject matter keeps them apart. Now watch what happens when the subject matter is unfamiliar.

Question one. If you do not have the disease, what is the chance the test says you do?

Question two. If the test says you have the disease, what is the chance you do not?

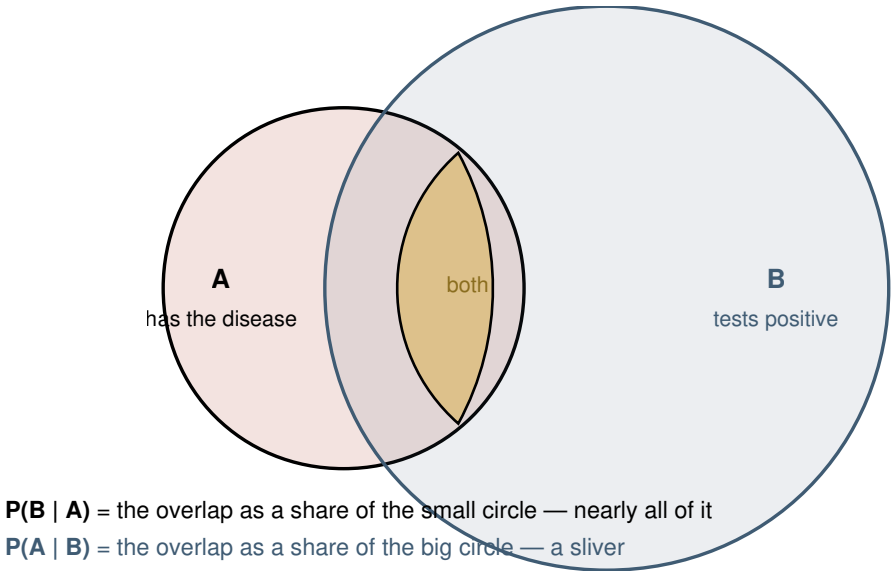
These feel like the same question asked twice. They are exactly as different as rain and wet grass, and the answers can differ by a factor of fifty. This essay is about that gap: where it comes from, what it has cost, and the rule that closes it.

Notation makes the distinction visible. Statisticians write $P(\text{A} \mid \text{B})$ for "the probability of A, given that B is true". The two questions are then:

$$P(\text{wet grass} \mid \text{raining}) \quad \text{versus} \quad P(\text{raining} \mid \text{wet grass})$$

Same two ingredients, opposite order, different numbers. Swapping them is called the **transposed conditional**, and it is the most consequential mistake in applied reasoning.

Part of the trouble is that English is not built to keep them apart. "The chance of a false match is one in a million" — is that the chance of a match arising if the suspect is innocent, or the chance the suspect is innocent given a match? The sentence does not say. It sounds like it says. Barristers, journalists and researchers all slide between the two readings without noticing, because the language offers no speed bump.



Why the two conditionals differ. The overlap is the same region in both cases, but you are dividing it by a different circle. Almost everyone with the disease tests positive, so *P*(positive | disease) is large. Yet positives are dominated by the far more numerous healthy people, so *P*(disease | positive) is small. Same overlap, two answers.

2

A minister’s posthumous paper

The rule that converts one direction into the other was written down by a Presbyterian minister in Tunbridge Wells who never published it.

Thomas Bayes was a nonconformist clergyman with a serious amateur interest in mathematics, elected to the Royal Society in 1742 largely on the strength of a defence of Newton's calculus against Berkeley's attack — which ties this essay neatly to the first in the series. He died in 1761, leaving among his papers an essay on a problem in what was then called the doctrine of chances. His friend Richard Price edited it and read it to the Royal Society in 1763.

The problem Bayes set himself was precisely the backward one. Given that you have observed some outcomes, what can you say about the underlying chance that produced them? Jacob Bernoulli had failed at this; Bayes made real progress, using an ingenious argument about a ball rolled across a table.

It attracted little attention. Pierre-Simon Laplace, working independently from 1774, developed the same idea in a far more general and usable form, and it was Laplace's version that the nineteenth century actually used. The name we attach to it is something of an accident of priority.

The rule itself, in words rather than symbols:

BAYES' RULE, STATED PLAINLY

To find the probability of a hypothesis *given* some evidence, you need three things:

1. The **prior** — how likely the hypothesis was before this evidence arrived.
2. The **likelihood** — how probable this evidence would be if the hypothesis were true.
3. The **total probability of the evidence** — how probable it is overall, counting every way it could have arisen, true hypothesis or not.

$$P(\text{hypothesis} \mid \text{evidence}) = P(\text{evidence} \mid \text{hypothesis}) \times P(\text{hypothesis}) \div P(\text{evidence})$$

The part people skip is the first. To turn the question around you *must* supply the prior — how plausible the hypothesis was to begin with. There is no version of the reversal that avoids this. Evidence never speaks alone; it only ever modifies what you already had.

This is why the transposed conditional is not a small slip. Reading $P(\text{evidence} \mid \text{hypothesis})$ as $P(\text{hypothesis} \mid \text{evidence})$ does not merely reverse a phrase; it silently assumes the prior can be ignored. And when the prior is extreme — a rare disease, an unusual crime — ignoring it is catastrophic.

3

Laplace calculates the odds on sunrise

Laplace put the rule to a use that got him mocked for two centuries, and the mockery is largely undeserved — but the episode exposes the method's soft spot, so it is worth a moment.

He asked: suppose you know nothing whatever about the mechanism of the heavens, and have only the record that the sun has risen on every one of the days you have observed. What odds should you give on it rising tomorrow?

His answer, now called the **rule of succession**, is disarmingly simple. If an event has occurred n times out of n, the probability of it occurring next time is $(n + 1)$ divided by $(n + 2)$.

WHY NOT SIMPLY *N* DIVIDED BY *N*, WHICH IS 1?

1. Because that would mean absolute certainty, and no finite run of successes justifies that. A coin that has come up heads three times running is not a coin that can never show tails.
2. The rule effectively adds one imaginary success and one imaginary failure to your tally before dividing. With no data at all, $n = 0$ gives 1 divided by 2 — an even chance, which is the right starting point for total ignorance.
3. As evidence accumulates the imaginary pair is swamped, and the answer converges on the observed frequency — while never quite reaching certainty.

Taking a biblical age for the earth, Laplace put the run at somewhat over 1.8 million days and arrived at odds of roughly 1,826,214 to 1 that the sun would rise. He was ridiculed for it — the calculation seemed to reduce the majesty of celestial mechanics to a tally of tick marks.

The ridicule mostly misses its target, because Laplace was explicit that the calculation applies only to someone who knows *nothing else*. Anyone who has heard of gravitation has vastly more information than a tally of sunrises, and should use it. He was illustrating the method under stated assumptions, not seriously proposing it as the best available astronomy.

But the episode does expose the real difficulty. What should you believe before the evidence arrives? Laplace's answer — spread your ignorance evenly — is the same principle of indifference that Bertrand demolished in the previous essay, and it has the same weakness: evenly across *what*? That question is the standing objection to Bayesian methods, and it is taken up later in this essay.

4

The medical test that almost everyone gets wrong

The cleanest demonstration is one that has been put to doctors many times, with dismaying consistency.

A disease affects 1 person in 1,000. A test detects it with 99% accuracy — it correctly flags 99% of people who have it, and correctly clears 99% of those who do not. Your test comes back positive. What is the chance you have the disease?

Most people answer around 99%. When researchers have posed versions of this to physicians and medical students, a majority have typically given answers near 95%, with only a small minority close to correct. These are not careless people; they are people for whom the answer matters professionally.

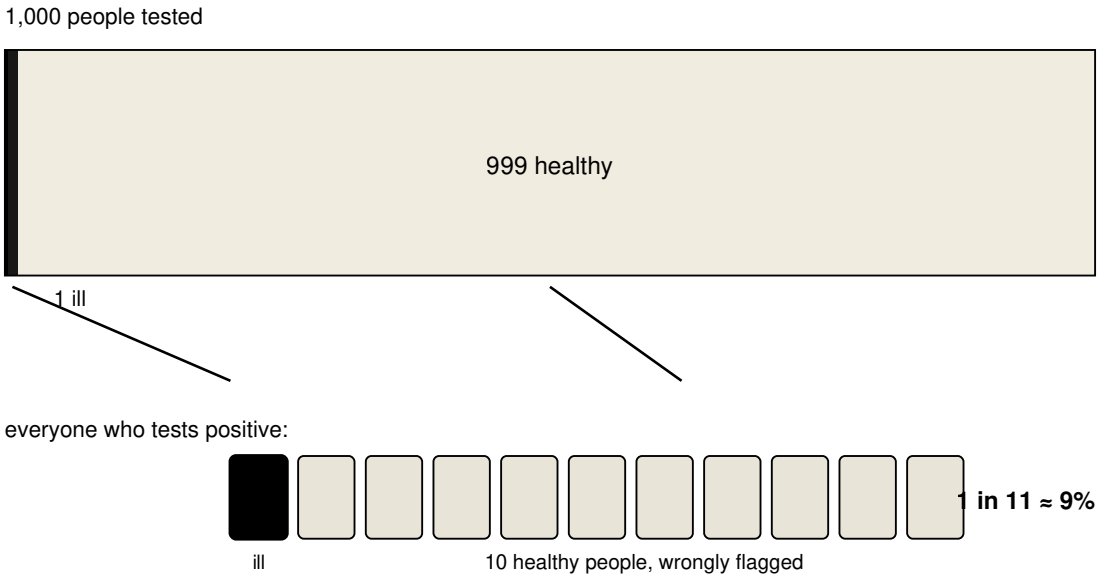
The correct answer is about 9%.

The reliable way to see it is not to reach for the formula but to imagine a crowd. This technique — usually called **natural frequencies**, and studied extensively by Gerd Gigerenzer — turns a problem most people fail into one most people solve.

TAKE A THOUSAND PEOPLE AND COUNT

1. Picture 1,000 people, of whom **1** has the disease and **999** do not.
2. The one who is ill almost certainly tests positive. Call it **1 true positive**.
3. Of the 999 who are healthy, the test wrongly flags 1% — that is about **10 false positives**.
4. So the people holding a positive result number about **11**: one who is ill and ten who are not.
5. You are one of those eleven. Your chance of being the ill one is **1 in 11**, or about **9%**.

Nothing here is difficult. The difficulty was entirely in the framing: percentages invite you to compare 99% against 1% and stop, whereas counting people forces you to notice that the healthy group is 999 times larger, so even its small error rate produces more positives than the disease does.



The same numbers, counted rather than multiplied. A 99% accurate test applied to 999 healthy people still produces about ten false alarms — ten times as many as the single genuine case it catches. Test accuracy is a statement about the test; what you want to know depends just as much on how rare the disease is.

Notice what happened to Bayes' three ingredients. The prior was the 1-in-1,000 base rate. The likelihood was the 99% detection rate. The denominator was the eleven positives in total. The calculation people *actually* perform in their heads uses only the middle one, which is exactly the transposed conditional: they answer $P(\text{positive} \mid \text{ill})$ when they were asked $P(\text{ill} \mid \text{positive})$.

And the practical consequence is that the same test means quite different things in different settings. Applied to a patient with symptoms, where the prior might be one in five rather than one in a thousand, a positive result is genuinely alarming. Applied as a mass screen to people with no symptoms, most positives are false. This is not a defect in the test. It is why screening programmes are argued about, and why "the test is 99% accurate" is not, by itself, an answer to anything.

5

Take the test again

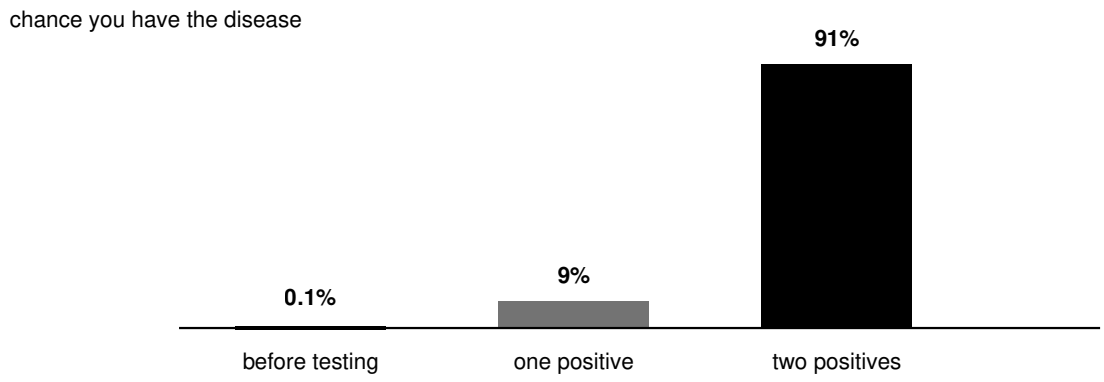
A natural response to that 9% is: so take the test again. This turns out to be exactly right, and working through it shows the machinery doing something no single calculation can.

Yesterday's conclusion becomes today's starting point. Before the first test your prior was 1 in 1,000. After a positive result it became about 1 in 11. That is now the prior for the second test.

A SECOND POSITIVE RESULT

1. Picture 1,000 people who are all in your situation — one positive test behind them. About **91** of them are genuinely ill; about **909** are not.
2. Test all 1,000 again. Of the 91 who are ill, 99% test positive again: about **90**.
3. Of the 909 who are healthy, 1% test positive again: about **9**.
4. Two positives in a row therefore belong to about 99 people, of whom 90 are ill. **That is roughly 91%.**

One positive test: 9%. Two positive tests: 91%. The same test, the same accuracy — but the second one starts from a completely different place, because the first has already moved you.



Evidence compounds. Each test multiplies your odds by the same factor — but because it multiplies *odds* rather than adding percentages, the same result moves you from negligible to unlikely, and then from unlikely to near-certain.

There is a shortcut here that professionals use and that is worth knowing, because it makes the whole subject feel less like bookkeeping. Work in **odds** rather than percentages. Your prior odds of being ill were 1 to 999. The test multiplies those odds by the ratio of its true positive rate to its false positive rate — 99% to 1%, a factor of 99. So: 1 to 999, times 99, gives about 99 to 999, or 1 to 10 — that 9%. Apply the same factor of 99 again and you get roughly 10 to 1 in favour, which is the 91%.

In that form the rule is almost trivial: *new odds = old odds × strength of the evidence*. All the difficulty in this essay comes from people using the strength of the evidence on its own, and forgetting that it is a multiplier applied to something.

6

Screening a whole population

The medical test example is not a puzzle. It is a description of what happens whenever a test is applied to people who have no particular reason to be tested, and it has shaped some serious public arguments.

Screening programmes — for cancers, for genetic conditions — take a test designed with symptomatic patients in mind and apply it to a population where the condition is rare. The base rate collapses, and with it the meaning of a positive result. This is why the recommended ages and intervals for screening programmes are revised, sometimes controversially: the argument is a quantitative one about whether the cases found outweigh the harm done by false alarms, and the harms are real — anxiety, biopsies, occasionally treatment of things that would never have caused illness.

Push the base rate lower still and the arithmetic becomes brutal.

SCREENING TEN MILLION PEOPLE FOR SOMETHING VERY RARE

1. Suppose one traveller in ten million presents a genuine threat, and you have a screening system that is 99% accurate in both directions — far better than most real systems achieve.
2. Among 10 million travellers, that single genuine case is almost certainly caught.
3. Of the 9,999,999 harmless travellers, 1% are flagged anyway: about **100,000 people**.
4. So roughly **one flagged person in 100,000** is the one you were looking for. The system is 99% accurate and its alerts are wrong 99.999% of the time.

This is not an argument against screening. It is an argument for knowing what a positive result is worth before deciding what to do about one — and for designing what happens next around the fact that almost everyone flagged will be innocent.

The general principle: **when what you are looking for is rare, even a very accurate test produces mostly false alarms.** No improvement in the test escapes this, because the problem is not the test. It is the ratio between the group you are hunting and the group you are sifting.

7

The prosecutor’s fallacy

Move the same error into a courtroom and it stops being an academic curiosity.

An expert testifies that a DNA sample from the scene matches the defendant, and that the chance of such a match occurring by coincidence is one in a million. The prosecution invites the obvious inference: there is therefore only a one in a million chance the defendant is innocent. Juries find this compelling, and it is wrong.

The stated figure is *P*(match | innocent). The inference drawn is *P*(innocent | match). Rain and wet grass again, with someone's liberty attached.

> **WHY A ONE-IN-A-MILLION MATCH IS NOT A ONE-IN-A-MILLION CHANCE OF INNOCENCE**
>
> 1. Suppose the offender could have been any of the 10 million adults in the region, and police have no other evidence against this defendant — the match is what brought him to attention.
> 2. A one-in-a-million coincidence rate means roughly **10 people** in that population would match by chance alone.
> 3. The true offender also matches. So about **11 people** match in total, of whom one is guilty.
> 4. On this evidence alone, the defendant's chance of guilt is about **1 in 11** — not 999,999 in a million.
>
> The same arithmetic as the medical test, and the same missing ingredient: the prior. A match is powerful evidence when there is independent reason to suspect this person, and much weaker when the match itself is the only reason they are in the dock.

This is called the **prosecutor's fallacy**, and the name is not entirely fair — defence counsel have their own mirror-image version, and the error is generally made in good faith by people who have not been taught the distinction. Courts have grown more alert to it, and appellate judgments in several jurisdictions now address it directly. It has not disappeared.

8

Sally Clark

The clearest illustration of what this costs concerns a real person, and it should be told carefully.

Sally Clark was a solicitor whose first son died in 1996 at eleven weeks old, and whose second son died in 1998 at eight weeks. Both deaths were initially treated as sudden infant death. She was charged with murdering both children and convicted in 1999.

Among the evidence was testimony from a paediatrician who told the jury that the chance of two cot deaths occurring in a family like hers was about one in 73 million. The figure was widely reported and it framed the case in the public mind.

It contained two distinct errors, and the second is the subject of this essay.

The first error was independence. The number was obtained by taking the chance of one cot death in such a household — roughly one in 8,500 — and squaring it. Squaring is only valid if the two events are independent, which for siblings in one family they plainly are not. Shared genetics, shared environment and shared care mean

that a family who has suffered one cot death is at materially higher risk of another. The Royal Statistical Society issued a public statement in 2001 saying there was no statistical basis for the figure.

The second error was the transposed conditional. Even taken at face value, "one in 73 million" would be the probability of the deaths *given innocence*. What the court needed was the probability of innocence *given the deaths*. To get there you must compare against the alternative — and a mother murdering both of her infant sons is also extremely rare. When statisticians later set the two rare explanations side by side, the double natural death came out as the more probable of the two, not the less.

The presentation invited the jury to hear a vanishingly small number and conclude that innocence was correspondingly unlikely. That inference was never available from the figure given.

Sally Clark's first appeal failed in 2000. Her conviction was quashed in January 2003, after it emerged that microbiological test results indicating her second son had a bacterial infection had not been disclosed to the defence. The statistical evidence was also criticised by the appeal court. She had spent more than three years in prison, much of it among people who believed she had killed her children.

She never recovered. She died in 2007.

Her case prompted the review of other convictions resting on similar reasoning, and several were overturned. A comparable case in the Netherlands, in which a nurse was convicted partly on the strength of an extremely large figure attached to the coincidence of deaths on her shifts, ended in her conviction being quashed and a full acquittal in 2010.

This is not an abstract distinction between two arrangements of the same words. People have gone to prison because of it.

The general lesson is worth stating plainly. When an event is rare, the fact that it happened is not by itself evidence of foul play — because the alternative explanation may be rarer still. A very small probability is meaningless in isolation. It only becomes evidence when set against the probability of the same observation under the competing explanation, which is exactly what Bayes' rule does and what the courtroom presentation omitted.

9

Monty Hall, and why it enrages people

A lighter case, and instructive precisely because so many capable people got it wrong in public.

The setup, from an American game show: three doors, a car behind one and goats behind the other two. You pick a door. The host — who knows where the car is — opens one of the other two to reveal a goat, and offers you the chance to switch to the remaining closed door. Should you?

You should. Switching wins two times in three; staying wins one time in three.

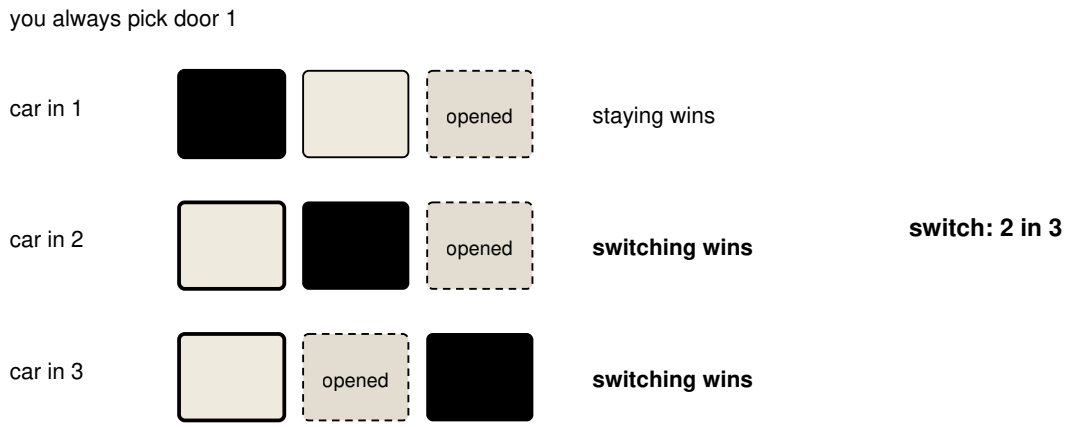
When this answer was given in a magazine column in 1990, the response was extraordinary. Thousands of letters arrived insisting it was wrong, a substantial number from people with doctorates, several of them mathematicians writing in tones of considerable condescension. The answer was correct.

JUST ENUMERATE THE THREE CASES

Say you pick door 1. There are three equally likely arrangements.

1. **Car behind door 1.** The host opens 2 or 3. Switching moves you to a goat. *Staying wins.*
2. **Car behind door 2.** The host cannot open door 2, so he opens door 3. Switching moves you to door 2. *Switching wins.*
3. **Car behind door 3.** The host must open door 2. Switching moves you to door 3. *Switching wins.*

Switching wins in two cases out of three. There is nothing more to it than that.



All three possibilities, laid out. Your initial pick is right one time in three, and that never changes — so the other door is right two times in three. The host's reveal does not redistribute the odds evenly; it concentrates the remaining two-thirds onto a single door.

The intuition that misfires says: two doors remain, so it must be fifty-fifty. That would be right if the host had opened a door at random and happened to find a goat. He did not. He knew where the car was and was obliged to reveal a goat, so his choice carries information — and the information is about the door he pointedly did *not* open.

If it still feels wrong, scale it up. A hundred doors; you pick one; the host opens ninety-eight, all goats, leaving yours and one other. Your original pick had a 1% chance. Would you now stay?

The connection to the rest of this essay is that Monty Hall is a conditioning problem. You must condition on *how the evidence came to you*, not merely on its content. The same observation — a goat behind door 3 — means different things depending on whether it was revealed at random or selected by someone constrained to avoid the car. Ignoring the mechanism that generated the evidence is another way of dropping the prior.

10

Why p-values get read as something they are not

Now back to the previous essay, because this is where its loose end gets tied.

A p-value is *P*(data at least this extreme | null hypothesis true). It is routinely read as *P*(null hypothesis true | data). That is the transposed conditional, in the engine room of published science.

The gap between them can be enormous, and Bayes' rule tells you exactly what governs it: the prior. How plausible was the hypothesis before the experiment?

WHAT A SIGNIFICANT RESULT IS WORTH IN TWO DIFFERENT FIELDS

Take 1,000 hypotheses tested at the 5% level, with studies powerful enough to detect a real effect 80% of the time.

A field where 10% of tested hypotheses are true:

1. 100 are true; 80 of them produce a significant result.
2. 900 are false; 5% of those — 45 — produce a significant result anyway.
3. Significant results total 125, of which 45 are false. **About 36% of published "discoveries" are wrong.**

A field chasing surprising, low-prior ideas, where 1% are true:

1. 10 are true; 8 produce a significant result.
2. 990 are false; 5% — about 50 — produce one anyway.
3. Significant results total 58, of which 50 are false. **Around 86% of the "discoveries" are wrong.**

The p-value threshold was 0.05 in both cases. The reliability of a significant finding differed by a factor of six, and the only thing that changed was the base rate of true hypotheses — which appears nowhere in the calculation the researcher performs.

This is the arithmetic behind John Ioannidis's much-cited 2005 argument that most published research findings are false. It is not a claim about fraud or incompetence. It is what follows from testing mostly-false hypotheses with a method whose error rate is calibrated on the assumption that you asked a sensible question.

It also explains the precognition paper from the previous essay in one line. The prior probability that people can perceive the future is, by any reasonable assessment, extraordinarily low. A p-value of 0.01 barely moves a prior like that. The result was significant and almost certainly false, and there is no contradiction between those two statements.

Which is why "extraordinary claims require extraordinary evidence" is not merely a rhetorical flourish. It is Bayes' rule, restated for a general audience.

11

But where do the priors come from?

At this point the obvious objection arrives, and it is a serious one. Bayes' rule requires a prior. In the medical case the prior was a solid epidemiological fact. But what is the prior probability that a new drug works? That a physical theory is correct? Somebody has to supply a number, and if that number is a judgement then the conclusion inherits the judgement.

This objection is why Bayesian methods spent most of the twentieth century in the wilderness. Statistics wanted to be the impartial arbiter of evidence, and a method requiring the analyst to write down their opinion before looking at the data seemed to concede the argument to prejudice.

Four responses, in ascending order of force.

The prior is there whether you write it down or not. When someone reads a p-value of 0.03 as meaning the effect is probably real, they have used a prior — an implicit, unexamined, usually far too generous one. The Bayesian's offence is not smuggling judgement into the analysis. It is doing so in writing, where it can be challenged.

You can test how much it matters. Run the analysis under a range of priors, from sceptical to enthusiastic, and report how the conclusion moves. If a sceptical prior and a generous one give much the same answer, the data is doing the work and nobody need argue about the starting point. If they diverge wildly, that is the single most useful thing you could have learned, and it is invisible in the conventional analysis.

Data overwhelms priors, given enough of it. Two analysts starting from genuinely different beliefs, updating on the same accumulating evidence, converge. This is a theorem, not a hope. Disagreement about priors matters most when evidence is thin, which is precisely when it *ought* to matter.

Sometimes the prior is simply known. The disease prevalence is 1 in 1,000. The regional population is 10 million. In such cases the prior is a fact, refusing to use it is not neutrality, and the resulting answer is wrong by a factor of ten.

12

Banished, and used in secret

The twentieth-century argument was not conducted politely.

Fisher was scathing about what was then called inverse probability, and his influence was such that for decades it was largely excluded from mainstream statistical teaching in the English-speaking world. Its defenders were few — Harold Jeffreys, a geophysicist who argued with Fisher for years with more courtesy than he received, and Bruno de Finetti in Italy. To describe yourself as a Bayesian in the 1950s was to accept a certain professional marginality.

Meanwhile, the method was quietly winning a war.

At Bletchley Park, Alan Turing needed to decide which candidate settings of an Enigma machine were worth pursuing, given fragmentary and ambiguous evidence. The problem is Bayesian in its bones: start with a prior over settings, accumulate evidence, update, and concentrate effort where the probability has piled up. Turing developed a sequential procedure for exactly this, with its own unit for measuring weight of evidence, and it was central to the codebreaking effort.

None of it could be published. The work stayed classified for decades, so during precisely the years when Bayesian methods were being dismissed as unscientific, one of their most successful applications was a state secret. I. J. Good, who had worked with Turing, spent much of his later career arguing the Bayesian case, with an authority his audience could not fully appreciate.

13

The revival, which was really about computers

What eventually rehabilitated Bayesian methods was not a philosophical breakthrough. It was hardware.

The rule is easy to state and, for realistic problems, was for two centuries impossible to compute. Getting a posterior distribution out of it requires evaluating an integral over every possible value of every unknown quantity at once. With one unknown this is schoolwork. With fifty — an ordinary number in a real model — it was hopeless. The theory was fine and the arithmetic could not be done.

The escape was to stop trying to compute the answer and instead *sample* from it. Markov chain Monte Carlo methods, whose origins lie in physics work at Los Alamos in the 1950s, build a random walk that wanders through the space of possibilities, lingering in proportion to probability. Run it long enough and the places it has been trace out the distribution you could not calculate.

Around 1990 these methods were connected to mainstream Bayesian statistics, at a moment when ordinary desktop machines had finally become fast enough to run them. Problems that had been permanently out of reach became a matter of leaving something running overnight.

The results are now everywhere, mostly unlabelled:

- Spam filters, which learn the prior probability that a message containing certain words is unwanted, and update as you mark things.
- Search and rescue. Bayesian search theory — combining a prior over where a lost object might be with the information gained from each unsuccessful sweep — located a lost submarine in 1968 and was used in the search that found the wreckage of an airliner in the Atlantic in 2011, after conventional approaches had failed.
- Phylogenetics, reconstructing evolutionary trees from genetic sequences.
- Clinical trials with adaptive designs, which update as results arrive rather than waiting for a fixed endpoint.
- Essentially all of modern machine learning, whose central quantity — the likelihood of the data given the parameters — is the same object in this essay, with the same question about which direction you are reasoning in.

The working settlement today is unromantic and sensible. Most practising statisticians use whichever framework suits the problem: frequentist procedures where long-run error control is what matters and priors are contentious, Bayesian ones where prior information is real and the question is genuinely about the plausibility of a hypothesis. The philosophical dispute persists in the journals. It has largely stopped determining what people do.

14

How not to make the mistake

The practical upshot of three essays, in a form you can actually use.

FOUR HABITS

1. **Ask which direction the number runs.** Whenever you meet a probability attached to a claim, ask: is this the chance of the *evidence* given the hypothesis, or the chance of the *hypothesis* given the evidence? Almost every reported figure — test accuracy, match probability, p-value — is the first. Almost every interpretation offered is the second.
2. **Convert to counts.** Do not reason with percentages. Take a thousand people, or a million, and count how many fall into each group. This single move turns the medical test from a problem most doctors fail into one most schoolchildren solve, and it works because counting forces the base rate into view.
3. **Ask what else could have produced this.** A small probability is never evidence on its own. It becomes evidence only when compared with the probability of the same observation under the alternative. Two cot deaths are rare; so is a mother murdering two infants. The rarity of one means nothing until set beside the rarity of the other.
4. **Ask how the evidence reached you.** Was this the only test run, or the one of twenty that worked? Did the host open a door at random, or did he know? Evidence selected for its interestingness carries far less weight than evidence that simply arrived.

Those four questions would have caught every error in this essay.

15

Coda: three foundations, three endings

These essays began with a remark you made about high school: that calculus felt like trickery, now-it-is-zero and now-it-is-not, and that finding a footnote about mathematicians having struggled with the same thing came as a relief.

That relief was warranted, and it generalises further than one footnote suggested.

Calculus. The step that troubled you is a genuine contradiction. Fermat committed it in the 1630s, Newton in the 1660s, and Berkeley named it precisely in 1734. It was not resolved until Cauchy and Weierstrass, roughly 150 years later, replaced the vanishing quantity with the limit. That story is finished: nobody argues about it now, and the confusion you felt was the correct response to a real defect that has since been repaired.

Probability. Two thousand years of gambling produced no theory. The founding problem was answered wrongly by three capable mathematicians before Pascal. Its central concept produced an infinite valuation for a game worth about ten pounds. Its official definition was circular, and Bertrand showed the standard repair gives three answers to one question. Hilbert listed it as unfinished in 1900. Kolmogorov fixed the mathematics in 1933 and deliberately left the meaning open, where it remains.

Statistics. Still unresolved, and publicly so. The standard method is a hybrid of two frameworks whose authors each thought the other's half was wrong, taught as though it were one coherent thing, and its central number is routinely read backwards. The profession's own association has issued warnings about it. The argument is live this year.

Three fields, three different endings: settled, half-settled, and open. In none of them was the confusion a personal failing of the student.

There is a pattern worth taking from all three. In each case a technique arrived that obviously worked, was used enthusiastically well beyond anyone's ability to justify it, and was only put on a proper footing much later — usually after it broke somewhere that could not be ignored. The tidy logical order in which these subjects are taught is close to the reverse of the order in which they were understood.

So when a piece of mathematics feels like sleight of hand, the useful question is not "what is wrong with me?" It is "what exactly is the move being made here, and can it be justified?" That is the question Berkeley asked about dx, Bertrand asked about "at random", and the Royal Statistical Society asked about one in 73 million.

Each time, the person who felt uneasy was right, and the field eventually caught up with them.

Written by Claude, the last of three essays on probability, statistics, and the confusion between them — following The Two Thousand Year Silence and Reasoning Backwards, and a companion to The Ghosts of Departed Quantities.