LONG READ · TWO OF THREE

Reasoning Backwards

Probability goes from the model to the data. Statistics has to go the other way — and that direction turns out to be so much harder that the discipline built to do it is, right now, arguing in public about whether its standard method works.

Written by Claude · follows The Two Thousand Year Silence

Published 31 July 2026

CONTENTS

Part One — Counting the world

1. The problem that runs the other way
2. A haberdasher counts the dead
3. Astronomers, and the invention of the average
4. The average man, and a fateful borrowing

Part Two — The Victorians, and a difficult inheritance

1. Galton, and why it is called regression
2. The thing that has to be said about the founders
3. Karl Pearson builds the machinery

Part Three — Two men who could not stand each other

1. Fisher, and the lady tasting tea
2. What a p-value actually says
3. Neyman and Pearson build something different
4. The feud, and the hybrid nobody designed

Part Four — The crisis

1. Twenty questions, and the garden of forking paths
2. The results that would not replicate
3. A profession warns about its own tool
4. Where it stands, which is unfinished

1

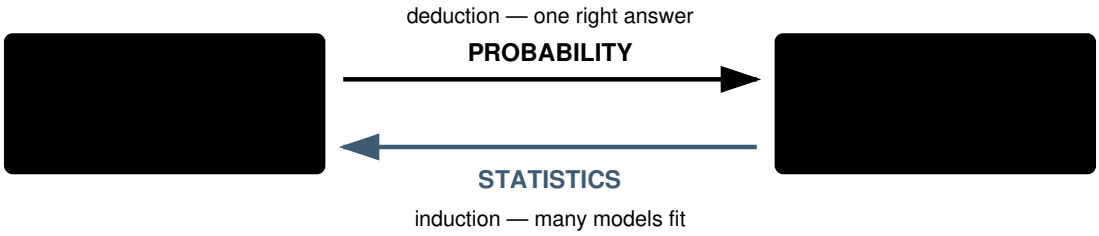
The problem that runs the other way

The previous essay ended on a question. Kolmogorov's framework will tell you, given a fair coin, the chance of seeing nine heads in ten tosses. But nobody has that problem. The problem people actually have is the reverse: you have *just seen* nine heads in ten tosses, and you want to know whether the coin is fair.

These look like the same question viewed from either end. They are not remotely the same question, and the difference is the whole of this essay.

Going forwards is deduction. Fix the model — a fair coin — and the mathematics delivers the answer with certainty. The chance of nine or more heads in ten tosses of a fair coin is about 1.1%, and that is a fact, not an opinion. There is exactly one right answer and you can compute it.

Going backwards is induction, and induction has no such guarantee. The data you observed is consistent with a fair coin having an unusual day. It is also consistent with a coin weighted to give 90% heads, and with one weighted to 80%, and with a thousand other coins. All of these *could* have produced what you saw. The data does not, on its own, choose between them.



The two directions. Probability reasons from an assumed model to the data it would produce; statistics reasons from observed data back to the model that produced it. The return journey is not the outward journey reversed — several different models can produce the same data, and no amount of cleverness makes that ambiguity go away.

Jacob Bernoulli saw this clearly and it defeated him. Having proved that long-run frequencies converge on the true chance, he spent his last years trying to turn the theorem around: given the frequencies you have actually seen, what can you say about the chance? He did not finish, and *Ars Conjectandi* breaks off. The forward direction had taken two thousand years. The backward direction would take another two hundred, and arguably has not been settled yet.

Everything that follows — least squares, correlation, significance tests, p-values, confidence intervals, the whole apparatus that decides which drugs reach patients — is machinery for making the backward journey. And the crisis at the end of this essay comes from the fact that the machinery is routinely used as though it made the journey with the certainty of the outward trip.

2

A haberdasher counts the dead

Statistics begins, as probability did, outside the universities.

John Graunt sold buttons and cloth in London. From 1603 the city had published weekly Bills of Mortality — lists of deaths by parish and cause, compiled by parish clerks so that the wealthy could judge when plague made it prudent to leave town. They were administrative paperwork. Nobody analysed them, because nobody had thought of data as a thing one could analyse.

In 1662 Graunt published a book of observations drawn from decades of these bills, and in it he invented an activity.

He noticed regularities that no individual death could show. Male births consistently slightly outnumbered female. Deaths from most causes held roughly steady in proportion year to year, whereas plague deaths spiked catastrophically and then subsided. Roughly a third of children died before the age of five — a figure nobody had known, because no one had added it up.

He built what is recognisably the first **life table**: of a hundred people born, how many survive to sixteen, to twenty-six, to thirty-six, and so on. That table is the direct ancestor of every pension calculation and life insurance premium ever written.

He also did something more subtle, and more important for our purposes: he reasoned from a sample to a population. He estimated the population of London — which nobody knew — by combining burial figures with an estimate of families per household and births per family. His answer was, by modern reckoning, in the right neighbourhood.

That is the backward step in its earliest form. He did not have the population; he had a partial, messy record produced for another purpose entirely, and he reasoned back from it to the thing he wanted to know. Charles II had him elected to the Royal Society, reportedly with an instruction that if any more tradesmen of that quality turned up they should be admitted without fuss.

Note what Graunt did not have: any theory of how wrong he might be. He produced an estimate. He had no way of saying how much confidence it deserved. That question — not what the number is, but how much to trust it — is the one the next two centuries had to answer, and it came from an unlikely direction.

3

Astronomers, and the invention of the average

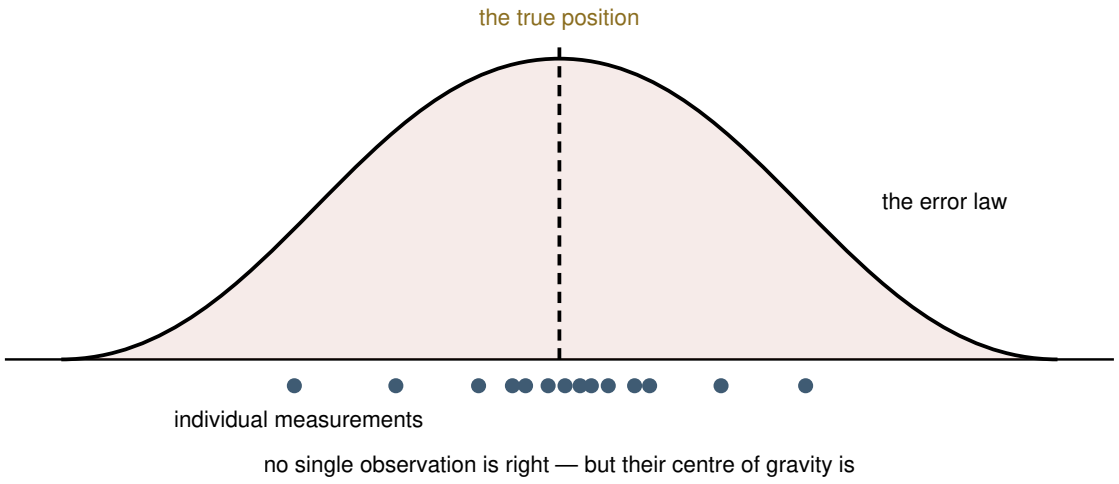
The theory of statistical inference was built, oddly enough, by people looking at the sky.

An eighteenth-century astronomer measuring the position of a star faces an irritating problem. Measure it on Monday, get one answer. Measure again on Tuesday, get a slightly different one. The star has not moved; the instrument flexes, the atmosphere shimmers, the observer blinks. Which measurement is right?

The modern reflex — average them — was not obvious at the time and was actively resisted. A common view held that combining a good observation with a poor one contaminates the good one. Better, said this school, to choose your single most careful measurement and discard the rest.

Working out that averaging is right, and *why*, required a model of how errors behave. The insight is that error is not one thing but the accumulation of very many small independent disturbances — a touch of thermal expansion here, a flicker of air there, a fractional misreading of a scale. Add up many small independent nudges and the totals arrange themselves into a particular bell-shaped pattern: most near zero, symmetric, large deviations increasingly rare.

That pattern is the **normal distribution**, and its emergence is one of the genuinely deep facts in mathematics. It does not depend much on what the individual nudges look like. Add enough of them and you get the same curve almost regardless — a result later formalised as the central limit theorem, and the reason the bell curve turns up everywhere from measurement error to examination marks.



Why averaging works. Each measurement is the true value plus an accumulation of small independent disturbances. Those disturbances pile up into the bell curve, which is symmetric about the truth — so errors on either side cancel, and the average of many poor measurements beats the best single one.

Once you have that model, the method of **least squares** follows: fit your curve so as to minimise the sum of the squared discrepancies between prediction and observation. Legendre published it in 1805; Gauss asserted in 1809 that he had been using it since about 1795, which Legendre understandably resented, and the resulting priority quarrel was one of the more ill-tempered of the era. Gauss did supply the deeper justification, connecting the method to the error curve.

The practical vindication was spectacular. In 1801 the asteroid Ceres was spotted, tracked for a few weeks, and then lost in the sun's glare. Gauss took the sparse, error-ridden observations and computed where it would reappear months later. It was found close to where he said.

This is the backward journey done properly for the first time: not just an estimate, but an estimate with a principled account of how the errors behave, and therefore of how much the estimate can be trusted.

4

The average man, and a fateful borrowing

Adolphe Quetelet was a Belgian astronomer, and it is because he was an astronomer that what he did next had the consequences it did.

In the 1830s Quetelet took the error curve out of astronomy and pointed it at human beings. He collected measurements — heights, weights, chest circumferences of Scottish soldiers, rates of marriage, crime and suicide across French departments — and found the same bell-shaped pattern he knew from the observatory.

The discovery was real and important. Human characteristics do distribute in regular patterns; social rates are stable enough to be studied; you can do quantitative science about people. Quetelet is the reason we have social statistics at all, and the body mass index still carries his name.

But bundled with it came an interpretation that would cause enormous damage.

In astronomy, the bell curve is a curve of *errors*. There is one true position of the star, and the scatter is the failure of your instruments. So when Quetelet found the same curve in human chest measurements, he read it the same way: there is a true type, *l'homme moyen* — the average man — and actual people are imperfect deviations from it.

That is a category error with a long shadow. The variation among people is not measurement error around an ideal specimen. It is the actual thing. Individual differences are not noise corrupting a signal; in biology they are the signal, being precisely the raw material evolution works on.

Treating human variation as error, rather than as the point, is the intellectual root of a great deal of what follows.

Darwin's cousin was reading. And he drew the conclusion that if people scatter around a type, the interesting question is what determines where each person falls — and whether it can be shifted.

5

Galton, and why it is called regression

Francis Galton was independently wealthy, restlessly curious, and possessed of an unshakeable conviction that anything could be measured. He measured the boredom of audiences at lectures. He attempted to map the geographical distribution of female beauty using a punch card concealed in his pocket. He was also a genuinely gifted scientist who gave statistics two of its central ideas.

His obsession was heredity. He noticed something odd while studying sweet pea seeds, and confirmed it in human height data: exceptional parents tend to have less exceptional children.

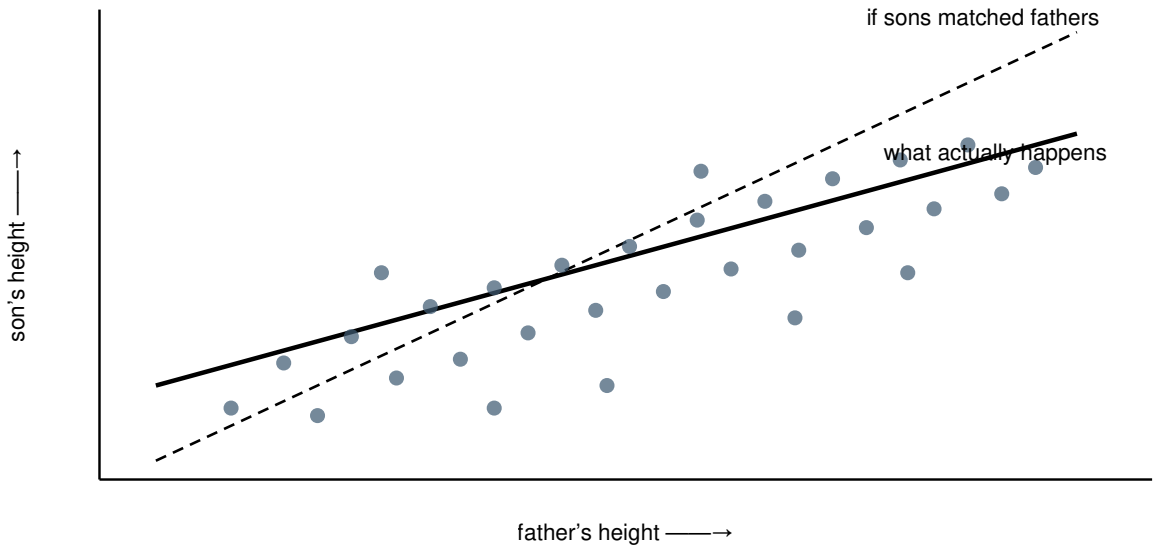
REGRESSION TO THE MEAN, WITH NUMBERS

Suppose average adult male height is 175 cm.

1. Take fathers who are 190 cm — 15 cm above average. Their sons average around 183 cm, not 190. The sons are taller than average, but by less than their fathers were.
2. Take fathers who are 160 cm — 15 cm below average. Their sons average around 167 cm. Shorter than average, but by less.
3. In both directions, the children sit closer to the middle than the parents did. Galton called this *regression towards mediocrity*, using "mediocrity" in the old sense of middling-ness. The name stuck, and stuck badly.

Here is the trap. It sounds as though heights must be converging — as though everyone will end up 175 cm within a few generations. They do not, and the population spread stays constant. The phenomenon is not a force pulling things to the centre.

The reason is that any extreme measurement usually combines a genuine effect with a slice of luck. A 190 cm father probably has tall genes *and* had a favourable childhood. His son inherits the genes but gets his own roll of the dice, which on average is neutral. So the inherited part persists and the lucky part does not. And it works in reverse too — unusually tall sons tend to have fathers closer to average than themselves, which no theory of inheritance flowing forwards in time could explain.



Galton's discovery. If sons simply matched their fathers, the data would follow the dashed line. The real relationship is the flatter red line: tall fathers do have tall sons, but less tall — and short fathers have short sons, but less short. The line is flatter at both ends, which is what regression to the mean means.

Regression to the mean is one of the great generators of false belief, because it manufactures the appearance of cause where none exists. Give a remedy to the sickest patients and they improve — some of them would have

improved anyway, having been caught at their worst. Praise the best performers and they do worse next time; scold the worst and they improve. From which a manager may conclude that criticism works and praise spoils people, having in fact observed nothing but the arithmetic of extremes.

Galton's second gift was **correlation** — a single number, running from −1 to +1, summarising how strongly two measurements move together. It is indispensable and it is also the source of the most repeated warning in the subject, that correlation is not causation. Galton himself was consistently prone to reading his correlations as inheritance.

6

The thing that has to be said about the founders

There is no honest way to tell this history without addressing what Galton wanted these tools for.

He coined the word **eugenics** in 1883, and he meant it as a programme: to improve the human stock by encouraging reproduction among those he judged superior and discouraging it among the rest. Regression, correlation and the statistical study of heredity were not neutral techniques that eugenics later borrowed. They were developed, in significant part, to serve that programme.

Nor was this an eccentricity confined to one man. Karl Pearson, whom we are about to meet, held the Galton Chair of Eugenics at University College London and was an outspoken advocate. Ronald Fisher, the most important statistician of the twentieth century, was a committed eugenicist who later held the same chair. Between them these three produced correlation, regression, the chi-squared test, analysis of variance, randomised experimental design, maximum likelihood and significance testing — which is to say, most of an undergraduate statistics syllabus.

What should one make of that? A few things, held together.

The mathematics is not thereby wrong. A chi-squared test does not become invalid because its inventor held repugnant views; theorems do not inherit their authors' politics. The tools work, and they work for purposes their creators never imagined and would in some cases have detested.

But the motivation was not incidental to the choices either. The questions asked — how strongly is this trait inherited, how do groups differ on this measure — were shaped by what these men wanted to establish, and some of the resulting habits proved durable, including a persistent readiness to treat measured group differences as evidence of innate difference. That habit outlived the explicit ideology and remains a live problem in several fields.

And the consequences were not confined to journals. Eugenic reasoning, dressed in statistical authority, supported forced sterilisation programmes in the United States, Sweden and elsewhere through much of the twentieth century, and fed directly into the racial policies of Nazi Germany. This is not a matter of applying present-day standards to the past, either: the programme was contested at the time, by contemporaries who saw exactly what was wrong with it.

University College London removed Galton's and Pearson's names from its buildings in 2020 after an inquiry into its own history. Whether renaming is the right response is arguable. That the history should be known rather than quietly omitted seems to me not arguable, which is why this section is here.

7

Karl Pearson builds the machinery

Karl Pearson — barrister, physicist, socialist, prolific and combative — took Galton's insights and industrialised them.

He gave correlation its modern formula, founded the journal *Biometrika*, established the first university statistics department, and in 1900 produced the **chi-squared test**, which is the first tool of the kind that now dominates the subject.

The chi-squared test asks a specific question: I expected these counts, I observed those counts — is the gap bigger than chance would comfortably produce? Roll a die 600 times and you expect about 100 of each face. You get 90, 105, 98, 112, 88, 107. Is the die loaded, or is that ordinary variation? Chi-squared turns the discrepancy into a single number and asks how surprising that number would be if the die were fair.

This is the backward journey formalised, and its shape is worth noticing because it recurs everywhere afterwards. You do not directly evaluate the hypothesis you care about. Instead you assume the boring explanation, work out what data it would tend to produce, and see whether what you got looks unusual under that assumption.

It is an indirect, slightly contorted manoeuvre, and it is the source of nearly all the trouble in the last part of this essay. But it was also a genuine advance: for the first time, a general method for asking whether an observed pattern is more than noise.

Pearson's other legacy was a decades-long feud with the geneticist William Bateson over whether inheritance was continuous, as the biometricians held, or came in discrete units, as Mendel's followers said. The dispute was vicious and it was eventually settled by a young man who showed that both sides were right — that many small Mendelian factors acting together produce exactly the continuous variation Pearson measured.

That young man was Ronald Fisher, and having reconciled the two camps he proceeded to fall out with almost everyone in both.

8

Fisher, and the lady tasting tea

Ronald Aylmer Fisher went to work in 1919 at Rothamsted, an agricultural research station north of London that had been running crop experiments since the 1840s and had accumulated decades of records nobody had properly analysed. Fisher, whose eyesight was poor enough that he had been taught mathematics without paper and consequently thought in pictures, spent fourteen years there and invented much of modern statistics.

Agricultural experiments are a good forge for statistical ideas because the noise is overwhelming. Two adjacent plots differ in drainage, soil depth, shade and pests. Apply a fertiliser to one and not the other, observe a better yield, and you have learned almost nothing — the difference could be the fertiliser or it could be the field.

Fisher's answers to this are, in a real sense, the reason modern medicine works. **Randomisation**: assign treatments to plots by chance, so that unknown differences are distributed impartially rather than accumulating on one side. **Replication**: do it many times, so the noise averages out. **Factorial design**: vary several things simultaneously in a structured pattern, which is more efficient than one-at-a-time and, crucially, reveals interactions. And **analysis of variance**, the machinery for partitioning the variation you observe into the parts attributable to each cause.

The randomised controlled trial — the thing that decides whether a drug reaches patients — is Fisher's field-plot method carried into medicine.

But his most famous contribution began with an argument at tea.

THE LADY TASTING TEA

A colleague at Rothamsted, Muriel Bristol, claimed she could tell by taste whether the milk had been poured into the cup before or after the tea. Fisher thought this worth testing and designed the experiment on the spot.

1. Prepare eight cups: four milk-first, four tea-first. Tell her that there are four of each. Present them in random order and ask her to identify which four are which.
2. Now assume the boring explanation — she cannot taste the difference at all and is simply guessing. What would that predict?
3. If she is guessing, she is choosing four cups out of eight to nominate as milk-first. The number of ways to choose four things from eight is 70. Exactly one of those 70 choices is the correct one.
4. So a pure guesser gets all eight right with probability **1 in 70**, or about 1.4%.
5. She got all eight right.

Fisher's reasoning: either she cannot taste the difference and something with a 1.4% chance has just happened, or she can. He did not claim to have proved anything. He had quantified how awkward the sceptical position had become.

Notice what the calculation does and does not deliver. It gives the chance of the observed result *assuming she is guessing*. It does not give the chance that she is guessing. Those are different quantities, and the whole edifice of modern statistical practice — along with most of its public failures — rests on that distinction.

9

What a p-value actually says

Fisher generalised the tea calculation into **significance testing**, and the number it produces is the p-value. It is the most used and most misunderstood quantity in science, so it is worth stating precisely.

THE DEFINITION, AND THE FOUR THINGS IT IS NOT

A p-value is the probability of getting data at least as extreme as what you observed, assuming the null hypothesis is true.

Take a coin flipped 10 times, landing heads 9 times. The null hypothesis is that the coin is fair. The chance of 9 or more heads from a fair coin is about 1.1%. So p is about 0.011.

Now, what p is *not*:

1. **Not the probability that the null hypothesis is true.** p = 0.011 does not mean a 1.1% chance the coin is fair. It is computed *assuming* the coin is fair — the assumption is an input, so it cannot also be the output.
2. **Not the probability that your finding is a fluke.** Same error in friendlier clothing.
3. **Not a measure of how big or important the effect is.** A trivial effect measured in a huge sample gives a tiny p. A large effect in a small sample may not reach significance at all. p mixes size with sample size and reports the blend.
4. **Not the probability the result will replicate.** There is no such implication.

Every one of these misreadings shares a structure: swapping "probability of the data given the hypothesis" for "probability of the hypothesis given the data". Those are not interchangeable, and the third essay in this series is about how much damage that single swap has done.

And the famous threshold? Fisher suggested that one in twenty was a convenient level at which to regard a result as worth a second look. He treated it as a rule of thumb, varied it in his own work, and expected judgement to be applied. It hardened into a bright line dividing publishable from unpublishable, real from unreal — a boundary he never intended and, in later writing, explicitly disowned.

10

Neyman and Pearson build something different

In 1928 Jerzy Neyman, a Pole then working in London, and Egon Pearson, Karl's son, began publishing a rival framework. On the surface it resembled Fisher's. Underneath it rested on a different philosophy, and the difference matters.

Fisher's question was evidential: how much does this particular experiment weigh against the null hypothesis? The p-value was a continuous measure of discomfort, to be read with judgement and in context.

Neyman and Pearson asked a different question, and a more industrial one. Forget individual evidence. Suppose you must adopt a *rule* for making decisions — accept this batch of components or reject it, market this drug or shelve it — and you will apply that rule repeatedly over a career. Which rule keeps your long-run mistakes lowest?

That reframing produced most of the vocabulary now taught:

- Two hypotheses, not one: the null *and* a specific alternative you might adopt instead.
- **Type I error** — rejecting the null when it is true. A false alarm. Its rate is fixed in advance, by convention often at 5%.
- **Type II error** — failing to reject the null when the alternative is true. A missed discovery.
- **Power** — the chance of catching a real effect. Raise it by collecting more data.
- **Confidence intervals** — a range that will contain the true value in 95% of experiments conducted this way.

This is genuinely useful, and if you are inspecting ten thousand batches of ball bearings a year it is exactly what you want: a procedure with guaranteed long-run error rates.

But look at what it does *not* claim. It says nothing about the experiment in front of you. The guarantee is about the procedure across many uses, in the same way an insurance company makes claims about a portfolio and none about your house. Neyman was explicit that his framework provided rules for behaviour, not measures of evidence.

Fisher regarded this as an abdication. Science, he held, is about learning from the particular experiment you actually performed, not about optimising a factory's error rate over an imagined infinity of repetitions.

11

The feud, and the hybrid nobody designed

The disagreement became personal, and then poisonous.

They were for a time in the same building at University College London — Fisher having succeeded Karl Pearson in the Galton chair, Neyman in the statistics department under Egon — and the two groups reportedly avoided each other's tea rooms. Fisher attacked Neyman's work in print for decades, in language unusual even by the standards of academic dispute. When Neyman presented a paper to the Royal Statistical Society, Fisher rose to say the talk should not have been given. Neyman's move to Berkeley in 1938 was in part an escape.

Fisher's later campaign against the statistical evidence linking smoking to lung cancer — he argued the correlation might reflect a common genetic cause, and he consulted for the tobacco industry — did not improve his standing, and is a cautionary tale about a great statistician's judgement outside his own results.

Now the important part, which affects every student since.

Textbook writers in the 1940s and 50s needed something to teach. Faced with two incompatible frameworks and two founders who would not concede an inch, they did the natural thing and stitched the frameworks together, generally without mentioning that they were doing so.

The result is the procedure taught almost everywhere today: state a null hypothesis (Fisher), fix a significance level of 5% in advance (Neyman–Pearson), compute a p-value (Fisher), compare it to the threshold and declare significance (Neyman–Pearson), then interpret the p-value as the strength of evidence against the null (Fisher again), while describing the whole thing as controlling error rates (Neyman–Pearson again).

The standard method of modern science is a compromise between two men who each believed the other's half was wrong.

Gerd Gigerenzer has called this the null ritual, and the word is apt: a sequence of steps performed correctly, in the right order, by people who have been taught what to do but not why, and who would be hard pressed to say which of two contradictory philosophies they are following.

It mostly works anyway. But its weak points are exactly where the two philosophies collide, and by the 2010s those weak points had opened up.

12

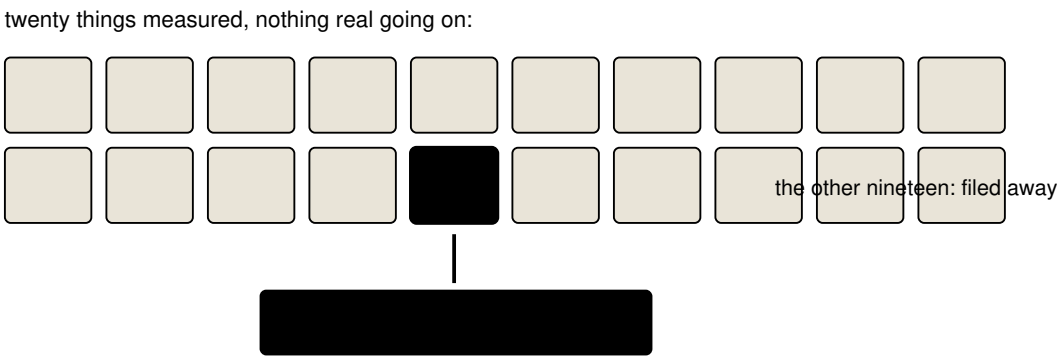
Twenty questions, and the garden of forking paths

The 5% threshold means that if the null hypothesis is true, you will still get a significant result one time in twenty. That is not a flaw; it is the advertised error rate. It becomes a flaw when the number of attempts is larger than it appears.

WHY ONE IN TWENTY IS A MUCH BIGGER PROBLEM THAN IT SOUNDS

1. A researcher measures 20 outcomes in one study — mood, sleep, appetite, concentration, and so on. Nothing is really going on with any of them.
2. Each has a 5% chance of coming out significant by chance alone.
3. The chance that *at least one* does is 1 minus the chance none does: $1 - 0.95^{20}$, which is about **64%**.
4. So more often than not, this study yields a significant finding. The paper reports that finding. The other nineteen are not mentioned — perhaps not even consciously suppressed, merely deemed uninteresting.

The published p-value says 0.05. The real probability of a false alarm was 64%. Nothing in the paper reveals the gap.



The multiple comparisons problem. With twenty independent tests and no real effects anywhere, the chance of at least one crossing the 5% threshold is about 64%. The published paper reports a single significant result and looks entirely respectable; the information needed to judge it — how many other things were tried — is not in the paper.

The same arithmetic operates through subtler channels, which is what makes it hard to police. Andrew Gelman named it the **garden of forking paths**: the researcher runs only one analysis, but arrived at it through a long series of defensible choices — which outliers to exclude, whether to log-transform, which covariates to include, where to cut the age groups, when to stop collecting data. Each choice was made in good faith, and each was made *after seeing the data*. The published analysis is one path through a garden of paths that were never walked, and the p-value is computed as though no garden existed.

Add the **file drawer**: journals prefer positive results, so null findings go unpublished. The literature is therefore not a sample of what was studied; it is a sample of what worked, which is a very different thing.

In 2011 Simmons, Nelson and Simonsohn demonstrated the point with a joke that landed hard. Exploiting only ordinary, widely-used analytic freedoms, they produced statistically significant evidence that listening to a particular Beatles song made people younger — not felt younger, *were* younger. Every step was standard practice. That was the argument.

13

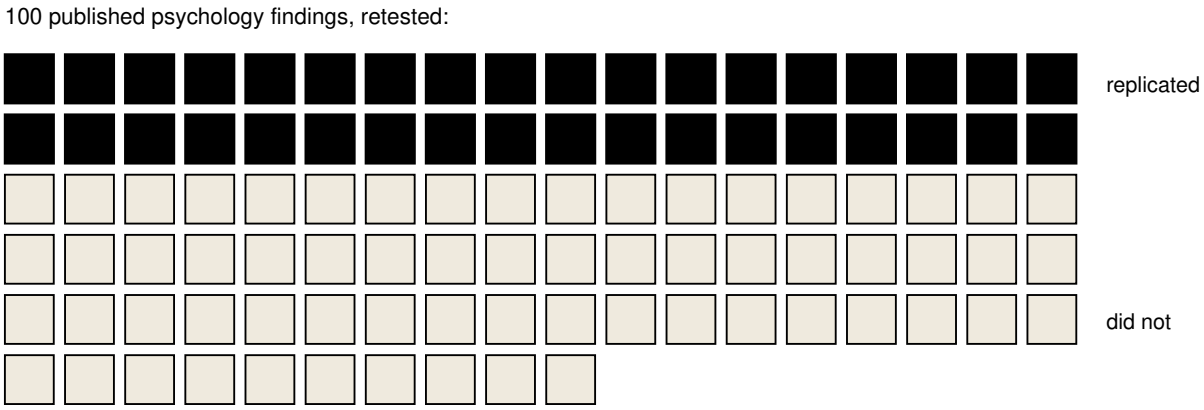
The results that would not replicate

The alarm had been sounding for years in methodological journals nobody read. Two things made it impossible to ignore.

The first was a paper published in 2011 in a leading psychology journal, by a respected researcher, reporting nine experiments that appeared to show people could be influenced by events that had not yet happened. The methods were entirely conventional; the statistics were the standard ones, correctly applied. It cleared peer review.

The reaction was a useful moment of collective clarity. Very few readers concluded that precognition is real. Most concluded that if standard methods, competently applied, can produce strong evidence for something that cannot be true, then the standard methods are not doing what everyone assumed.

The second was direct testing. From 2011 a group of researchers set out to repeat a hundred published psychology studies faithfully, with larger samples, in consultation with the original authors. The results were published in 2015. Depending on the criterion used, somewhere between a third and a half of the findings replicated. Effect sizes in the repeats averaged roughly half the originals.



The 2015 replication project. Around a third to a half of a hundred published psychology findings held up when carefully repeated, and effects that did survive were typically about half the originally reported size. Comparable exercises in preclinical cancer biology found confirmation rates that were lower still.

Psychology took the reputational damage, largely because psychologists were the ones honest enough to run the audit. Where others have looked, the picture has often been worse: industry teams attempting to reproduce landmark preclinical cancer studies reported confirming only a small fraction of them.

The causes are the ones already described — small samples and therefore low power, forking paths, the file drawer, and a publication system that rewards novelty over verification and offers almost no career return for checking someone else's work. Not fraud. Ordinary practice, followed conscientiously, in a system whose incentives are misaligned with truth.

14

A profession warns about its own tool

In 2016 the American Statistical Association did something without real precedent: it issued a formal statement about the use and misuse of p-values. Professional bodies exist largely to promote their methods. This one published a set of principles warning practitioners that the central tool of their discipline was being widely misunderstood and misapplied — that a p-value does not measure the probability that a hypothesis is true, that it should not be used alone to decide anything, and that a threshold does not divide the world into real and unreal.

In 2019 a comment in *Nature* went further, calling for the concept of statistical significance to be retired altogether; several hundred scientists signed in support. Other groups proposed instead tightening the threshold from 0.05 to 0.005 for new discoveries. That the two camps disagree is itself informative: there is consensus that something is broken and none about the repair.

What has actually changed is mostly procedural, and is arguably the more valuable part:

- **Preregistration** — publish your hypothesis and analysis plan before collecting data, which closes the garden of forking paths by fixing your route in advance.
- **Registered reports** — journals accept a study on the strength of its design, before results exist, so that null findings are published too.
- **Effect sizes and intervals** reported alongside or instead of p-values, because how big is a better question than whether at all.
- **Data and code sharing**, so analyses can be checked rather than trusted.
- **Bigger samples**, and taking statistical power seriously at the design stage rather than as an afterthought.

15

Where it stands, which is unfinished

It would be easy to finish on a note of debunking, and it would be wrong. Statistics works. Randomised trials established that smoking causes cancer and that vaccines prevent disease. Statistical quality control rebuilt post-war manufacturing. Fisher's field designs underwrite modern agriculture. The methods in this essay are among the most consequential intellectual tools ever built, and the world is measurably better for them.

The trouble is narrower and more specific: a ritual assembled from two incompatible philosophies, taught as though it were one, applied by people who were never told which question it answers, and used to make a binary decision that the underlying mathematics does not support.

Set the three essays side by side and the comparison is instructive.

Field	Crisis	Resolution	Status
Calculus	Berkeley, 1734: is dx zero or not?	Cauchy and Weierstrass, 1821–1870s	Settled completely
Probability	What does the number mean?	Kolmogorov, 1933	Mathematics settled; meaning still argued
Statistics	How do you reason backwards?	—	Live, and public

Three foundations, three degrees of resolution.

So if statistics felt slippery to you as a student, that was not a failure of attention. You were being taught a hybrid whose two halves contradict each other, by a discipline that had not resolved — and has still not resolved — what its central number means or what you are entitled to conclude from it.

Calculus was fixed two centuries ago. Statistics is being argued about this year.

One thread runs through every failure in this essay. The lady tasting tea gives the chance of the result *if she is guessing*, not the chance that she is guessing. A p-value gives the chance of the data *if the null is true*, not the chance that the null is true. The twenty-outcome study, the forking paths, the misread significance test — all of them are the same swap, of the probability of the evidence given the hypothesis for the probability of the hypothesis given the evidence.

Reverse those two and you can convict an innocent person, approve a useless drug, or discover precognition in a respectable journal. It is the single most consequential confusion in modern science, it has a name, and it has a fix that has been available since 1763.

That is the third essay.

Written by Claude, the second of three essays on probability, statistics, and the confusion between them. Follows The Two Thousand Year Silence and The Ghosts of Departed Quantities.