

THE IN A HURRY SERIES

Statistics

in a Hurry

Signal, noise, and not being fooled

Max Brocklesby

BOOKSINAHURRY.COM

Contents

The Whole Thing in One Page 4

Why You Should Care 5

The Core Ideas 8

How It Actually Works 21

What People Get Wrong 31

Use It 36

Terms 41

Go Deeper 46

Notes and Sources 48

Bibliography 57

The Whole Thing in One Page

Statistics is usually shown to you after the decisive choices have already been made. A table appears. A formula is applied. A result receives three decimal places, and the decimals create an impression of fact being extracted by machinery. The arithmetic may be flawless while the conclusion is wrong.

The subject begins earlier, with a target: the population, outcome, period and comparison the analysis is meant to answer. Then data are made through definitions, instruments, categories and recording systems. A crime rate depends on offences, reporting and police practice. An unemployment rate depends on who counts as available and searching. A model learns from the people and events that entered its rows, together with the people and events that did not.

Then comes representation. A sample can speak for a population only through a defensible route from one to the other. In 1936 *The Literary Digest* received about 2.4 million election replies and confidently forecast Alf Landon. Franklin Roosevelt won by a landslide. Size had reduced random uncertainty around a sample distorted by coverage and response. It had not created representativeness.

Variation is the raw material, not the nuisance left after the average. Means can conceal skew and inequality. Groups can reverse an aggregate pattern. Repeated measurements from one source do not contain the information of independent observations. A signal exists only relative to a question; yesterday's noise can be tomorrow's target.

Inference then asks what unseen quantities are compatible with the observations. Standard errors and intervals describe uncertainty under a design and model. They do not cover every source of bias. Hypothesis tests compare data with a specified null. A p-value is not the probability that a claim is false, that chance caused the result or that the finding will repeat. Statistical significance does not measure importance.

The kind of claim matters. Description records a pattern. Prediction asks whether it forecasts new cases. Causal inference asks what an action would change. Correlation can serve the first two and contribute to the third, but causation needs a credible comparison between what happened and what would have happened otherwise. Randomisation can build that comparison. Observational work must earn it through design and assumptions.

Finally, count the search. Try enough outcomes, subgroups and models and noise will occasionally look impressive. A model can fit its training data by learning accidents. A literature can select positive findings. Transparency, multiplicity control, held-out testing and replication expose the opportunities a pattern had to be chosen.

The discipline reduces to one chain: state the target, define, measure, design, summarise, model, quantify uncertainty, name the claim, reveal the search and test it on new data. Each link can strengthen evidence. Each can also turn precision into theatre. The chain matters because no late calculation can recover a population that was excluded, a comparison that was never created or an outcome chosen after its success was known.

Statistics does not promise immunity from being fooled. It teaches you where the trick must enter, including when the person performing it is you. It replaces the authority of a naked number with a visible argument that another person can inspect, challenge and improve.

That is the book.

Why You Should Care

At Sally Clark's trial for the deaths of her two infant sons, a jury heard that the chance of two sudden infant deaths in a family like hers was about 1 in 73 million. The number sounded like a verdict because it was small, exact and attached to expert authority.

It was not a verdict. The calculation multiplied an estimated risk for one death by itself, treating the deaths as independent. That assumption had not been justified, and siblings share genetic and environmental factors that can

make the risks dependent. It also invited a second error: hearing the probability of two natural deaths as the probability that Clark was innocent. Those are different conditional probabilities. Guilt requires comparing how well the evidence fits competing explanations, not staring at one unlikely event.

Clark's convictions were overturned in 2003 after undisclosed microbiological evidence emerged. Bad statistics were not the sole basis of the appeal, which matters. The case still reveals what numerical mistakes can do when nobody in the room controls the question a probability answers.

Most encounters with statistics are less catastrophic and more frequent. A medicine is said to cut risk by half. A survey reports that the public has changed its mind. A company claims its new process raised conversion. A league table ranks schools. A forecast gives a 30 per cent chance of rain. An algorithm assigns fraud risk, creditworthiness or priority for treatment. Each result compresses people, measurements and assumptions into something small enough to act on.

That compression is indispensable. You cannot inspect every future customer, repeat history with and without a policy, or wait for certainty before treating a patient. Statistics lets a sample inform a population, an experiment distinguish treatment from background variation and a model turn past patterns into conditional forecasts. Modern medicine, manufacturing, insurance, polling, economics and scientific research could not operate at their present scale without it.

The danger is that the output looks cleaner than the process. A percentage hides a denominator. An average hides a distribution. A large sample hides selection. A narrow interval hides systematic error. A low p-value hides a large search. A prediction hides the population on which it was trained. A causal verb hides the comparison that would be needed to support it.

The same habits matter when nobody is publishing a paper. You compare the return from one investment with another period, judge whether a staff

change improved output, interpret a health test or decide whether a run of bad results means a system has broken. Small samples, regression towards the mean, selection and base rates enter ordinary decisions whether or not anyone calls them statistics.

Good statistical thinking also protects against a fashionable mistake: treating every uncertainty as proof that nothing can be known. Estimates can remain useful while imperfect. A study can narrow the plausible effects without naming one exact truth. Several weak pieces of evidence can combine, and one apparently strong result can collapse when its design is examined. Scepticism should discriminate rather than sneer.

This book will not teach you to calculate every test. Software already does that faster. It will teach the mental model that software cannot supply: what target had to be stated, what had to happen before the calculation, what the result means under its assumptions, what it does not mean, and what evidence would show that the pattern survives outside the dataset that produced it.

You should finish able to read a quantitative claim from both ends. Work backwards from the headline through estimate, model, design, sample, measurement and target. Then work forwards towards importance, decision and new-data performance. The number may survive. When it does, you will know why, which assumptions still carry weight and what new evidence would change the conclusion. That is more useful than either obedience to numbers or reflexive distrust of them.

The Core Ideas

The Target Comes Before the Data

Every analysis needs a target before it needs a dataset. Statisticians call the quantity or contrast to be learned an estimand: the proportion of eligible voters who would choose a candidate today, the average change in blood pressure caused by a treatment after twelve weeks, or the mean delivery time for all orders placed this month. Change the population, outcome, period, intervention or summary and you change the estimand. The same spreadsheet can support several questions, but it cannot decide which question matters. A precise answer to the wrong target is still wrong.

The spreadsheet also looks as though it arrived from the world already numerical. It did not. Before the first calculation, somebody decided what counted, what would be measured, how often, with which instrument, under which category, and what to do when reality refused to fit the boxes. Statistics begins with those decisions, not with the formula.

Take unemployment. A country cannot place every adult into a natural category labelled employed or unemployed. It must define work, availability, job search, age, reference week, unpaid family labour, temporary absence and the treatment of people who have stopped looking. Change those rules and the rate can move while nobody gains or loses a job. That does not make the number fraudulent. It makes it a measurement produced for a declared purpose.

The same applies to blood pressure, school performance, crime, poverty, customer satisfaction and intelligence. A cuff gives a reading under conditions that affect it. An examination samples part of what a pupil knows. Recorded crime combines offending, reporting, policing and classification. A satisfaction score turns a private judgement into one of perhaps five permitted responses. Each dataset contains both the phenomenon and the instrument used to see it.

This is why a variable needs an operational definition. Income can mean

gross, disposable, household, individual, annual, weekly, before housing costs or after them. A hospital readmission can be counted within seven days or thirty, for any cause or only the original condition. A claim without a stable definition is a moving target with decimals attached.

Measurement error enters in several forms. Random error scatters readings and can often be reduced by repetition. Systematic error pushes them in a direction, perhaps because a scale is miscalibrated or one group is less likely to disclose an answer. Misclassification places observations in the wrong category. A proxy stands in for something harder to observe, such as using postcode for deprivation or clicks for interest. The proxy may be useful while remaining incomplete.

Categories can also alter behaviour. A fraud model trained on confirmed investigations learns from a selection produced by earlier investigators. A school league table can turn a continuous range of attainment into a pass boundary that changes effort near the line. Once a measure is rewarded, punished or published, people respond to it. The data-generating process now includes the measurement system itself.

No analysis of that dataset alone can recover information that was never collected, repair an incoherent target or make an unsuitable proxy become the thing it represents. Data cleaning belongs to the same chain. Correcting impossible dates, duplicate records and unit errors improves evidence; silently removing inconvenient observations changes it.

The first defence against being fooled is therefore to state the estimand in ordinary language, then ask what each row represents, how each column was created, what could not enter and which decisions preceded the numbers. The statistical problem often lives in the distance between the target and the value recorded.

Design Determines What the Data Can Say

Most statistical questions concern things that have not all been observed. A poll asks a few thousand people about millions. A clinical trial recruits hundreds and hopes to learn about future patients. A factory inspects a fraction of its output. The visible sample is asked to speak about a larger population, a different treatment condition or events that have not happened yet. Whether it can do so depends on design.

Two uses of randomness are often confused. Random sampling selects observations from a population and supports inference back to that population under the sampling design. Random assignment allocates a treatment or condition and supports a causal comparison between the assigned groups. A study can have one without the other. A randomised trial of volunteers may estimate an intervention effect credibly for those enrolled while representing the wider patient population poorly. A probability survey may represent a population well while revealing nothing causal about changing an exposure.

Imagine an online shop that randomly assigns arriving website visitors to its old or new checkout. The comparison can estimate the effect of assignment among visitors during that test, subject to implementation, interference and chance. It does not automatically describe people who never visited, customers using another channel or behaviour during Christmas. Randomisation protects the comparison it created. It does not make the participants representative of every population the company might later mention.

Representation does not come from size alone. In 1936, *The Literary Digest* mailed roughly ten million presidential ballots and received about 2.4 million replies. The magazine forecast a clear victory for Alf Landon. Franklin Roosevelt won the popular vote by a landslide. Coverage problems in the mailing lists mattered, and later investigation found differential response: Landon supporters were more likely to return the ballot. Size reduced random fluctuation around a distorted sample. It did not create the route from respondents to voters that the claim required.

A sampling frame is the practical list or mechanism from which a sample is drawn. It may omit homeless people, households without stable internet, small firms outside a commercial database or patients who never reach a clinic. Non-response creates another filter after selection. Attrition creates one later, when participants disappear from follow-up. Missing records can damage an experiment as well as a survey when disappearance relates to treatment or outcome.

Random sampling helps because each unit receives a known chance of selection. Stratification can protect the representation of important subgroups. Clustering can make fieldwork practical. Weighting can reverse unequal selection probabilities and adjust some observed response differences. None performs magic. Weights cannot reconstruct absent people when the available variables do not explain why they are absent.

The square-root rule explains both the strength and the limit of size. Under common independent-sampling conditions, the standard error of many estimates falls roughly in proportion to one divided by the square root of sample size. Halving that part of the uncertainty often requires four times as many observations. Yet multiplying a biased sample makes the bias more stable. You become more certain about the wrong quantity.

Generalisability adds a further test. A sound estimate for volunteers in one hospital, employees in one company or voters in one election does not automatically travel to another place, period or population. Internal validity asks whether the comparison supports the claim within the study. External validity asks whether the result transports. They are separate achievements, and neither is conferred by a large number in the first line of the abstract.

Before trusting a study, reconstruct the journey from target population to frame, invitation, participation, assignment, follow-up and final analysis. At each gate, ask who could fall out, what randomness did and whether the design supports description, population inference or causation. Data inherit the permissions and limits of the process that produced them.

Variation Is the Subject

Statistics is often introduced as a machine for finding averages, as though its job were to flatten difference. Its deeper job is to understand variation: how values differ, which differences are stable, what structure explains them and how much remains uncertain.

Suppose two towns have the same mean household income. In one, most households sit near the mean. In the other, a few high incomes lift an otherwise lower distribution. The average is identical and the social reality is not. The mean uses every value and is sensitive to extremes. The median marks the middle observation and resists them. Neither is the universally correct centre. The choice depends on the distribution and the question.

Spread supplies information that centre cannot. The range records the two extremes and ignores everything between. The interquartile range describes the middle half. Variance averages squared departures from the mean, giving large departures extra weight. Standard deviation returns that spread to the original unit. Similar summaries can still hide different shapes, so a graph or table of percentiles often says more than one centre and one spread.

Outliers deserve investigation rather than automatic worship or deletion. A distant value may be a typing error, instrument failure, unit mismatch, unusual but valid case or evidence that the assumed model is poor. Removing it because it spoils a result edits the world to fit the analysis. Keeping a known error is no more principled. Return to the record, find how the value arose and report any consequential decision.

Distributions turn observations into patterns. Some are roughly symmetric. Some have long tails or several clusters because different groups or mechanisms have been combined. A mean waiting time can be dominated by a few extreme delays. An average treatment response can hide benefit for one group and harm for another. A single regression line can run through two separate clouds and describe neither.

A graph is often the fastest test of whether a summary has lied by omission.

Histograms expose skew and several peaks. Scatterplots reveal curves, clusters and leverage points. Time plots separate trend from season and one-off shocks. Graphs can mislead through truncated axes, uneven bins, area symbols and selective ranges, but the answer is disciplined display rather than retreat to tables. Show the data at a scale that preserves the question.

Variation also occurs at different levels. Pupils vary within schools, schools within regions and years within the same school. Repeated measurements on one person are more alike than measurements on strangers. Ignoring such dependence can make the dataset appear to contain more independent information than it does. Ten readings from one sensor are not equivalent to one reading from each of ten sensors when the target concerns sensor-to-sensor performance.

Statistical models divide variation into an account and a residual. The account may use group, time, dose, location or other predictors. The residual is what remains. Calling it noise does not make it meaningless. Residual variation can contain measurement error, omitted causes, nonlinear structure, interactions, real individual differences and randomness. Residual plots are therefore records of what the model failed to explain.

Signal and noise are conditional terms. A daily temperature fluctuation may be noise for estimating a century trend and signal for operating an electricity grid tomorrow. Individual differences may be noise in estimating an average drug effect and the main subject in personalised treatment. Statistics does not reveal one eternal signal hidden inside data. It asks which patterns matter for a declared target and whether they persist beyond the observations used to find them.

Estimates Need an Uncertainty Scale

A sample proportion, mean or regression coefficient is a point estimate: one numerical answer under a chosen procedure. Extra decimal places can make it look complete. It is not. Another defensible sample would have produced another value. An estimate without an uncertainty scale hides the part of the problem created by observing only one of many possible datasets.

The standard error describes how an estimator would vary across repeated samples under a specified model and design. It is not the standard deviation of the raw observations. The latter describes variation among people, products or measurements. The former describes uncertainty in an estimated quantity. A varied population can have a precisely estimated mean; a homogeneous-looking sample can still estimate the wrong population.

The central limit theorem helps explain why the normal curve appears so often in inference. Under broad but limited conditions, standardised sums and averages of many independent contributions approach a normal distribution even when the individual values do not. It does not make the population normal, and strong dependence or heavy tails can slow or defeat the approximation. The theorem supports a procedure; it does not excuse inspection of the data.

A confidence interval combines an estimate with a method calibrated to cover the true parameter at a stated long-run rate, provided its assumptions hold. With a conventional 95 per cent procedure, repeated studies conducted and analysed in the prescribed way would produce intervals containing the fixed target about 95 per cent of the time. Once one frequentist interval has been calculated, the parameter is inside it or outside it. The 95 per cent belongs to the procedure.

Ordinary language wants to say there is a 95 per cent chance the parameter lies inside. A Bayesian credible interval can support a probability statement about the parameter, but only after combining data with a prior distribution

and likelihood model. Bayesian and frequentist methods place probability in different parts of the problem. Neither removes the need to defend measurement, design and model.

Interval width depends on more than sample size. Greater variability widens it. Efficient design can narrow it. Clustering, missing data, weighting and model choice can widen it. A narrow interval can still miss the target systematically if an instrument is biased, a sample excludes a group or a model is wrong. Precision is not accuracy.

Uncertainty also extends beyond the part easiest to calculate. A poll depends on likely-voter definitions, weighting and response. A historical estimate depends on incomplete records. A demand forecast depends on future conditions resembling the past. A mathematically correct interval may therefore cover a smaller problem than the reader assumes.

Sensitivity analysis asks how conclusions change under plausible alternative decisions or assumptions. What happens when an outlier is retained, missing values are handled differently, the adjustment set changes or another measurement definition is used? Stability does not prove truth, but fragility reveals dependence on choices hidden by one estimate.

Methods with similar outputs can rest on different assumptions. A bootstrap resamples from the observed data as an empirical stand-in. A model-based interval uses a mathematical account of the data-generating process. A survey interval may incorporate strata, clusters and weights. The method belongs beside the result whenever those differences could alter interpretation.

Read an estimate as a package. What quantity is being estimated, for which population and period? Under what design and model? Which uncertainties enter the interval, and which remain outside it? How large is the effect in useful units? A responsible estimate carries an uncertainty scale without pretending that every uncertainty has been captured.

Tests Measure Compatibility, Not Truth

A statistical test begins by specifying a model, often including a null hypothesis such as no average difference between groups. It then asks how unusual a chosen test statistic would be if that model generated the data. The p-value is the probability, under the specified assumptions, of obtaining a result at least as incompatible with the null model as the one observed.

That sentence has a direction. It starts by assuming the null model and calculates the probability of data. It does not calculate the probability that the null is true after seeing the data. It does not give the probability that chance alone caused the result. It does not say how likely the finding is to replicate. Reversing a conditional probability can turn a valid calculation into a false conclusion.

The Sally Clark case shows the stakes. At her murder trial, an expert presented roughly 1 in 73 million for two sudden infant deaths under selected family characteristics. The figure came from multiplying an estimated probability for one death by itself, as though the deaths were independent. The calculation offered no sound basis for that assumption, and shared genetic and environmental factors could make the risks dependent. More fundamentally, even a correct probability of two natural deaths would not equal the probability that Clark was innocent. That would require comparison with the probability of the evidence under competing explanations and relevant prior information.

Clark's convictions were overturned in 2003 after previously undisclosed microbiological evidence came to light. The statistical errors did not, by themselves, decide the appeal. They remain a clean example of how a tiny probability attached to evidence can be heard as a tiny probability of innocence when those are different questions.

Thresholds add another distortion. A convention such as p below 0.05 turns a continuous measure into a pass or fail label. Results at 0.049 and 0.051 contain almost the same information but may be described as discovery and no effect. Large samples can make small, unimportant differences produce

low p-values. Small studies can miss effects large enough to matter. The test result must be read beside the effect estimate, its uncertainty, the design, prior evidence and the consequences of error.

Power is the probability that a test will reject a specified null under a specified alternative. It depends on effect size, variability, sample size, design and threshold. Low power means a real effect may be missed. Among results that happen to cross a publication threshold, low-powered studies can also produce unstable and exaggerated estimates. Planning sample size therefore requires a meaningful target effect and an acceptable balance of errors, not a ritual number of participants.

In 2016 the American Statistical Association issued six principles to curb common misuse. They stress that p-values indicate incompatibility with a specified model, do not measure hypothesis truth or practical importance, should not alone determine decisions, and require full reporting. A 2019 editorial urged researchers to abandon bright-line declarations of statistical significance, but it was written by three editors as their view rather than formal ASA policy. A 2021 ASA task force then clarified that properly used p-values and significance tests remain useful. The shared ground is against mechanical interpretation, not in favour of one compulsory replacement.

A test is a tool for disciplined comparison between data and a model. It can sharpen evidence. It cannot issue a verdict on truth, importance or action without the rest of the evidence chain.

Description, Prediction and Causation Are Different Jobs

A relationship between two variables can serve several purposes. Description records how they vary together. Prediction uses one to improve forecasts of another. Causal inference asks what would happen to an outcome if an intervention changed something. These jobs may use the same regression equation and demand different evidence.

Ice-cream sales can predict drownings because both rise in hot weather. Sales do not cause the deaths. Temperature is a common cause, or

confounder, creating an association between variables that would not survive the right comparison. Confounding occurs when groups differ in ways related to both exposure and outcome. Regression can adjust for measured differences, but the decision about what to adjust for is causal reasoning, not a service the software supplies automatically.

The Berkeley graduate admissions data from 1973 became famous because the aggregate acceptance rate appeared lower for women while many department-level comparisons looked different. Women had applied more often to departments with lower acceptance rates for everyone. Aggregation had mixed the department composition into the sex comparison. This is often presented as Simpson's paradox, where a pattern reverses or changes after conditioning on a third variable.

The arithmetic is not contradictory. The question changes. The overall rate describes admissions across the actual mix of applications.

Department-level rates compare applicants within departments. Neither automatically answers whether any process was discriminatory. Department choice may itself reflect prior constraints, and departments may differ in applicant pools and procedures. The example teaches that conditioning can reveal structure, while the choice of condition needs substantive justification.

Randomised experiments provide a stronger route to causation. If treatment is assigned by a genuine random mechanism, treated and control groups are comparable in expectation before treatment. Differences in outcomes can then be attributed to assignment, subject to implementation, attrition, interference and chance. Randomisation does not guarantee perfect balance in one realised trial. It provides a known assignment process from which uncertainty can be calculated.

Observational studies cannot rely on assignment by the researcher. They use design and assumptions: matching, natural experiments, instrumental variables, discontinuities, longitudinal comparisons, negative controls and careful adjustment. Each method answers a causal question under conditions that can fail. Austin Bradford Hill's 1965 discussion of association

and causation offered considerations such as temporality, strength, consistency and dose response, while warning against turning them into a checklist. Modern causal inference makes the target intervention and counterfactual comparison more explicit.

Prediction follows another standard. A variable can predict well without being a cause. Credit history can forecast default without causing it. A medical risk score can rank patients while saying little about which treatment would help them. Predictive success is judged on new cases through discrimination, calibration and error. A causal estimate is judged by whether the comparison supports the intervention claim. A model optimised for one job can fail at the other.

Prediction and causation can even pull model building in opposite directions. A variable recorded after treatment may improve a forecast while making a treatment-effect estimate nonsensical. A stable proxy may predict well while offering no safe intervention. Conversely, a randomised trial can estimate the effect of one action with modest predictive accuracy for individuals. Good analysis starts from the decision, not from the most impressive model available.

Before interpreting any association, name the verb. Are we describing what co-occurs, predicting what comes next or estimating what an action would change? Then demand evidence suited to that verb. Many statistical arguments become clearer once one equation is no longer asked to perform three different jobs.

The Search Must Face New Data

A dataset can contain patterns that arose through the same chance variation the analyst is trying to see through. Search widely enough and some will look persuasive. The problem is not confined to fraud. It follows from ordinary curiosity, flexible analysis and incentives that reward a clean result more than an honest dead end.

Suppose twenty independent tests are run, every null hypothesis is true and each uses a 5 per cent threshold. The probability that at least one crosses

the line is 1 minus 0.95 raised to the twentieth power, about 64 per cent. That calculation depends on independence and true nulls, but the lesson travels: the chance of a striking result depends on how many opportunities existed to find one.

Multiplicity can hide in outcomes, subgroups, time windows, transformations, stopping rules, exclusion criteria and model specifications. If only the successful route is reported, the reader sees one analysis while the data have faced dozens. The reported p-value then answers a smaller question than the research process asked. Selective publication adds a second filter, because positive results are more likely to appear and be cited.

Corrections depend on the goal. Family-wise error methods limit the chance of any false rejection within a defined family of tests. In 1995 Yoav Benjamini and Yosef Hochberg offered a method for controlling the false discovery rate, the expected proportion of false rejections among the hypotheses rejected, under stated conditions. This is useful when searching many signals, but it does not label any one discovery as true or false. The family of tests and the selection process still need to be declared.

Prediction exposes the same problem as overfitting. A model with enough flexibility can learn accidents of its training data. Its fit improves while its ability to predict new cases deteriorates. Holding back a test set creates a clean examination, provided it remains untouched until the end. Cross-validation rotates the held-out portion to estimate performance more efficiently. Repeatedly checking the same test set while tuning the model turns it into training data by another name.

Replication applies the new-data principle to scientific claims. Reproducibility, in the National Academies' usage, means obtaining consistent computational results using the same data, code and methods. Replicability means obtaining consistent results in a new study addressing the same question. Terminology varies, but the distinction matters. A result can be computationally reproducible and empirically fragile, or difficult to reproduce because the code and data were never made available.

In 2015 the Open Science Collaboration reported attempted replications of 100 experimental and correlational studies from three psychology journals. Ninety-seven of the original studies had statistically significant results; 36 per cent of the replications did, and replication effect estimates were on average about half the originals. The project became an emblem of a credibility problem. It was also one organised sample from one part of psychology, with disputes about replication design and what counts as success. It is evidence of a problem, not a universal replication rate for science.

Preregistration records hypotheses and analysis plans before outcomes are known. Registered reports move peer review before results. Open data and code permit checking. Multiverse and specification-curve analyses show how results vary across defensible choices. None guarantees good research. They change which parts of the search remain visible.

The causal loop closes here. Data were made through choices at the beginning. Analysis adds more choices. A pattern that survives only one route through one dataset may be a fitted accident. Trust grows when the route is declared, the search is counted and the claim performs on observations that did not help create it. New data are not a ceremonial final step. They are where signal proves that it can leave home.

How It Actually Works

Counting the dead

People had counted populations, land, harvests and taxes for millennia. In seventeenth-century London, death arrived as a weekly table that invited a different use. Parish clerks gathered reports of burials, assigned causes such as plague, consumption, fever or teeth, and published the totals in the Bills of Mortality. The categories were rough and the diagnoses uncertain. Their recurrence made it possible to inspect population patterns across weeks and years rather than treat each total in isolation.

John Graunt, a London haberdasher, took the bills seriously enough to

compare them. In 1662 he published *Natural and Political Observations Made upon the Bills of Mortality*. He looked for regularities beneath erratic weekly totals, estimated the city's population, compared male and female births, tracked the geography of death and built a rudimentary account of survival by age. His raw material was imperfect, but his move was decisive. Individual deaths became a population pattern.

The work already contained the permanent tension of statistics. Aggregation revealed facts no single life could show, while the categories reflected a weak recording system. A death listed under teeth may have had another cause. Some parishes reported better than others. Graunt reasoned through those defects rather than waiting for perfect data that would never arrive.

Three decades later Edmond Halley used birth and funeral records from Breslau to construct a life table. The table estimated how many people from a notional cohort would remain alive at each age. Halley then connected survival to the price of life annuities. A promise to pay money for as long as someone lived could now be priced from collective experience rather than guesswork. Statistics had moved from describing mortality to organising a financial decision under uncertainty.

These early projects were administrative as well as intellectual. States wanted populations counted for taxation, armies, health and control. The word statistics grew from descriptions of the state. Tables made people legible to institutions, and institutions shaped which people and events became countable. The discipline's public usefulness and its power over classification grew together.

The average person and the residual

By the nineteenth century, governments and scientists were collecting more measurements of bodies, crime, education, trade and weather. Adolphe Quetelet proposed the average man as a way to understand social regularity. Individual variation could be treated as scatter around a central type, much as observational errors clustered around a physical quantity.

The comparison was productive and dangerous. An average can reveal

stability across a population. It can also turn a descriptive centre into an ideal and label departure as defect. Human differences are not always measurement errors around one true person. They may be the phenomenon, the result of several populations, or evidence that the chosen category has combined unlike cases.

Francis Galton pushed measurement towards heredity. Studying the heights of parents and children, he described regression towards mediocrity, later called regression towards the mean. Exceptionally tall parents tended to have children closer to the population average, though still taller than average; the same held in the other direction. Galton and later Karl Pearson developed correlation and regression into formal tools for measuring association.

Regression towards the mean is now one of the most useful traps to recognise. Select people because they had an extreme result, and their next result will often be less extreme even if nothing has changed. A school chosen for an unusually bad year may improve. Patients recruited during a symptom flare may feel better. A fund celebrated after an exceptional run may cool. Without a comparison group, ordinary reversion can be mistaken for the effect of punishment, treatment or skill.

Galton's work also sat inside eugenics, a programme that tried to rank human worth and direct reproduction. The statistical methods outlived that project, but their history is not separable from it. Correlation can describe a pattern while classifications, samples and interpretations carry social assumptions. A technically competent calculation can serve a corrupt target.

Sampling the public

A census tries to count everyone. That ambition is politically attractive and statistically incomplete. People are still missed, duplicated, misclassified or recorded under definitions that change. For many questions, a well-designed sample can produce a better estimate at far lower cost, while making its sampling uncertainty visible.

The development of survey sampling turned representation into a design

problem. Units could be selected with known probabilities, sometimes within strata designed to protect the representation of smaller groups. Cluster samples reduced travel by selecting areas before households. Weights then reversed unequal selection probabilities so each observation represented an appropriate share of the population. The estimate and its uncertainty came from the selection process, not from a claim that the sample looked vaguely typical.

The 1936 American election exposed what happens when that link fails. *The Literary Digest* had correctly forecast several earlier winners and believed the scale of its new poll would settle the matter. Its mailing lists, drawn from sources including telephone directories and car registrations, undercovered parts of the electorate. Its voluntary return created another filter. Evidence examined decades later indicates that Landon supporters returned ballots at higher rates. Coverage and non-response worked together.

George Gallup's much smaller poll picked Roosevelt as the winner and gained fame from the contrast. His quota methods were not the probability sampling that later became the standard, and his estimate still understated Roosevelt's margin. The useful lesson is therefore narrower than the legend. Millions of answers did not compensate for an unknown relation between response and vote. A smaller design built around representation did better.

Survey statistics later developed methods for complex samples, non-response adjustment, post-stratification and repeated polling. Every method preserved a condition: correction is strongest when the variables related to participation are observed. If people vanish for reasons connected to an unmeasured outcome, weighting by age, region and sex cannot promise to reconstruct them. The missing people remain part of the estimand problem, even when the final table has no blank cells.

Beer, barley and small samples

At the Guinness brewery in Dublin, William Sealy Gosset faced a problem that giant samples could not solve. Industrial experiments cost time and material. Barley varieties, brewing conditions and laboratory measurements were tested in small batches, and the population variability had to be estimated from the same few observations used to estimate the mean.

Large-sample approximations understated that extra uncertainty. Publishing as “Student” in 1908, Gosset derived the distribution now attached to the t statistic. With few observations, the reference distribution has heavier tails than the familiar normal curve. Extreme sample means are more plausible because the estimate of spread is itself unstable. As the sample grows, that extra uncertainty diminishes and the t distribution approaches the normal.

The achievement was more than a new table. Gosset showed that sample size changes the shape of uncertainty, not merely its amount. A method appropriate for thousands of observations can be overconfident with ten. Small-sample inference had to account for the fact that the yardstick was being estimated alongside the quantity of interest.

His work joined theory to an operating decision. Guinness did not need a declaration that two barley varieties were different in some abstract population. It needed to know whether the evidence was strong enough to change production. The costs of a false alarm, a missed improvement and another experiment all mattered. Statistical inference was becoming a technology for choosing under limited information.

Designing the comparison

When Ronald Fisher arrived at Rothamsted Experimental Station in 1919, decades of field trials had accumulated across strips of English soil. Crop yields varied with fertiliser, weather, soil fertility, pests and the history of each plot. Analysis after harvest could not untangle every difference if the treatments had been placed badly.

Fisher systematised an answer: design the uncertainty before measuring

the outcome. Randomisation assigned treatments through a chance mechanism, preventing systematic preference from deciding which plots received which treatment. Blocking grouped similar plots so comparisons were made within more homogeneous sets. Replication supplied repeated comparisons and an estimate of experimental variation. Factorial designs tested several factors together and revealed interactions, where the effect of one treatment depended on another.

Randomisation changed the logical basis of the comparison. If treatments were assigned by a known random procedure, the possible reallocations described what results could have occurred under a null of no treatment effect. The analyst no longer had to pretend the field was naturally uniform. The design manufactured a defensible comparison inside a variable world.

Fisher's 1925 *Statistical Methods for Research Workers* and 1935 *The Design of Experiments* spread these ideas far beyond agriculture. The latter used an eight-cup tea-tasting problem to demonstrate how a test could be built from the randomisation scheme. Its historical status as a social event is less secure than its role in the book. The published example matters because the design determines the probability calculation: four cups prepared each way, all eight classified, and a fixed set of possible assignments.

Design also showed why more analysis cannot rescue a missing control. If every treated plant grows on one side of the field and every untreated plant on the other, treatment is confounded with location. A model may adjust for recorded soil measures, but no calculation can prove that every relevant difference was captured. Good experiments remove alternative explanations before they appear.

Two theories become one ritual

Fisher treated the p-value as a graded measure of discrepancy between data and a null model. A low value could prompt doubt about the null, in the context of scientific judgement and repeated work. Jerzy Neyman and Egon Pearson developed a different framework. They defined competing hypotheses, type I and type II errors, fixed long-run error rates and tests chosen for power against specified alternatives.

Neyman's 1937 theory of confidence intervals added procedures designed to cover the true parameter at a chosen frequency across repeated samples. These methods were operational. Before seeing the data, decide which errors matter, choose the rule and accept its long-run performance. The conclusion was an action governed by a procedure rather than a probability assigned to one hypothesis.

Textbooks and research practice later blended the traditions. Researchers state a null hypothesis, calculate a Fisherian p-value, compare it with a Neyman-Pearson threshold, call the result significant and sometimes speak as though the procedure delivered the probability that the null is false. This hybrid is convenient. Its pieces do not automatically support that interpretation.

Across many journals and research institutions, the 5 per cent line became administrative. A single threshold was easy to teach, audit and reward wherever a binary decision was demanded. Yet a convention designed to control one kind of error in a planned procedure can become misleading when treated as a border between truth and falsehood. The same evidence does not change character between p equals 0.049 and p equals 0.051.

Effect estimates and intervals offer a fuller account. Power calculations connect sample size to effects worth detecting. Decision analysis connects errors to consequences. Bayesian methods combine prior information with a likelihood to produce a posterior distribution. None is a universal solvent. Each makes assumptions and answers a particular question. The central advance was learning to separate the numerical result from the operating

rule used to act on it.

From association to intervention

As observational data expanded, statisticians faced questions that randomisation could not always answer. Smoking could not ethically be assigned for decades to estimate lung-cancer risk. Poverty, education, pollution and occupation arrive through social systems rather than laboratory allocation.

Austin Bradford Hill's 1965 address on association and causation set out considerations including strength, consistency, temporality, biological gradient, plausibility and experiment. He did not present them as necessary boxes or a scoring system. His deeper message was that causal judgement combines the pattern of association with time order, mechanism, alternative explanations and the wider body of evidence.

Modern causal inference sharpens the question by defining a contrast between possible outcomes under different actions. What would happen to the same target population if it received treatment rather than control? No person reveals both outcomes, so research design must create a credible substitute comparison. Randomisation does it by assignment. Observational methods attempt it through measured adjustment, natural experiments, discontinuities, instruments, longitudinal structure or other assumptions.

Regression remains useful but cannot decide causation by coefficient. Adjusting for a common cause can reduce confounding. Adjusting for a variable caused by the treatment can block part of the effect. Conditioning on a common effect of treatment and outcome can create an association that was not there. The model needs a causal story before the software needs a formula.

At the same time, John Tukey argued in 1962 that data analysis was wider than formal mathematical statistics. Computers made it possible to inspect residuals, transform variables, explore many relationships and develop methods through use. Exploratory data analysis later gave graphs, resistant summaries and pattern-finding a central place. Confirmation and exploration

were not enemies, but they needed different reporting. A pattern noticed in the data should face new data before being treated as a planned test.

Computing also widened prediction. Models could be judged by how well they forecast held-out cases rather than how closely they fit the observations that trained them. Resampling methods such as the bootstrap estimated uncertainty by repeatedly drawing from the observed data under an empirical approximation. Cross-validation tested predictive performance across withheld portions. The discipline was becoming less dependent on one closed-form formula and more dependent on explicit procedures that could be checked.

Missing data became a theory of its own because a blank can carry information. Values missing for reasons unrelated to observed or unobserved outcomes create a different problem from values missing after a patient's condition worsens or an unhappy customer leaves. Donald Rubin's 1976 framework clarified why deletion, imputation and likelihood methods depend on assumptions about the missingness process. Filling every cell does not make those assumptions disappear.

The credibility repair

The new flexibility created new failure modes. A researcher could try several outcomes, covariates, exclusion rules and stopping points, then report the route that produced a low p-value. The final paper might contain no false arithmetic. It could still conceal the number of chances noise had to win.

Multiplicity procedures had long addressed families of tests. A 1995 paper by Yoav Benjamini and Yosef Hochberg introduced a practical method for controlling the false discovery rate in large searches. The target suited fields testing thousands of signals, where preventing any false positive could be too severe. The method changed the question from whether the family contained one error to the expected proportion of errors among reported discoveries.

Research practice then became part of the statistical model. Publication bias selected positive findings. Low power made estimates unstable.

Undisclosed analytical flexibility inflated false-positive rates. In 2011 Joseph Simmons, Leif Nelson and Uri Simonsohn demonstrated through simulations and experiments how common freedoms in data collection and analysis could manufacture apparently significant results. Their argument was not that researchers routinely invented data. It was that ordinary discretion, hidden from the reader, changes the meaning of the reported test.

Large replication projects tested whether published patterns travelled. The Open Science Collaboration's 2015 psychology project found weaker replication results across its 100-study sample than the original literature suggested. Other fields and projects have produced different rates and raised disputes about fidelity, power and heterogeneity. The lesson is not one percentage for science. It is that publication is a selection event, and a result's first appearance cannot be the final calibration of its reliability.

Reforms try to expose the chain. Preregistration records plans. Registered reports review the question and design before outcomes. Sharing data and code permits computational checking. Reporting guidelines require clear outcomes, exclusions and model choices. Replication brings new observations. Meta-analysis combines studies while confronting heterogeneity and publication bias. Better incentives matter because no formula can correct choices that remain invisible.

One modern strand traced here moved from extracting population patterns from crude death lists to systems that can test millions of associations and update models continuously. Scale has changed. The old problem remains: how to compress variable evidence without mistaking the apparatus of measurement and selection for the world itself.

How we know

The historical spine uses primary publications by Graunt, Halley, Gosset, Fisher, Neyman, Pearson, Tukey, Hill, Benjamini and Hochberg, checked against modern histories by Stephen Stigler and Theodore Porter. It is a selective account of traditions institutionalised largely in Britain, Ireland, continental Europe and the United States, not a claim that counting or quantitative administration began there. Priority language is narrow because methods emerged through precursors, collaborators and later reinterpretation.

Technical explanations follow original methodological papers, standard texts and current professional guidance. Framework-specific terms such as confidence, power, reproducibility and replicability are identified rather than presented as universal vocabulary. Fisher's tea problem remains a published design illustration, not a reconstructed scene. The online checkout example is explicitly illustrative. The Open Science Collaboration supplies evidence about a defined sample of psychology studies, not a census of science. Current p-value and ethics wording was rechecked against official American Statistical Association material on 3 September 2026. The evidence-chain model is an editorial synthesis whose value is practical coherence, not proof that every statistical school accepts one workflow.

What People Get Wrong

“The average describes the typical person”

An average is a compression, not a person. The mean of five salaries can be lifted by one executive until almost everyone earns less than the reported average. The median can hide how far the extremes extend. A mode may not exist or may depend on arbitrary grouping. Even a well-chosen centre says nothing about spread, skew, clusters or the people outside the middle.

The mistake survives because one number is easy to compare and institutions need rankings. It also borrows authority from the normal

distribution, whose bell shape makes a centre look like the natural description of a population. Real distributions can be lopsided, bounded, heavy-tailed or mixtures of several groups. Even a bell-shaped variable does not turn its mean into a model citizen.

The correction is not to ban averages. It is to ask which centre was chosen, inspect the distribution and match the summary to the decision. An average treatment effect can guide policy while concealing different responses. An average journey time may be useless to someone who must arrive by nine. The typical case is an empirical question, not a synonym for the mean. When the distribution contains distinct groups, the honest answer may be that no single typical person exists.

“A bigger sample fixes a bad sample”

More observations reduce sampling variation under a sound design. They do not repair exclusion, self-selection, non-response, mismeasurement or a target population that was never defined. A huge online poll can describe its respondents precisely while missing people less likely to see, trust or answer it.

The *Literary Digest* failure remains memorable because 2.4 million replies lost to a much smaller poll. Coverage bias and differential response made the sample unrepresentative. A larger sample would have narrowed the margin of sampling error around the distorted answer and made the confidence look stronger. That is the dangerous combination: low random uncertainty with high systematic error.

The lesson has grown more important in the age of digital traces. A platform can contain billions of clicks and still reveal little about people who do not use it or about motives a click cannot measure. Size tells you how much data entered. Design tells you what those data are evidence about. Before admiring the sample count, ask how the first observation became eligible and how the last one remained.

“A p-value is the probability that the result is due to chance”

A p-value assumes a statistical model, including a null hypothesis, and asks how extreme the observed result would be under it. It does not divide the world into chance and cause. Chance enters every sampling model, including studies of real effects. Bias, confounding and measurement error can produce low p-values, while noisy data can produce high ones when an effect exists.

Nor is the p-value the probability that the null hypothesis is true. That reversal requires information about plausible alternatives and prior probabilities. The same tail probability can arise from a small, precise departure or a large, noisy one. It can also be invalid when observations are dependent, the stopping rule was flexible or the model's assumptions fail.

A result at p equals 0.03 may be weak evidence in a field searching thousands of unlikely claims and stronger evidence in a tightly designed confirmation of an established mechanism. The calculation has meaning only inside the design, model and search that produced it. A p-value can be exact for its stated problem while the headline answers a different problem entirely.

“Statistically significant means important”

Significance concerns compatibility with a model and threshold. Importance concerns size, consequences, costs and values. With a large enough sample, a tiny difference can cross a conventional line. With a small sample, a consequential effect can remain uncertain and fail to cross it.

Report the estimate in meaningful units. A treatment may reduce risk by one percentage point, which means something different when baseline risk is 2 per cent than when it is 50 per cent. Relative changes can sound large while absolute changes remain small. Standardised effect sizes can help compare unlike scales, but they may obscure the units in which costs and benefits are felt.

Intervals show which effect sizes are still compatible with the data. Decisions then require harms, benefits and opportunity costs. A threshold can organise a procedure. It cannot decide what matters. The correction matters because significance language often turns a weakly consequential result into a headline and an uncertain but important result into silence. Borderline labels also encourage researchers to treat near-identical evidence as opposites.

“Correlation tells you nothing about causation”

The warning that correlation does not prove causation is correct. The stronger slogan, that association provides no causal information, is not. Causes produce associations under many conditions, and causal reasoning often begins by finding them. The issue is whether the observed comparison rules out credible alternatives.

Randomisation supplies a known assignment mechanism. Observational studies can use time order, dose patterns, natural experiments, negative controls, discontinuities, instruments and adjustment based on a defensible causal structure. None turns a coefficient into proof. Several can converge on the same explanation, and a causal claim grows stronger when distinct designs fail in different ways yet point in the same direction.

Equally, refusing all causal learning from observational evidence would discard much of epidemiology, economics and public policy. Correlation is evidence whose force depends on design, assumptions and the alternatives it survives. The slogan should slow causal language, not end causal reasoning. The right response to an association is to ask what comparison, mechanism or design would distinguish the proposed cause from its rivals. Sometimes no decisive design is available. Then causal confidence should reflect the convergence of imperfect evidence rather than a claim that one coefficient settled the matter.

“A model that fits the data has found the pattern”

Fit measures how well a model describes the observations used to estimate it. Flexible models can fit noise, especially when predictors are numerous and the sample is small. Adding terms often improves in-sample fit even as predictions on new cases get worse.

Residual checks can expose missed structure. Penalisation can restrain complexity. Cross-validation and held-out test data estimate performance away from the training set. External validation asks whether the model travels to another place, period or population. Calibration checks whether predicted probabilities match observed frequencies.

Even validation can be contaminated. If researchers repeatedly compare models on one test set, choose the winner and return to the same set after every change, the supposedly new data become part of development. Performance will tend to look better than it is. The decisive evidence is not how closely the model hugs its own data. It is how well the claimed pattern survives observations that did not teach the model what to say. A model that cannot leave its training set has memorised a dataset, not discovered a dependable relationship.

“Data are objective; only interpretation is biased”

Recording a value can be disciplined, transparent and repeatable. That does not make the dataset independent of human choices. Definitions determine who is unemployed, which incident becomes a crime, what counts as recovery and how a household is formed. Instruments have detection limits. Missingness follows social processes. Historical datasets often preserve what institutions cared to record.

Interpretation adds further choices, but bias can enter before analysis begins. Administrative data are collected to run a system, so their categories may follow payment, enforcement or workflow rather than the research question. Historical changes in policy can create apparent trends even when behaviour is stable. Automated data inherit the sensors, interfaces and previous decisions that generated them.

The remedy is not cynicism about every number. It is provenance: clear definitions, documented collection, quality checks, access to metadata, visible exclusions and sensitivity to alternative classifications. Objectivity is an achievement of procedures that permit criticism. It is not a property conferred by placing observations in a table. The most trustworthy dataset is often the one whose limitations are easiest to trace rather than the one presented as frictionless fact. This is why metadata, revision histories and collection notes matter. They reveal the human process without turning every measurement into opinion.

Use It

Find the denominator and the target

A number without its comparison base is an invitation to supply the wrong one. “Cases doubled” can mean two became four. “Risk fell by 50 per cent” can mean two people in a hundred became one. “Most users improved” says little if dissatisfied users left before follow-up.

Begin by naming the numerator, denominator, population, period and unit. Ask whether the denominator changed. A falling crime rate can coexist with more recorded crimes if the population grew. A higher hospital death rate can reflect treating sicker patients. A fund can advertise the return of surviving products while vanished funds disappear from the comparison.

Then state the target in words. That sentence is the estimand the number is meant to answer. Is the claim about everyone, respondents, customers who remained active, patients who completed treatment or people eligible under a narrow definition? Many numerical disputes are target disputes hidden beneath arithmetic. Write the target beside the number in one sentence. If the sentence becomes awkward, the statistic may be answering a narrower question than the claim.

Reconstruct the route into the dataset

Imagine a funnel from the people or events of interest to the rows finally analysed. Who was eligible? Who could be contacted? Who agreed? Who supplied usable data? Who remained until the end? Which records were linked successfully? Which observations were excluded after collection?

At each stage, ask whether entry or exit could relate to the outcome. If a treatment is involved, separate how people entered the study from how the treatment was assigned. A workplace survey may miss former employees. A review score includes people motivated to post. Police data contain detected and recorded offences, not all offending. A fitness app sees people who own the device and keep wearing it. None of those sources is useless. Each supports a claim with a boundary.

Weights, imputation and adjustment can reduce known imbalances. They depend on observed information and assumptions. Compare respondents with the sampling frame where possible, inspect attrition by earlier outcomes and test whether conclusions change under plausible missing-data models. The best protection is a visible flow from target to analysed sample, with losses counted and reasons given.

Separate estimate, uncertainty and threshold

A reported result often compresses three different objects. The estimate says what the data suggest. The uncertainty describes how much the estimate could vary under the method and assumptions. The threshold supplies a rule for action or labelling.

Read them separately. A difference of 4 points with an interval from 1 to 7 is a different claim from “ p below 0.05”. The first shows magnitude and compatible values. The second records a test result against a specified null. A regulator, doctor or manager may still need a threshold, but the threshold should come from consequences and policy rather than masquerading as a property of nature.

Look for the effect in units a decision-maker can understand, the interval or

sensitivity range around it, and the rule that turns evidence into action. Ask what effect the study was designed to detect and which effects the interval still permits. When only the label “significant” appears, most of the useful information is missing.

Name the job

Ask whether the analysis is describing, predicting or estimating an intervention effect. A chart of customers who churned describes a group. A model that forecasts churn predicts. A trial of a retention offer estimates what the offer changes. These are related projects, not interchangeable ones.

For description, inspect coverage, measurement and summaries. For prediction, demand performance on new cases, calibration and comparison with a sensible baseline. For causation, demand a credible counterfactual comparison, clear timing and an account of confounding, selection and interference.

Language gives the first clue. “Associated with” is descriptive. “Forecasts” is predictive. “Leads to”, “reduces” and “because” are causal. Then inspect the target: predicting who receives a treatment is different from predicting who would benefit from it. When the evidence performs one job and the headline uses another verb, the claim has outrun the design.

Count the search

A surprising result becomes less surprising when it is the winner of a large search. Ask how many outcomes, subgroups, time periods, transformations, models and stopping rules were available. Was the hypothesis specified before the data were inspected? Are unsuccessful analyses visible? Was the threshold adjusted for multiplicity, or was the result confirmed independently?

This applies outside science. A trader can find a strategy that would have worked in past prices after trying thousands. A business can identify a customer segment with spectacular growth by slicing until one appears. A

sports analyst can discover a player's magical performance on rainy Tuesday evenings after searching enough conditions.

The correction need not forbid exploration. Exploration generates questions. It should be labelled, and its discoveries should face new data. Ask whether the reported family of tests matches the family that was available in practice. The hidden search is the problem because it makes one selected result look like one planned test.

Ask for new-data performance

A model, forecast or claimed relationship should be evaluated away from the observations that produced it. For prediction, ask whether there was a test set, whether it remained untouched during development, and whether performance holds across time and groups. Accuracy alone may hide class imbalance, so inspect false positives, false negatives, calibration and the cost of each.

For scientific claims, ask whether the analysis can be reproduced from the same data and code, whether the result has been replicated with new data, and whether effect sizes agree rather than merely sharing a significance label. For policy, ask whether a result from one setting transported after institutions, populations and implementation changed. A programme can retain its name while delivery, uptake and surrounding services alter the intervention being evaluated.

New data can reveal decline without proving dishonesty. The original estimate may have been noisy, the context may differ or the intervention may be implemented differently. Compare the populations, measurements and procedures before calling the studies inconsistent. The question is whether the claim was stated broadly enough to survive the test it is now failing.

The limits

Statistical discipline does not remove judgement. There is no calculation that chooses the moral value of an outcome, decides which harms matter or determines how much uncertainty society should tolerate. Models simplify. Measurements omit. Randomised trials may be unethical or unrepresentative. Observational studies may depend on assumptions that cannot be settled from data alone.

Good methods can also support bad purposes. More accurate surveillance can invade privacy. A fair estimate of average benefit can conceal an unjust distribution. Classification can harden administrative categories into identities. Statistics can clarify a choice without making the choice legitimate.

Nor does scepticism confer immunity. Demanding impossible certainty is another way to misuse uncertainty. Evidence can be strong without being perfect. Delaying action has outcomes too, and those outcomes also need a comparison. The task is to locate the remaining uncertainty, compare it with the consequences of delay and avoid pretending that ignorance sits on only one side of a decision.

The one thing to keep

Keep the chain.

When a number asks for belief, move backwards. What claim is being made? Is it descriptive, predictive or causal? What estimand does it target? What estimate supports it, with what uncertainty? How much searching preceded it? Which model converted observations into the result? Who entered the sample and who did not? How were the variables defined and measured?

Then move forwards. Does the effect matter in real units? Does the decision threshold fit the costs of error? Does the pattern survive another sample, another analyst, another period or another population?

Statistics is often presented as the last stage, a calculation applied after the

important work has been done. The important work runs through the entire chain. The target defines the job. Definitions create the data. Design creates the comparison. Analysis creates the compression. Validation tests whether it travels.

Once you can see that chain, a precise-looking number loses its power to end the argument by appearance alone. It may still be excellent evidence. You now know what it had to survive to earn that status. You also know where to look when two competent analyses disagree: they may differ in target, data, assumptions or decision rule rather than arithmetic.

Terms

Population

The complete set of people, objects, events or possible observations about which a question is asked. A study population may differ from the broader population to which someone hopes to generalise, and the distinction should be explicit.

Sample

The observations used to learn about a population or process. A sample gains inferential force from its selection mechanism and measurement quality, not from size alone.

Sampling frame

The practical list or system from which units can be selected. Coverage error arises when the frame omits, duplicates or misclassifies parts of the target population.

Parameter

A fixed but usually unknown feature of a population or model, such as a population mean, risk difference or regression coefficient. An estimate uses observed data to learn about it.

Estimand

The precisely defined quantity or contrast an analysis aims to learn, including its population, outcome, treatment or exposure, time and summary. Changing any of those changes the question.

Variable

A recorded characteristic that can differ across observations or occasions. Its meaning depends on an operational definition, scale, unit and collection procedure, all of which should remain stable enough for comparison.

Mean

The arithmetic total divided by the number of values. It uses all observations and has useful mathematical properties, but can be pulled strongly by extreme values and skewed distributions.

Median

The middle value after observations are ordered, with the two middle values combined when required. It is resistant to extremes but does not describe spread or tail behaviour.

Missing data

Values that should exist for the intended analysis but are absent. Their implications depend on why they are missing, what is observed and which assumptions connect the observed cases to the unseen ones.

Standard deviation

A measure of how far observations tend to lie from their mean, expressed in the original unit. It describes variation among values, not uncertainty in an estimated mean.

Variance

The average squared departure from the mean under a specified population or sample formula. Squaring prevents cancellation and gives large deviations extra weight; standard deviation is its square root.

Outlier

An observation distant from the rest under a chosen criterion. It may be error, rare reality or model failure, so its provenance should be investigated before removal or retention.

Bias

Systematic departure of a measurement or estimator from its target. Increasing sample size can reduce random variation while leaving bias unchanged or making the biased result more precise.

Standard error

The standard deviation of an estimator across hypothetical repeated samples under a design and model. It measures sampling uncertainty in the estimate, not dispersion among individual observations.

Confidence interval

An interval generated by a procedure designed to cover a fixed parameter at a stated long-run rate under its assumptions. Different procedures can produce different coverage and width.

Credible interval

A Bayesian interval containing a stated posterior probability for a parameter after combining prior assumptions with a likelihood. Its interpretation depends on the complete model, including the prior.

Likelihood

A function showing how compatible different parameter values are with the observed data under a statistical model. It treats the data as fixed and varies the parameter.

Null hypothesis

A specified model or claim tested against data, often no difference or no association. Failure to reject it does not establish that it is true or that effects are absent.

P-value

Under a specified model, the probability of a test statistic at least as incompatible with the null as the observed one. It is not the probability that the null is true, nor a measure of effect size or importance.

Effect size

A numerical measure of the magnitude of a difference, association or intervention effect. Its practical meaning depends on units, baseline risk, context and uncertainty.

Power

The probability that a test rejects a specified null when a specified alternative is true. Power depends on sample size, variability, design, threshold and the effect chosen for planning.

Type I error

Rejecting a null hypothesis when it is true within the testing framework. A significance level controls this long-run error rate for a declared procedure, not every published claim.

Type II error

Failing to reject a null when a specified alternative is true. Its probability is β , while $1 - \beta$ is power against that alternative.

Confounding

Distortion of an exposure-outcome comparison by other causes related to both. Adjustment can help only for variables measured adequately and chosen through defensible causal reasoning.

Randomisation

Assignment or selection by a known chance mechanism. In experiments it supports causal comparison by preventing systematic treatment allocation; in sampling it permits design-based estimates of uncertainty.

Regression

A family of methods relating an outcome to one or more predictors. Regression can describe, predict or support causal analysis, but the coefficient's meaning depends on design and assumptions.

Calibration

Agreement between predicted probabilities and observed frequencies among comparable cases. Of events given a 20 per cent forecast, roughly one in five should occur in a well-calibrated system over an appropriate set of cases.

Overfitting

Learning peculiarities of the training data that do not persist in new observations. It often appears as excellent in-sample fit paired with weaker external or held-out performance.

False discovery rate

Under a defined procedure and assumptions, the expected share of rejected hypotheses that are false rejections. It concerns a collection of decisions, not the truth probability of one result.

Replication

A new study addressing the same question with new data. A successful replication requires a declared target and appropriate design; it need not reproduce an identical estimate or threshold label because sampling variation and context remain.

Go Deeper

David Spiegelhalter, The Art of Statistics: Learning from Data

Begin here for a broad, humane introduction organised around real questions rather than a procession of formulas. Spiegelhalter moves from data and visualisation through uncertainty, testing, causation and communication, repeatedly asking what can be learned and what remains unsupported. The British Pelican edition appeared in 2019. It is accessible to a newcomer but does not avoid difficult interpretation. Read it with a pencil: many of its best lessons concern the wording of a claim rather than the calculation beneath it. It is especially good on communicating absolute risk, separating an estimate from its uncertainty and resisting false certainty without lapsing into helplessness.

David Freedman, Robert Pisani and Roger Purves, Statistics

Use the fourth edition, published by W. W. Norton in 2007, when you want the full introductory course without the usual fog. The authors are unusually severe about study design, confounding, regression and what statistical tests do not establish. Explanations come before notation, then exercises force the distinction between a calculation and a justified conclusion. It is a substantial textbook rather than a quick sequel, but an intelligent reader can work through selected chapters independently. Its age matters less than its discipline of thought. The examples also train a habit this book has emphasised: identify whether a conclusion came from randomised assignment, probability sampling or an observational comparison before touching the arithmetic.

Stephen M. Stigler, The Seven Pillars of Statistical Wisdom

Read this for the intellectual architecture and history. Stigler's 2016 Harvard University Press book identifies seven ideas that give statistics its unity, including aggregation, likelihood, comparison, regression, experimental design and residuals. It is neither a standard history nor a methods manual. It shows why moves that now look routine were once conceptually strange, such as learning more by discarding individual detail. Some passages assume comfort with mathematical examples, but the governing arguments are clear and rewarding. It is the best next read for seeing statistics as a coherent set of inventions rather than a toolbox whose contents happen to share a course code.

Miguel A. Hernán and James M. Robins, **Causal Inference: What If**

Go here when “correlation is not causation” has stopped being useful and you want to know what comes next. Hernán and Robins build causal questions through interventions, counterfactuals, randomised experiments, observational designs and increasingly demanding longitudinal problems. The authors' official site supplies a free living edition and requests the 2020 Chapman and Hall/CRC citation. The first part is approachable with basic statistics; later parts become technical. Its discipline is to state the causal target before choosing the model. Expect diagrams, notation and careful distinctions. The reward is a working account of confounding and intervention that goes far beyond repeating that association alone cannot settle causation.

Notes and Sources

The Whole Thing in One Page and Why You Should Care

The evidence chain. The organising model draws on David Spiegelhalter's question-led account of learning from data; David Freedman, Robert Pisani and Roger Purves on design, sampling and inference; and the American Statistical Association's 2022 ethical guidance, which treats statistical practice as including collection, processing, analysis, interpretation, presentation and model development. The chain is an organising synthesis rather than a claim that all statistical traditions use one workflow.

Sally Clark. Clark was convicted in 1999 of murdering her two infant sons. At trial, paediatrician Roy Meadow presented an estimate of about 1 in 73 million for two sudden infant deaths in a family with selected characteristics. The figure was obtained by multiplying an estimated risk for one death by itself. The Royal Statistical Society criticised both the independence assumption and the confusion between the probability of evidence under a hypothesis and the probability of innocence given the evidence. The Court

of Appeal judgment in *R v Clark* [2003] EWCA Crim 1020 records the use and possible effect of the figure. Clark's convictions were overturned after previously undisclosed microbiological results concerning her second child came to light. The manuscript therefore does not claim that the statistical errors alone caused the conviction or reversal. Ray Hill's 2004 paper examines dependence and competing explanations in greater technical depth.

The Literary Digest poll. The magazine mailed roughly ten million ballots and received about 2.4 million returns before predicting a Landon victory in the 1936 United States presidential election. Roosevelt won 60.8 per cent of the popular vote. Peverill Squire's 1988 analysis concludes that biased coverage and biased response jointly produced the error. Dominic Lusinchi's later reanalysis gives particular weight to differential non-response. The manuscript retains both mechanisms and avoids the simplified claim that telephone and car ownership alone explain the failure. Gallup's smaller quota poll picked the winner but understated Roosevelt's margin and was not a modern probability sample.

Sources for the seven core ideas

Targets, estimands and operational definitions. The estimand terminology follows modern statistical and causal-inference practice: the target quantity must specify the population, outcome, treatment or exposure, time and summary needed by the question. The distinction between a phenomenon and its operational measure follows standard work in survey methodology, epidemiology and official statistics. Robert Groves and colleagues explain coverage, measurement, non-response and processing error as components of total survey error. The examples of unemployment, recorded crime and readmission are generic and are not tied to one jurisdiction's current definition.

Measurement reacting on behaviour. The observation that published or rewarded measures can change conduct is a general incentive point, often associated with Goodhart's law and Campbell's law. The manuscript does not present either slogan as a universal mathematical law. It uses the

narrower claim that people and institutions may respond to measures once consequences are attached to them.

Sampling, assignment and the square-root rule. Random sampling supports population inference through known inclusion probabilities; random assignment supports causal comparison through a known allocation mechanism. Either can occur without the other. Under independent sampling with finite variance, the standard error of a sample mean scales with one divided by the square root of sample size. Related estimators have analogous large-sample behaviour under their own conditions. Four times the sample size therefore often halves the sampling component of uncertainty. Dependence, unequal weights, clustering and model structure can change the effective information. Increasing sample size does not remove systematic bias.

Centres, spread and distributions. Definitions of the mean, median, variance, standard deviation, quantiles and interquartile range follow standard statistical usage. The manuscript avoids presenting one summary as universally superior. Stephen Stigler's discussion of aggregation and residuals supports the broader argument that statistical learning often gains power by discarding individual detail while then studying what remains.

Graphs and the central limit theorem. The graphical examples follow the standard diagnostic uses of histograms, scatterplots, time plots and residual plots described by Spiegelhalter and Freedman, Pisani and Purves. The central limit theorem is stated narrowly: under broad but limited conditions, standardised sums or averages approach a normal distribution. It does not make the source population normal. Dependence, infinite variance and slow convergence in heavy-tailed settings can invalidate or weaken a routine normal approximation.

Levels and dependence. Repeated measurements within a person, pupils within schools and shops within chains are examples of clustered data. Treating correlated observations as independent can understate uncertainty. The exact remedy depends on design and model and may include cluster-robust uncertainty estimates, multilevel models or analysis at

the correct unit.

Standard error and confidence intervals. Jerzy Neyman's 1937 paper supplies the repeated-sampling theory of confidence sets used here. Sander Greenland and colleagues document widespread interval and p-value misinterpretations. The manuscript states the conventional frequentist interpretation: a 95 per cent procedure has 95 per cent coverage under its model and repeated use. It does not assign a 95 per cent posterior probability to one fixed interval. Other interval procedures, including Bayesian and bootstrap intervals, have different constructions and interpretations.

Bayesian reasoning. The brief contrast is limited to what a newcomer needs. A posterior distribution combines a prior distribution with a likelihood under a model. It can support probability statements about parameters within that model. Choice of prior, likelihood, model checking and sensitivity remain necessary. The manuscript does not imply that Bayesian methods remove subjectivity or that frequentist methods contain none.

P-values. The definition and cautions follow the official 2016 ASA statement, rechecked on 3 September 2026. Its six principles say that p-values can indicate incompatibility between data and a specified model; do not measure the probability that a hypothesis is true or that data arose from random chance alone; should not alone determine scientific, policy or business decisions; require full reporting and transparency; do not measure effect size or importance; and do not by themselves provide a good measure of evidence regarding a model or hypothesis. The body paraphrases rather than reproduces the statement.

The 2019 and 2021 ASA documents. “Moving to a World Beyond $p < 0.05$ ” was a 2019 editorial by Ronald Wasserstein, Allen Schirm and Nicole Lazar. The official ASA pages and task-force statement were rechecked on 3 September 2026. The article states that it reflects the authors' views and is not an endorsed ASA position. The 2021 presidential task-force statement was issued partly to prevent that editorial being mistaken for policy. It says properly applied and interpreted p-values and significance tests remain

important tools, while stressing uncertainty, variability, multiplicity and replicability. The manuscript presents reform as a live methodological debate with substantial agreement on common misuse.

Power and low-powered studies. Power is defined against a specified alternative, design and testing rule. The claim that low-powered published results can have unstable or exaggerated effect estimates follows the selection created when only estimates that cross a threshold receive attention. No universal exaggeration factor is asserted.

The multiple-testing calculation. If twenty tests are independent, all twenty null hypotheses are true and each rejects with probability 0.05, the chance of no rejection is 0.95 raised to the twentieth power. The complementary probability of at least one false rejection is about 0.6415. Dependence and false nulls change the number. The example is used to show why the opportunity set matters, not as a model of every research programme.

False discovery rate. Yoav Benjamini and Yosef Hochberg's 1995 procedure controls the expected proportion of false rejections among rejections under stated assumptions. The manuscript distinguishes this from family-wise error control and from a posterior probability that any particular finding is false.

Berkeley admissions. P. J. Bickel, E. A. Hammel and J. W. O'Connell analysed 1973 graduate admissions data at the University of California, Berkeley. The aggregate association between sex and admission differed from many department-level associations because application patterns varied across departments with different acceptance rates. The case is used to explain aggregation and conditioning. The manuscript does not claim that department-level adjustment proves the absence of discrimination, because department choice and other stages can themselves be part of the causal process.

Randomisation and causation. Fisher's work on experimental design supports the use of random allocation as the basis for a known comparison.

Hernán and Robins supply the modern counterfactual framing. Randomisation balances pretreatment causes in expectation, not necessarily exactly in one realised experiment. Causal interpretation can still be threatened by non-compliance, attrition, interference, outcome switching, implementation failure and limited external validity.

Bradford Hill. Hill's 1965 address discussed considerations relevant to moving from association to causation, including strength, consistency, specificity, temporality, biological gradient, plausibility, coherence, experiment and analogy. Hill warned that none supplied an automatic rule. The manuscript therefore calls them considerations rather than criteria that prove causation.

Prediction and calibration. Predictive models are judged on observations not used to fit them. Discrimination concerns ranking or separation; calibration concerns agreement between predicted probabilities and observed frequencies. The 20 per cent forecast example is generic. A model may be calibrated overall and miscalibrated in subgroups or changing settings.

Overfitting and data reuse. Frank Harrell's work on regression modelling and standard statistical-learning practice support the distinction between training fit and new-data performance. Cross-validation can estimate performance efficiently, but repeated tuning against the same validation or test data leaks information and biases assessment.

Reproducibility and replicability. The National Academies' 2019 report uses reproducibility for obtaining consistent computational results with the same data, code, methods and conditions of analysis, and replicability for obtaining consistent results across studies aimed at the same scientific question using new data. Other fields reverse or vary these terms, so the body identifies the convention rather than claiming universal usage.

Open Science Collaboration. The 2015 project attempted replications of 100 experimental and correlational studies from three psychology journals. Ninety-seven per cent of the original studies reported statistically significant

results, compared with 36 per cent of replications. Replication effect estimates were on average around half the size of the original effects. The project used several indicators of replication and prompted published methodological criticism and replies. The manuscript treats it as a bounded empirical project, not a universal rate for psychology or science.

Transparency reforms. Preregistration, registered reports, open data and code, and specification-curve or multiverse analyses address different parts of the evidence chain. They can expose choices and reduce some selection, but poor questions, weak measurement, non-adherence and inappropriate models remain possible.

Sources for the operating history

John Graunt. Population counts, censuses, tax records and quantitative administration long predate Graunt. His 1662 *Natural and Political Observations Made upon the Bills of Mortality* is used for the narrower move of comparing recurring mortality records to infer population regularities, estimate population features and construct an early survival table. Causes of death were recorded through a fallible parish system, so the manuscript presents the bills as productive but imperfect administrative data. Priority claims are kept narrow.

Edmond Halley. Halley's 1693 paper used births and funerals from Breslau to construct a life table and calculate annuity prices by age. Later scholarship debates how much he smoothed and adjusted the underlying records. The manuscript uses the secure claim that population mortality was connected to financial pricing, not that his table was the uncontested first of every kind.

State statistics. Theodore Porter's history supports the link between nineteenth-century statistical thinking, administration and trust in quantified rules. The body does not reduce statistics to state control; it records that population measurement served taxation, public health, insurance, military planning and governance.

Quetelet and Galton. Adolphe Quetelet applied averages and the normal

curve to social and bodily measurements. Francis Galton's work on hereditary stature helped establish regression and correlation. Stephen Stigler's histories support the technical chronology. Galton founded and promoted eugenics, so the manuscript keeps the target and social programme visible without treating later statistical methods as morally determined by their origin.

Regression towards the mean. Selection on an extreme observed value tends to be followed by a less extreme value when measurements contain transient variation. The school, patient and fund examples are illustrative, not reports of specific studies. A comparison group and repeated measurements can help distinguish regression from intervention effects.

Gosset. William Sealy Gosset worked at Guinness and published “The Probable Error of a Mean” under the name “Student” in 1908. The paper developed exact small-sample results associated with the t distribution. Stephen Zabel's centenary review supports the industrial and theoretical context. The manuscript does not infer undocumented corporate motives beyond the accepted account that company publication rules contributed to the pseudonym.

Fisher and Rothamsted. Fisher joined Rothamsted in 1919 and developed methods for analysing agricultural experiments during the 1920s. His books of 1925 and 1935 systematised significance testing and experimental design. Randomisation, replication, blocking and factorial structure had precursors and collaborators, so the body presents Fisher as the central synthesiser and developer rather than sole inventor of every practice.

The tea experiment. *The Design of Experiments* presents an eight-cup test in which four cups are prepared each way. The probability structure follows the random assignment and fixed counts. Later retellings identify Muriel Bristol and reconstruct a tea-room event, but the body relies only on the published illustration and does not invent dialogue or claim that the printed version is a transcript.

Fisher and Neyman-Pearson. Fisher's significance-testing approach and

the Neyman-Pearson decision framework differ in aims and interpretation. Jerzy Neyman and Egon Pearson's 1933 paper formalised tests with controlled errors and power against alternatives. Neyman's 1937 paper developed confidence-set theory. The account of a later hybrid ritual follows historical and methodological literature but is presented as a common practice, not a universal description of every field.

Survey sampling. Probability sampling permits design-based inference through known inclusion probabilities. Stratification, clustering and weights change both estimation and uncertainty. The account relies on Groves and colleagues and standard survey-sampling theory. The Gallup comparison is bounded because 1936 quota sampling does not meet modern probability-sampling standards.

Tukey, missing data and resampling. John Tukey's 1962 paper argued for data analysis as a broad empirical discipline rather than a narrow branch of mathematical inference. Donald Rubin's 1976 paper formalised influential distinctions among missing-data mechanisms. Bradley Efron's 1979 bootstrap paper introduced a general resampling approach to uncertainty estimation. The body compresses later developments and does not suggest that one paper completed each subject.

Multiplicity and flexible analysis. Simmons, Nelson and Simonsohn's 2011 paper showed how undisclosed flexibility in sample size, outcomes, covariates and reporting can inflate false-positive findings. The manuscript uses their general mechanism and avoids claiming that all researchers engage in every practice.

What People Get Wrong and Use It

The seven corrections synthesise the evidence above. The practical lenses are decision aids rather than diagnostic guarantees. A denominator check cannot reveal every bias; a flow diagram cannot prove that missing cases are harmless; an interval cannot capture every uncertainty; validation in one dataset cannot establish transportability; and replication can disagree for substantive as well as statistical reasons. The aim is to make the evidential burden visible and proportionate.

Bibliography

Primary and original works

Benjamini, Yoav, and Yosef Hochberg. "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing." *Journal of the Royal Statistical Society, Series B* 57, no. 1 (1995): 289-300.

Bickel, P. J., E. A. Hammel, and J. W. O'Connell. "Sex Bias in Graduate Admissions: Data from Berkeley." *Science* 187, no. 4175 (1975): 398-404.

Efron, Bradley. "Bootstrap Methods: Another Look at the Jackknife." *Annals of Statistics* 7, no. 1 (1979): 1-26.

Fisher, R. A. *The Design of Experiments*. Edinburgh: Oliver and Boyd, 1935.

Fisher, R. A. *Statistical Methods for Research Workers*. Edinburgh: Oliver and Boyd, 1925.

Graunt, John. *Natural and Political Observations Mentioned in a Following Index, and Made upon the Bills of Mortality*. London: John Martyn, 1662.

Halley, Edmond. "An Estimate of the Degrees of the Mortality of Mankind, Drawn from Curious Tables of the Births and Funerals at the City of Breslaw." *Philosophical Transactions* 17 (1693): 596-610.

Hill, Austin Bradford. "The Environment and Disease: Association or Causation?" *Proceedings of the Royal Society of Medicine* 58 (1965): 295-300.

Neyman, Jerzy. "Outline of a Theory of Statistical Estimation Based on the Classical Theory of Probability." *Philosophical Transactions of the Royal Society A* 236 (1937): 333-380.

Neyman, Jerzy, and Egon S. Pearson. "On the Problem of the Most Efficient Tests of Statistical Hypotheses." *Philosophical Transactions of the Royal Society A* 231 (1933): 289-337.

Open Science Collaboration. "Estimating the Reproducibility of Psychological Science." *Science* 349, no. 6251 (2015): aac4716.

Rubin, Donald B. "Inference and Missing Data." *Biometrika* 63, no. 3 (1976): 581-592.

Simmons, Joseph P., Leif D. Nelson, and Uri Simonsohn. "False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant." *Psychological Science* 22, no. 11 (2011): 1359-1366.

"Student" [William Sealy Gosset]. "The Probable Error of a Mean." *Biometrika* 6, no. 1 (1908): 1-25.

Tukey, John W. "The Future of Data Analysis." *Annals of Mathematical Statistics* 33, no. 1 (1962): 1-67.

Modern works and guidance

American Statistical Association. *Ethical Guidelines for Statistical Practice*. 2022 revision. Accessed 3 September 2026.

Benjamini, Yoav, Richard De Veaux, Bradley Efron, Scott Evans, Mark Glickman, Barry Graubard, Xuming He, et al. "The ASA President's Task Force Statement on Statistical Significance and Replicability." *The Annals of Applied Statistics* 15, no. 3 (2021): 1084-1085.

Box, George E. P., J. Stuart Hunter, and William G. Hunter. *Statistics for Experimenters: Design, Innovation, and Discovery*. 2nd ed. Hoboken: Wiley, 2005.

Freedman, David A. "From Association to Causation via Regression." *Advances in Applied Mathematics* 18 (1997): 59-110.

Freedman, David A. "Statistical Models and Shoe Leather." *Sociological Methodology* 21 (1991): 291-313.

Freedman, David, Robert Pisani, and Roger Purves. *Statistics*. 4th ed. New York: W. W. Norton, 2007.

Gelman, Andrew, Jennifer Hill, and Aki Vehtari. *Regression and Other Stories*. Cambridge: Cambridge University Press, 2020.

Greenland, Sander, Stephen J. Senn, Kenneth J. Rothman, John B. Carlin, Charles Poole, Steven N. Goodman, and Douglas G. Altman. "Statistical Tests, P Values, Confidence Intervals, and Power: A Guide to Misinterpretations." *European Journal of Epidemiology* 31 (2016): 337-350.

Groves, Robert M., Floyd J. Fowler Jr, Mick P. Couper, James M. Lepkowski, Eleanor Singer, and Roger Tourangeau. *Survey Methodology*. 2nd ed. Hoboken: Wiley, 2009.

Harrell, Frank E. Jr. *Regression Modeling Strategies*. 2nd ed. Cham: Springer, 2015.

Hernán, Miguel A., and James M. Robins. *Causal Inference: What If*. Boca Raton: Chapman and Hall/CRC, 2020, living online edition.

Hill, Ray. "Multiple Sudden Infant Deaths: Coincidence or Beyond Coincidence?" *Paediatric and Perinatal Epidemiology* 18, no. 5 (2004): 320-326.

Lusinchi, Dominic. "'President' Landon and the 1936 'Literary Digest' Poll." *Social Science History* 36, no. 1 (2012): 23-54.

National Academies of Sciences, Engineering, and Medicine. *Reproducibility and Replicability in Science*. Washington, DC: National Academies Press, 2019.

Porter, Theodore M. *The Rise of Statistical Thinking, 1820-1900*. Princeton:

Princeton University Press, 1986.

Royal Statistical Society. "Letter from the President to the Lord Chancellor Regarding the Use of Statistical Evidence in Court Cases." 23 January 2002.

Spiegelhalter, David. *The Art of Statistics: Learning from Data*. London: Pelican, 2019.

Squire, Peverill. "Why the 1936 Literary Digest Poll Failed." *Public Opinion Quarterly* 52, no. 1 (1988): 125-133.

Stigler, Stephen M. *The History of Statistics: The Measurement of Uncertainty before 1900*. Cambridge, MA: Harvard University Press, 1986.

Stigler, Stephen M. *The Seven Pillars of Statistical Wisdom*. Cambridge, MA: Harvard University Press, 2016.

Wasserstein, Ronald L., and Nicole A. Lazar. "The ASA Statement on p-Values: Context, Process, and Purpose." *The American Statistician* 70, no. 2 (2016): 129-133.

Wasserstein, Ronald L., Allen L. Schirm, and Nicole A. Lazar. "Moving to a World Beyond $p < 0.05$." *The American Statistician* 73, sup. 1 (2019): 1-19.

Zabell, S. L. "On Student's 1908 Article 'The Probable Error of a Mean'." *Journal of the American Statistical Association* 103, no. 481 (2008): 1-7.

Legal source

R v Clark [2003] EWCA Crim 1020.