Check for updates

The Royal Society of Medicine

Journal of the Royal Society of Medicine; 2026, Vol. 119(8) 237–247
DOI: 10.1177/01410768261479691

The historical development of sample size estimation: from Huygens and Bernoulli to the present

Kristian Thorlund[1,2] and Christian Gluud[2,3]
[1]Department of Health Research Methods, Evidence and Impact, Faculty of Health Sciences, McMaster University, Hamilton, ON, Canada
[2]Copenhagen Trial Unit, Centre for Clinical Intervention Research, The Capital Region, Copenhagen University Hospital – Rigshospitalet, Copenhagen, Denmark
[3]Department of Regional Health Research, The Faculty of Health Sciences, University of Southern Denmark, Odense, Denmark
Corresponding author: Kristian Thorlund. Email: thorluk@mcmaster.ca

Introduction

In the realm of empirical research, determining the appropriate sample size is a critical component of study design. Whether estimating a population proportion, testing a new treatment or validating a machine learning model, researchers must decide how much data to collect to ensure results are reliable, reproducible and scientifically meaningful.

Sample size estimation is not just a statistical technicality; it reflects broader epistemological commitments. Collect too little data, and the study's findings lack statistical power, risking false negative or false positive inferences. Collecting too much may waste resources or expose more participants than necessary to experimental or control risks. Finding the sample size that is 'just right' is both an art and a science.

Historically, the development of sample size estimation has mirrored the evolution of probability and statistics, from early philosophical inquiries into uncertainty to increasingly formalised methods for planning empirical research. This paper traces these developments through key periods and thinkers, beginning with Christiaan Huygens quantitative framing of uncertainty and Jacob Bernoulli's pioneering work on the 'law of large numbers'. It then follows the maturation of frequentist sample size calculation through hypothesis testing and experimental design, alongside later Bayesian approaches to uncertainty and planning, while briefly noting areas in which methodological development remains ongoing (Figure 1).

The origins of probabilistic thinking

Christiaan Huygens was among the first to frame uncertainty in quantitative terms. In *De Ratiociniis in Ludo Aleae*,[1] Huygens sought to understand how fair outcomes could be mathematically reasoned in games of chance. Building on earlier correspondence between Blaise Pascal and Pierre de Fermat in 1654, Huygens introduced the concept of expected value as a rational basis for decision-making under uncertainty. Although his work did not address sample size directly, Huygens' ideas established a new way of thinking about randomness, thereby shifting probabilistic inquiry from philosophy to computation, and laying the groundwork for future efforts to relate empirical data to underlying truths.

Jacob Bernoulli and the 'law of large numbers'

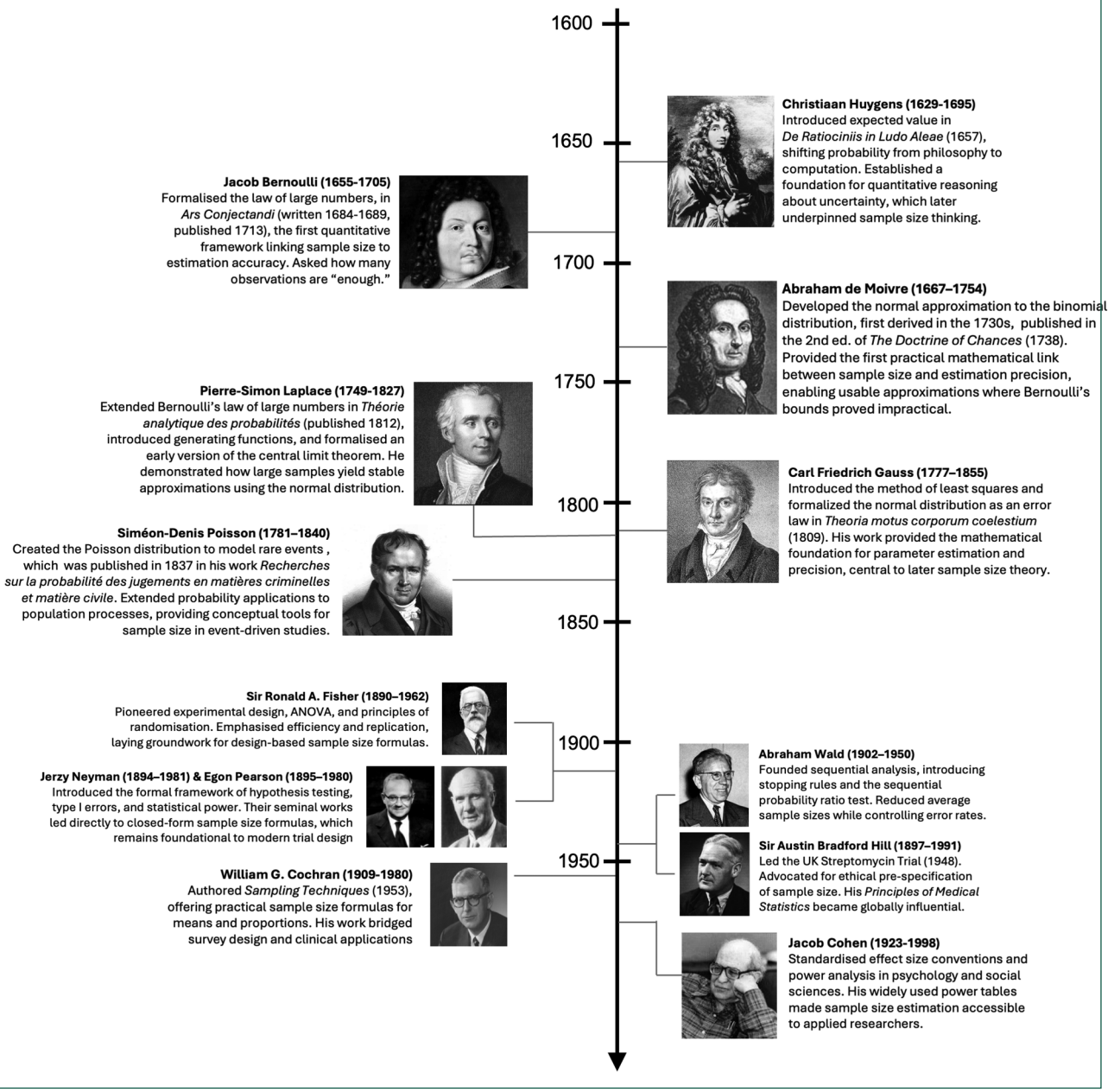
The roots of sample size estimation can be traced to Jacob Bernoulli, a Swiss mathematician from the famous Bernoulli family. In his posthumously published masterpiece, *Ars Conjectandi*,[2] Bernoulli articulated one of the most important results in probability theory: the law of large numbers, laying the mathematical foundations for the view of probability as a long-run relative frequency over an infinite series of repeated trials.

Bernoulli was interested in how many observations were needed to estimate the true probability of a success in repeated, independent experiments. For example, how many times a coin must be flipped to ensure that the observed proportion of heads approximates the true probability. He formalised this question using 'Bernoulli trials' – experiments with binary outcomes (success or failure) – and showed that the observed proportion converges to the true probability as the number of trials increases.

Bernoulli went further by attempting to derive a numerical bound on the number of trials required to estimate an unknown probability with a specified degree of accuracy

© The Author(s) 2026
Article reuse guidelines: sagepub.com/journals-permissions

Figure 1. Historical development of sample size estimation.



and confidence. In *Ars Conjectandi*,[2] he posed what later became known as 'Bernoulli's problem', applying his theory to the question at the core of sample size determination: how many observations are needed to ensure that the observed proportion of successes lies within a given margin of the true probability with high probability. Bernoulli illustrated this using an analogy of drawing coloured stones from a jar, where repeated random draws progressively reveal the underlying proportion, much like more patients entering a clinical trial allow estimation of the true treatment response proportion with increasing precision.

However, the numerical bounds Bernoulli derived for this purpose were both extremely conservative and methodologically crude. As Karl Pearson later demonstrated in a detailed historical and mathematical reassessment, Bernoulli relied on a loose system of inequalities rather than analytic approximations to the binomial distribution, resulting in sample size requirements that were exaggerated by several fold relative to later solutions.[3] In the case of Bernoulli's original jar example, to estimate a probability of 0.6 within ± 0.02 with odds of 1000:1, Bernoulli's method yielded a required sample size of more than

25,000 observations, a result so impractical that it seems to have deterred Bernoulli from publishing *Ars Conjectandi*; it appeared 8 years after his death.

Importantly, the inverse square root relationship between estimation accuracy and sample size, frequently attributed to Bernoulli in later textbooks, does not appear in *Ars Conjectandi*. As Karl Pearson showed, this relationship emerges only with Abraham de Moivre's normal approximation to the binomial distribution, developed in the 1730s and later formalised through the central limit theorem.[3,4] The subsequent attribution of this result to Bernoulli reflects a historical conflation rather than a faithful reading of the original work.

Bernoulli's lasting contribution therefore lies not in providing a practical solution to sample size determination, but in clearly formulating the problem itself: the recognition that a quantitative relationship exists between sample size, uncertainty and the reliability of empirical inference. This conceptual insight that large samples may be required to achieve high degrees of certainty remains a foundational lesson for statistical reasoning, even though the mathematical tools needed to operationalise it were supplied by later developments.

The 18th–19th century: probability theory matures, but sample size lags

After Bernoulli, the 18th and 19th centuries saw a flowering of probability theory, with major contributions from Pierre-Simon de Laplace, Carl Friedrich Gauss and Siméon-Denis Poisson. Laplace developed early versions of the 'central limit theorem' and used normal approximations to estimate probabilities.[5] Gauss introduced the method of least squares and assumed normally distributed errors in measurements.[6] Poisson developed his namesake distribution to model rare events in large populations.[7] Furthermore, several prominent researchers, including Jules Gavarret and Thomas Graham Balfour saw the dangers later called type I and type II errors in small numbers.[8]

Despite these advances, sample size estimation remained at best implicit and even then, approximate and based on pragmatic considerations such as availability of patients or therapy. Most applications involved censuses or observational studies with fixed sample sizes. There was little incentive or infrastructure to ask how many observations were enough.

The Neyman–Pearson revolution and the formalisation of hypothesis testing (1920s–1930s)

A turning point came with the work of Jerzy Neyman and Egon Pearson in the 1920s and 1930s. Their framework introduced formal concepts of hypothesis testing, type I and type II errors and statistical power.[9] These innovations enabled researchers to calculate the minimum sample size needed to achieve a desired probability of detecting a true effect, given an acceptable false-positive rate.

Their work culminated in the Neyman–Pearson 'lemma',[9] which provided a general theoretical foundation for constructing the most powerful tests for simple hypotheses. Although the Neyman–Pearson framework itself is distribution-agnostic, its application to commonly used parametric settings led, in subsequent methodological work, to closed-form sample size equations for standard scenarios. For example, in comparing two population means under a normality assumption with unknown variance, the required sample size per group is often approximated as:

$$n = \left(\frac{Z_{1-\alpha/2} + Z_{1-\beta}}{\delta / \sigma} \right)^2$$

where δ is the minimum detectable difference between groups, σ is the assumed standard deviation, and α is the type I error rate (probability of a false positive) and β is the type II error rate (probability of a false negative). Although this formula is derived using normal-theory approximations, it is routinely used to plan studies analysed using two-sample *t*-tests, particularly when σ must be estimated from the data.

Similarly, for tests of proportions, widely used sample size formulas are based on large-sample normal approximations to the binomial distribution,[10,11] for example:

$$n = \frac{Z_{1-\alpha/2}\sqrt{2p(1-p)} + Z_{1-\beta}\sqrt{p_1(1-p_1) + p_2(1-p_2)}^2}{(p_1 - p_2)^2}$$

where p_1 and p_2 are the proportions in the two groups, and p is a pooled or average proportion (often taken as $(p_1 + p_2)/2$) used in the variance approximation under the null hypothesis. These formulas allowed researchers to pre-specify sample sizes based on expected effect sizes, substantially advancing the prospective planning of clinical trials, survey experiments and industrial quality testing. Notably, these formulations make explicit the problem that Bernoulli had only glimpsed: that achieving higher levels of precision or confidence requires disproportionately larger sample sizes.

From a historical perspective, both formulas implicitly rely on the law of large numbers and the central limit theorem (unless normality is already assumed). The law of large numbers ensures that, as the sample size increases, the sample mean or proportion converges to its true population value, making the estimates more stable. The central limit theorem is used to justify approximating the sampling distribution of these statistics by a normal distribution, even when the underlying data are not perfectly normal. This normal approximation is what allows the use

of *z*-scores in the formulas and makes closed-form sample size calculations feasible (direct calculation of required sample sizes using specific algebraic formulas).

The broader conceptual shift introduced by Neyman and Pearson was therefore not only technical but also methodological. Statistical inference was no longer treated as a purely post hoc exercise applied after data collection, but a prospective design choice requiring explicit specification of acceptable error rates, effect sizes and study objectives. This reorientation made it possible to determine sample sizes in advance in a principled and reproducible way, aligning study design with scientific aims, ethical considerations and practical constraints. Neyman–Pearson-based sample size calculations thus formalised and approach to study planning focused on controlling inferential error rates when testing clearly specified hypotheses.

Fisher's influence and the rise of experimental design

In parallel with Neyman and Pearson, Sir Ronald Aylmer Fisher revolutionised empirical science by formalising the principles of experimental design in agriculture and biology. In his seminal work, *Design of Experiments*,[12] Fisher emphasised techniques like randomisation, blocking and factorial designs to reduce bias and improve precision. His introduction of the analysis of variance (ANOVA) enabled researchers to partition variability and test multiple hypotheses simultaneously.

Although Fisher did not advocate for power analysis in the way that Neyman and Pearson did, his methods implicitly addressed data sufficiency. For example, in ANOVA, the *F*-statistic compares the variance between groups to the variance within groups:

$$F = \frac{\mathrm{MS}_{\text{between}}}{\mathrm{MS}_{\text{within}}}$$

where MS denotes the mean square. The non-centrality parameter of the *F*-distribution, which is a function of sample size and effect size, determines the power of the test. As later statisticians such as Oscar Kempthorne as well as William Gemmell Cochran and David Roxbee Cox expanded Fisher's designs, they derived explicit sample size formulas to detect main effects or interactions in balanced factorial experiments.[11,13]

For instance, in a one-way ANOVA with k groups and equal group sizes n, the sample size needed to detect a standardised effect f with significance level $\alpha/2$ and power $1-\beta$ for a two-sided test is approximately:

$$n = \frac{(k-1)(Z_{1-\alpha/2} + Z_{1-\beta})^2}{kf^2}$$

Fisher's emphasis on efficiency, replication and systematic control of variability laid the conceptual and mathematical groundwork for sample size optimisation. While Fisher favoured inductive inference and resisted Neyman–Pearson's binary decision framework, his contributions were foundational to later developments in power analysis, resource allocation and design-based sample size planning.

Early references to sample size estimation for clinical studies

Early efforts to translate statistical principles into practical guidance for clinical research can be found in the work of Donald Mainland and Austin Bradford Hill in the 1930s.[14,15] Mainland had already addressed problems of chance and numerical adequacy in clinical work in a 1936 *BMJ* paper, prepared with statistical advice from Fisher, before expanding these ideas in his 1938 book.[14,16] Although neither author presented explicit formula-based derivations, both provided numerical guidance on the relationship between sample size, detectable effect and statistical uncertainty that is broadly consistent with modern theory.

In his 1938 book, Mainland considered both small studies, using exact probability arguments and larger studies, where normal approximations were invoked 1930s.[14] His calculations, developed with input from Fisher, illustrate an early attempt to quantify the number of observations required to detect differences of practical relevance under varying assumptions. Hill, in the first edition of *Principles of Medical Statistics*, focused on larger studies and provided approximate sample size requirements for detecting effects of different magnitudes.[15] His recommendations correspond closely to conventional significance thresholds and modest power, often effectively targeting a probability of detection near 50% and rounding sample sizes upward for practical use.

Taken together, these contributions represent an early 'broad-brush' approach to sample size planning, where the emphasis was on approximate detectability rather than formal error control or explicit optimisation. They also highlight that, despite the availability of theoretically grounded methods, sample size determination in clinical research remained largely pragmatic, shaped by feasibility and context rather than strict adherence to formal calculation.

The post-World War II era (1): operational research, clinical trials and standardised procedures (1940s–1960s)

By the post-World War II period, statistical thinking had already begun to permeate areas such as public health

and agriculture, particularly following Fisher's work at Rothamsted and was increasingly extending into clinical medicine. The rise of randomised trials marked a pivotal shift towards rigorous causal inference in medicine. Earlier controlled studies, including quasi-randomised or alternation-based designs, had been conducted in the 19th and early 20th centuries, illustrating growing awareness of the need to reduce bias, although true randomisation was not yet consistently implemented.[17,18] This shift is generally regarded as having begun with the 1948 Medical Research Council streptomycin trial whose principal statistician was Austin Bradford Hill.[19] As noted above, Hill had pointed out the need to address sample size issues in his hugely influential textbook '*Principles of Medical Statistics*'.[15] Despite this, however, there is no explicit mention of the calculation of sample size in the 1948 trial, which – like subsequent randomised trials – appears to have been dictated primarily by pragmatic considerations. Even the landmark 1954 Salk poliomyelitis vaccine trial determined its sample size pragmatically on epidemiological and logistical grounds, without reference to formal power calculations or detectable-effect criteria.[20] Consistent with this broader pattern, early quantitative reasoning about sample size in clinical trials most often appeared retrospectively, through post hoc power calculations or illustrative discussions of detectable differences, rather than as prospective, protocol-defining design criteria. Table 1 provides early illustrative examples of randomised trials.

By the 1950s, practical planning tools for clinical and observational research began to become more widely available. Cochran's *Sampling Techniques*[10] provided explicit sample size formulas for estimating population means and proportions with specified precision. For example, the sample size n required to estimate a population mean with margin of error E and standard deviation σ at confidence level $1-\alpha/2$ is:

$$n = \left(\frac{Z_{1-\alpha/2} \cdot \sigma}{E} \right)^2$$

This formula became widely used in public health surveys and clinical studies, where estimation rather than hypothesis testing was the main goal. For binary outcomes, Cochran and others popularised the normal approximation to the binomial distribution for planning prevalence surveys:

$$n = \frac{Z^2_{1-\alpha/2} \cdot p(1-p)}{E^2}$$

where p is the anticipated proportion and E is the desired margin of error. These equations enabled pre-study planning based on anticipated variability, target precision and confidence thresholds.

The post-World War II era (2): ethical rigour, sequential efficiency and the contributions of Hill and Wald (1940s–1960s)

In addition to advocating for the use of randomisation and blinding in clinical trials for the elimination of allocation bias, Hill emphasised the ethical and scientific imperative of pre-specifying trial objectives and statistical parameters, including sample size, to avoid misleading post hoc interpretations. He outlined practical guidance for calculating sample sizes based on detectable treatment differences and clinically meaningful effect sizes, long before these practices were codified in regulatory guidance. Hill's influence extended globally, shaping the methodology of early therapeutic trials in cardiology and infectious diseases.

Meanwhile, in the realm of operational research, Abraham Wald developed the theory of sequential analysis, which provided a radically different approach to determining how much data to collect. Instead of fixing the sample size in advance, Wald's method allowed data to be evaluated as it accumulated, with formal rules for stopping early if results were sufficiently strong. His most well-known contribution, the sequential probability ratio test (SPRT), minimised the average number of observations needed while controlling error rates.[30] For example, in the binary case, Wald derived boundary conditions for likelihood ratios such that sampling could be stopped as soon as the accumulated evidence crossed a critical threshold. Though originally applied in industrial quality control during World War II,[31] SPRT and its extensions were soon adapted to clinical trials, vaccine testing and epidemiological surveillance, particularly in settings where minimising patient burden or cost was paramount.[32] The work further set the stage for the group sequential design methodologies such as O'Brien-Fleming monitoring boundaries[33] and Lan-DeMets[34] alpha-spending that are now used so widespread in large well-designed clinical trials.[35]

Together, Hill and Wald demonstrated two complementary philosophies of sample size planning: one grounded in fixed-sample ethical rigour and trial design, the other in adaptive efficiency and probabilistic control. Both helped to cement the idea that sample size is not merely a logistical detail but a central element of scientific validity. At the same time, these post-war developments also exposed a broader conceptual issue that would later be articulated more explicitly. While both fixed-sample and sequential methods refined control of statistical errors and efficiency, they largely

Table 1. Examples of 'early' randomised clinical trials incorporating formal sample size planning and closed form calculations – either directly citing seminal methods papers or aligning with their frameworks.

Trial	Year	Clinical area	Type of primary outcome	Sample size/power rationale stated	Methodological reference citation or alignment	Notes
Zubrod et al.,[21] *J Chronic Dis*	1960	Oncology (chemotherapy)	Dichotomous (response/no response)	Quantitative discussion of sample size requirements for comparing two proportions ($\alpha = 0.05$, 90% power assurance) with worked numerical examples; no a priori calculation for the reported trial	Aligned with binomial variance framework from Cochran[10]	Transitional example: explicit articulation of type I and II error logic and detectable differences for binary outcomes; sample sizes chosen pragmatically rather than prospectively powered
Chouinard et al.,[22] *Biol Psychiatry*	1985	Psychiatry (acute mania)	Continuous (mania symptom scale)	Formal power calculation for detectable mean difference (two-sided $\alpha = 0.05$, 90% power); post hoc evaluation of power to detect smaller clinically meaningful differences	Direct citation of Cohen[23]	Post hoc power analysis used to interpret null findings; explicit invocation of standardised power framework for continuous outcomes
Werzberger et al.,[24] *N Engl J Med*	1992	Infectious disease (hepatitis A prevention)	Rates/event counts (incident hepatitis A cases over follow-up)	Power based on expected annual incidence and assumed vaccine efficacy; event-count modelling assumes Poisson processes	Direct citation of Brown and Green[25]	Explicit Poisson/ event-count framing and exact conditional inference; strong exemplar of rate-based planning
Earl et al.,[26] *Br J Cancer*	1991	Oncology (SCLC)	Time-to-event (overall survival)	Formal sample size planning based on required number of events for log-rank comparison; event-driven survival design	Direct citation of Freedman[27]	Explicit citation of log-rank-based sample size tables; clear application of asymptotic survival power theory in trial planning
Scandinavian Simvastatin Survival Study (4S),[28] *Lancet*	1994	Cardiology (CHD)	Time-to-event (mortality)	Event-driven design targeting a prespecified number of deaths; group comparison by log-rank test; relative risk via Cox regression	Methods consistent with Schoenfeld[29] and Freedman[27]	Landmark survival trial using formal event-driven planning and log-rank/Cox methods, though without explicit citation of underlying sample size formula papers

CHD: coronary heart disease; SCLC: small cell lung cancer.

retained an explanatory orientation, focusing on hypothesis testing and estimation under controlled conditions. In a seminal paper, Schwartz and Lellouch subsequently distinguished between explanatory trials, designed to test biological hypotheses and pragmatic trials, designed to support treatment decisions under real-world conditions, a distinction with direct implications for how sample size should be determined.[36]

Statistical standardisation, expansion by outcome types and the era of software (1970s–1990s)

By the 1970s, the field of sample size estimation became increasingly standardised, driven by widely adopted textbooks and statistical training. The most influential work during this period was Jacob Cohen's *Statistical Power Analysis for the Behavioral Sciences*,[23,37] which formalised the use of effect size conventions and provided ready-to-use power tables for common designs. Cohen introduced thresholds for small (0.2), medium (0.5) and large (0.8) standardised mean differences (Cohen's *d*), which could be paired with formulas such as:

$$n = \left(\frac{Z_{1-\alpha/2} + Z_{1-\beta}}{d} \right)^2$$

to estimate the sample size required for a two-sample *t*-test under standard normal assumptions. He also provided analogous formulas for correlations, proportions and ANOVA, enabling psychologists and social scientists to conduct a priori power analyses even without deep statistical training.

As sample size became embedded in applied research, general principles of precision and error control increasingly gave rise to outcome-specific formulas tailored to different types of outcomes. In settings characterised by varying follow-up times or recurrent events, attention shifted from risks to event rates, naturally leading to models based on counts of events observed over person-time. These settings are well described by the Poisson distribution, which provides a likelihood-based framework for incidence rates.

By the mid-1970s to early 1980s, methodological work had extended to models and closed form formulas for event counts and incidence rates, typically derived using large-sample approximations to the distribution of estimated rates or log rate ratios.[25,38] Related developments addressed time-to-event outcomes, where sample size and power calculations were formulated in terms of the expected number of events rather than the number of participants, drawing on asymptotic properties of log-rank tests and proportional hazards models,[27,29,39] which were increasingly cited in applied trial planning. Some examples of early clinical trials employing formal sample size planning and calculations for different outcome types are shown in Table 1. To our knowledge, it is not currently possible to identify with certainty the first clinical trial that employed a formal a priori sample size calculation, reflecting both incomplete reporting and the gradual adoption of these methods in practice.

Concurrently, regulatory agencies began to formalise the requirement that clinical trials justify their sample size based on expected effect size, variance, significance level and desired power. Key milestones included the International Conference on Harmonization (ICH) E9 guideline[40] and the Food and Drug Administration's (FDA's) Guidance for Industry documents,[41] which mandated transparent documentation of sample size logic in trial protocols. These documents shifted sample size estimation from an academic best practice to a regulatory obligation. This institutional shift was reinforced by later meta-epidemiological work showing that smaller randomised trials, on average, report lower methodological quality and yield systematically different treatment effect estimates than larger trials in meta-analyses.[42,43] At the same time, evidence suggests that the adoption of formal sample size calculations often preceded their routine reporting in the medical literature.[44]

The emergence of dedicated statistical software in the 1980s and 1990s played a transformative role in the dissemination and standardisation of sample size estimation (Table 2). Commercial platforms automated increasingly complex calculations, facilitated exploration of alternative design assumptions and expanded access to advanced methods such as interim monitoring, sample size re-estimation and adaptive trial designs. Collectively, these tools reduced technical barriers and consolidated sample size planning as a routine component of applied research across disciplines. As sample size estimation became embedded within regulatory, educational and computational infrastructures, frequentist approaches reached a high degree of methodological maturity and routinisation.

The 2000s to the present: from computational flexibility to open-source accessibility

In the early 2000s, advances in computing power and statistical software made it practical to apply simulation-based methods for sample size estimation in increasingly complex study and trial designs. Monte Carlo simulation and resampling techniques were widely adopted to evaluate operating characteristics such as power and precision in hierarchical, non-linear and time-to-event models, particularly in settings where closed-form planning formulas were unavailable.[6] Although initially developed largely within a frequentist framework, these computational approaches also provided the foundation for later Bayesian planning methods.

During this period, Bayesian approaches to sample size determination gained wider adoption, with planning criteria typically defined in terms of posterior informativeness rather than fixed error rates. Common strategies

Table 2. Evolution of software and packages for sample size estimation and trial design 1990s–2020s.

Era	Software / Package(s)	Approx. introduction	Primary functions	Contribution to the field
Commercial	PASS	Early 1990s	Extensive sample size and power calculations across study designs	Expanded access to specialised sample size methods
Commercial	StatXact	Early 1990s	Exact methods for small-sample and non-parametric analyses	Enabled planning for studies where asymptotic formulas were unreliable
Commercial	East	1990s	Group sequential designs, adaptive trials, interim monitoring	Facilitated implementation of modern adaptive clinical trial designs
Commercial	ADDPLAN	1990s	Adaptive design planning, sample size re-estimation, interim analyses	Advanced practical adoption of adaptive methodologies
Commercial	SAS PROC POWER	Late 1990s	Sample size and power calculations for common statistical tests	Integrated sample size planning into mainstream statistical workflows
Commercial	nQuery	1990s	Sample size and power calculations	Widely adopted in pharmaceutical trial planning
Open-source ecosystem	R	1990s	Statistical programming environment	Democratised access to advanced sample size methods through community-developed packages
R package	pwr	2000s	Classical sample size and power calculations	Popularised routine power analysis in academia
R package	TrialSize	2000s	Sample size calculations	Extended access to trial-specific planning methods
R package	gsDesign	2000s	Group sequential and adaptive designs	Open-source alternative to commercial adaptive-design software
R package	rpact	2010s	Confirmatory adaptive clinical trial design	Regulatory-oriented open-source adaptive design platform
Open-source ecosystem	Python	2000s	Statistical computing and simulation	Expanded simulation-based and machine-learning-oriented planning approaches
Python package	statsmodels	2010s	Power analyses and classical statistical methods	Brought sample size estimation into the broader data science ecosystem
Python package	scikit-power (or equivalent)	2020s	Sample size and power calculations	Emerging open-source support for trial planning and power analysis
GUI platforms	JASP, jamovi, BlueSky Statistics	2010s	Point-and-click statistical analyses and power calculations	Reduced barriers for non-programmers
Web-based interfaces	Shiny-, Streamlit- and Dash-based applications	2010s–2020s	Interactive trial simulation and sample size exploration	Enabled broad dissemination of advanced methods through web applications

included targeting a desired posterior credible interval width or ensuring a high posterior probability that a treatment effect exceeded a clinically meaningful threshold.[45] These methods generally focused on inferential precision and evidence thresholds, rather than explicit optimisation of expected utility or loss.

In parallel, adaptive and group-sequential designs saw increasing use in the first decade of the 21st century, supported by increasingly sophisticated software platforms (Table 2). These developments broadened prospective study planning beyond fixed, closed-form formulas through interim analyses, sample size re-estimation and other adaptive methodologies, while largely retaining an inferential rather than decision-analytic orientation.

In recent years, the rapid expansion of open-source statistical ecosystems, particularly R and Python, together

with increasingly accessible graphical user interfaces, has transformed access to advanced sample size and power analyses (Table 2). Methods that were once largely confined to specialised commercial software, such as simulation-based estimation, Bayesian planning, adaptive designs and trial simulation, have become widely available through community-developed packages and web-based applications. The net effect has been the democratisation of advanced trial planning methods, making rigorous sample size estimation accessible to a global audience without requiring costly licences or advanced programming expertise. This trend shows no sign of abating, with continued development of open-source tools likely to further expand accessibility and functionality.

Taken together, however, these developments suggest an uneven maturity across frameworks. Frequentist sample size methods for explanatory trials are highly standardised, and Bayesian approaches have increasingly mature tools for planning based on posterior precision or decision thresholds. By contrast, as of 2026, sample size determination for explicitly pragmatic, decision-analytic trials, where utilities, costs and consequences must be weighed, remains comparatively underdeveloped, with few widely adopted methods or software implementations.[46,47] Related Bayesian developments have also begun to incorporate uncertainty in model performance and decision-relevant criteria such as assurance probabilities and value of information, particularly in the context of prediction model validation.[48] Lastly, recent work has begun to address this gap by proposing decision-analytic frameworks that link sample size to acceptable regret and the probability of incorrect treatment selection, often yielding substantially smaller or more interpretable sample sizes in pragmatic settings.[49]

Concluding remarks

In today's data-rich environment, the central question is no longer simply 'how much data is enough', but rather 'enough for what purpose'? For explanatory research, the statistical foundations of sample size estimation are largely mature. For pragmatic, decision-oriented studies, however, the problem remains open, echoing Schwartz and Lellouch's call to align design, analysis and sample size with the ultimate goal of the investigation. Future methodological and reporting guidance (e.g. CONSORT) may benefit from more explicitly recognising these distinctions between explanatory and pragmatic objectives when discussing sample size determination.

From Huygens' philosophical exploration of uncertainty, through Bernoulli's formalisation of the law of large numbers, to the statistical revolutions of the 20th century, the history of sample size estimation reflects the broader evolution of probability and applied statistics (Figure 1). Each era brought conceptual and practical advances: Neyman and Pearson's framework transformed hypothesis testing into a tool for pre-study planning; Fisher and his successors established principles of design efficiency and post-war methodologists linked statistical rigour to real-world constraints in medicine, economics and other disciplines. This institutionalisation was further reinforced by the introduction of reporting standards such as the CONSORT statement,[50–52] which require explicit reporting of sample size calculations in randomised trials. Despite these requirements, empirical evaluations suggest that a substantial proportion of published trials still lack complete or reliable sample size justifications, even in high-impact journals.[53]

The 21st century has built on these foundations with more advanced approaches such as adaptive designs (e.g. early stopping or sample size re-estimation), Bayesian approaches including informative priors sample size evaluations, as well simulation-based approaches to address more complex sample size estimations. Advances in computing and open-source ecosystems have expanded access to powerful analytical tools, lowering barriers for researchers across disciplines.

Collectively, these developments have transformed sample size estimation into a multifaceted discipline that informs inference, prediction and decision-making. Rather than converging on a single universal principle, modern sample size planning increasingly reflects trade-offs among statistical precision, feasibility, ethical constraints and decision relevance. As methods continue to evolve, the enduring challenge is not merely technical optimisation, but thoughtful integration of statistical theory with the practical and ethical realities of scientific inquiry.

For more information on the importance of having an adequate sample size for a clinical trial, see this entry in the Catalogue of Bias: Catalogue of Bias Collaboration et al.[54]

Declarations

Acknowledgements: The authors thank Sarah Klingenberg (Copenhagen Trial Unit) for her diligent support in identifying relevant 'early' randomised trials for Table 1. The authors also thank Ofir Harari (Redwood AI) for his thorough review of later versions of the manuscript, ensuring accuracy and correctness in the historical description of statistical and probabilistic theory and the two reviewers (Benjamin Djulbegovic and Robert Matthews) and the James Lind Library editors for their many helpful suggestions.

Competing interests: The authors declared no potential conflicts of interest with respect to the research, authorship and/or publication of this article.

Funding: KT received no financial support for the research, authorship and/or publication of this article. CG was supported by the Copenhagen Trial Unit.

Ethical approval: No ethical approvals were required for this paper.

Guarantor: KT.

Contributorship: KT and CG both conceptualised the paper. KT wrote the first draft and both contributed equally to subsequent revisions. KT drafted Figure 1 and Table 2 and contributed to revision of Table 1. CG drafted Table 1 and contributed to revisions of Figure 1 and Table 2.

Provenance: Invited article from the James Lind Library.

Use of generative AI: Generative AI tools were used for limited editorial assistance. The authors take full responsibility for all content, analyses, interpretations and conclusions presented in this manuscript.

References

1. Huygens C. *De Ratiociniis in Ludo Aleae* [On Reasoning in Games of Chance]. (English translation by Beebee Thomas O. (2006)). In: David FN (ed.) *Games, Gods and Gambling*. Dover Publications, 1657: 69–74.
2. Bernoulli J. *Ars Conjectandi (The Art of Conjecturing)*. Basel: Thurneysen Brothers, 1713.
3. Pearson K. James Bernoulli's Theorem. *Biometrika* 1925; 17: 201–210.
4. de Moivre A. *The Doctrine of Chances: Or, A Method of Calculating the Probabilities of Events in Play*. 2nd ed. London: W. Pearson, 1738.
5. Marquis de Laplace PS. *Théorie analytique des probabilités*. Paris: Courcier, 1812.
6. Chow S-C, Shao J and Wang H. *Sample Size Calculations in Clinical Research*. 2nd ed. Boca Raton, FL: Chapman and Hall/CRC, 2008.
7. Hanley JA and Bhatnagar S. The 'Poisson' distribution: History, reenactments, adaptations. *Am Stat* 2022; 74: 363–371.
8. Campbell MJ. Doing clinical trials large enough to achieve adequate reductions in uncertainties about treatment effects. *J R Soc Med* 2013; 106: 68–71.
9. Neyman J and Pearson ES. On the problem of the most efficient tests of statistical hypotheses. *Philos Trans R Soc A* 1933; 231: 289–337.
10. Cochran WG. *Sampling Techniques*. New York: John Wiley and Sons, 1953.
11. Cochran WG and Cox GM. *Experimental Designs*. 2nd ed. New York: John Wiley and Sons, 1957.
12. Fisher RA. *The Design of Experiments*. Edinburgh: Oliver and Boyd, 1935.
13. Kempthorne O. *The Design and Analysis of Experiments*. New York: John Wiley and Sons, 1952.
14. Mainland D. *The Treatment of Clinical and Laboratory Data: An Introduction to Statistical Ideas and Methods for Medical and Dental Workers*. Edinburgh: Oliver and Boyd, 1938.
15. Hill AB. *Principles of Medical Statistics*. London: The Lancet, 1937.
16. Mainland D. Problems of chance in clinical work. *Br Med J* 1936; 2: 221–224.
17. Hrobjartsson A, Gøtzsche PC and Gluud C. The controlled clinical trial turns 100 years: Fibiger's trial of serum treatment of diphtheria. *BMJ* 1998; 317: 1243.
18. Stolberg M. Inventing the randomized double-blind trial: the Nürnberg salt test of 1835. *J R Soc Med* 2006; 99: 642–643.
19. Medical Research Council. Streptomycin treatment of pulmonary tuberculosis: a Medical Research Council investigation. *BMJ* 1948; 2: 769–772.
20. Francis T, Jr, Korns RF, Voight RB, Boisen M, Hemphill FM and Napier JA. An evaluation of the 1954 poliomyelitis vaccine trials: summary report. *Am J Public Health* 1955; 45: 1–63.
21. Zubrod CG, Schneiderman M, Frei E, III, Brindley C, Lennard Gold GL, Shnider BI, et al. Appraisal of methods for the study of chemotherapy of cancer in man: comparative therapeutic trial of nitrogen mustard and triethylene thiophosphoramide. *J Chronic Dis* 1960; 11: 7–33.
22. Chouinard G, Annable L, Turnier L, Holobow N and Szkrumelak N. A double-blind randomized clinical trial of lithium carbonate and haloperidol in the treatment of acute mania. *Biol Psychiatry* 1985; 20: 353–365.
23. Cohen J. *Statistical Power Analysis for the Behavioral Sciences*. 1st ed. New York: Academic Press, 1977.
24. Werzberger A, Mensch B, Kuter B, Brown L, Lewis J, Sitrin R, et al. A controlled trial of a formalin-inactivated hepatitis A vaccine in healthy children. *N Engl J Med* 1992; 327: 453–457.
25. Brown CC and Green SB. Additional power computations for designing comparative Poisson trials. *Am J Epidemiol* 1982; 115: 752–758.
26. Earl HM, Rudd RM, Spiro SG, Ash CM, Geddes DM and Souhami RL. A randomised trial of etoposide and carboplatin versus etoposide and cisplatin in small-cell lung cancer. *Br J Cancer* 1991; 63: 800–804.
27. Freedman LS. Tables of the number of patients required in clinical trials using the logrank test. *Stat Med* 1982; 1: 121–129.
28. Scandinavian Simvastatin Survival Study Group. Randomised trial of cholesterol lowering in 4,444 patients with coronary heart disease: the Scandinavian Simvastatin Survival Study (4S). *Lancet* 1994; 344: 1383–1389.
29. Schoenfeld D. The asymptotic properties of nonparametric tests for comparing survival distributions. *Biometrika* 1981; 68: 316–319.
30. Wald A. *Sequential Analysis*. New York: John Wiley & Sons, 1947.
31. Wald A. Sequential tests of statistical hypotheses. *Annals of Mathematical Statistics* 1943; 14: 117–186.
32. Armitage P. Restricted sequential procedures. *Biometrika* 1957; 44: 9–26.
33. O'Brien PC and Fleming TR. A multiple testing procedure for clinical trials. *Biometrics* 1979; 35: 549–556.
34. Lan KKG and DeMets DL. Discrete sequential boundaries for clinical trials. *Biometrika* 1983; 70: 659–663.
35. Gluud C and Thorlund K. Long overdue recognition of Klim McPherson's 1974 article on sequential analysis of trial data. *J R Soc Med* 2025; 118: 304–307.
36. Schwartz D and Lellouch J. Explanatory and pragmatic attitudes in therapeutic trials. *J Chronic Dis* 1967; 20: 637–648.
37. Cohen J. *Statistical Power Analysis for the Behavioral Sciences*. 2nd ed. Hillsdale, NJ: Lawrence Erlbaum, 1988.
38. Gail M. Power computations for designing comparative Poisson trials. *Biometrics* 1974; 30: 231–237.
39. Schoenfeld DA. Sample-size formula for the proportional-hazards regression model. *Biometrics* 1983; 39: 499–503.
40. International Conference on Harmonisation (ICH). E9 Statistical Principles for Clinical Trials. 1996. See https://www.ema.europa.eu/en/documents/scientific-guideline/ich-e-9-statistical-principles-clinical-trials-step-5_en.pdf (last checked June 28, 2026)
41. FDA. Guidance for Industry: E9 Statistical Principles for Clinical Trials. U.S. Department of Health and Human Services. 1998. See https://www.fda.gov/regulatory-information/search-fda-guidance-documents/e9-statistical-principles-clinical-trials (last checked June 28, 2026)
42. Kjaergard LL, Villumsen J and Gluud C. Reported methodologic quality and discrepancies between large and small randomized trials in meta-analyses. *Ann Intern Med* 2001; 135: 982–989.
43. Gluud LL, Thorlund K, Gluud C, Woods L, Harris R and Sterne JAC. Correction: reported methodologic quality and discrepancies

between large and small randomized trials in meta-analyses. *Ann Intern Med* 2008; 149: 219.
44. Soares HP, Daniels S, Kumar A, Clarke M, Scott C, Swann S, et al. Bad reporting does not mean bad methods for randomised trials: observational study of randomised controlled trials performed by the Radiation Therapy Oncology Group. *BMJ* 2004; 328: 22–24.
45. Gelman A, Carlin JB, Stern HS, Dunson DB, Vehtari A and Rubin DB. *Bayesian Data Analysis*. 3rd ed. Boca Raton, FL: CRC Press, 2013.
46. Harari O, Bingham DR, Dean A and Higdon D. Computer experiments: prediction accuracy, sample size and model complexity revisited. *Stat Sin* 2018; 28: 899–919.
47. Austin PC, Sapp RJ and Tu JV. Informing power and sample size calculations when using propensity-score weighting with survival outcomes. *Stat Med* 2021; 40: 4317–4332.
48. Sadatsafavi M, Gustafson P, Setayeshgar S, Wynants L and Riley RD. Bayesian sample size calculations for external validation studies of risk prediction models. *Stat Med* 2026; 45: e70389.
49. Hozo I, Hemkens LG and Djulbegovic B. Sample size determination for decision-centered pragmatic trials. *J Clin Epidemiol*. Epub ahead of print. DOI: 10.1016/j.jclinepi.2026.112397.
50. Begg C, Cho M, Eastwood S, Horton R, Moher D, Olkin I, et al. Improving the quality of reporting of randomized controlled trials: the CONSORT statement. *JAMA* 1996; 276: 637–639.
51. Moher D, Schulz KF and Altman DG. The CONSORT statement: revised recommendations for improving the quality of reports of parallel-group randomized trials. *JAMA* 2001; 285: 1987–1991.
52. Schulz KF, Altman DG and Moher D. CONSORT 2010 statement: updated guidelines for reporting parallel group randomized trials. *BMJ* 2010; 340: c332.
53. Charles P, Giraudeau B, Dechartres A, Baron G and Ravaud P. Reporting of sample size calculation in randomised controlled trials: review. *BMJ* 2009; 338: b1732.
54. Catalogue of Bias Collaboration; Spencer EA, Brassey J, Mahtani K and Heneghan C. Wrong sample size bias. 2017. See https://catalogofbias.org/biases/wrong-sample-size-bias/ (last checked June 28, 2026)