PROOF-CARRYING HYPOTHESIS TESTING: CERTIFIED STATISTICAL INFERENCE BY AI

Anonymous authors
Paper under double-blind review

ABSTRACT

Progress in formalisation technology, such as Lean, has allowed advanced mathematical proofs to be confirmed mechanically by computers. As well as providing greater assurance around existing, human, mathematical proofs, an application of this technology is to validate mathematical proofs by AI. In this paper, we consider the corresponding problem of certifying statistical inference by AI. We focus on testing a parameter (or regression contrast) under the linear model. We propose a statistical protocol, proof-carrying hypothesis testing, in which an AI produces a test with a formal proof of validity against the specific null hypothesis, design matrix, and assumption class set by the scientist, who withholds the data. An important aspect of the design of the protocol is to confront major tensions and statistical paradoxes which are pertinent to this hypothesis testing problem, namely the Bahadur–Savage theorem, Lindley's paradox, and the notion that "all models are wrong". We demonstrate the protocol on thirteen datasets across science, policy, medicine and AI, with finite-sample p-values and e-values derived and machine-checked by AI, finding strong evidence in several cases under declared variance and error bounds. At those sample sizes, asymptotic p-values based on the CLT cannot be supported without further assumptions.

1 INTRODUCTION

The central question motivating this work is: "how can we trust what an AI has determined to be true based on data?". This question is presumably on the mind of any person contemplating the use of AI in medical guidance and decisions, criminal investigation and justice, social and economic policy, or scientific research.

Substantial progress in formalisation technology, due in particular to an enormous community effort to build up mathematical theory in Lean, has allowed systematic verification of mathematical proofs by AI. The recent formalisation of Fermat's Last Theorem (Anthropic, 2026), and the resolution of Navier-Stokes existence and smoothness, a Millennium Prize problem (OpenAI, 2026), indicate that the capabilities of both technologies — formalisation and AI — now extend to some of the hardest mathematical problems.

Possibly dwarfing these efforts is the multilateral investment in AI for science, encompassing national and international programmes (The White House, 2025; Department for Science, Innovation and Technology, 2025; European Commission, 2025; Ministry of Science and Technology of China, 2023; MEXT, 2026), dedicated science initiatives at each of the major AI laboratories (MIT Technology Review, 2026; Anthropic, 2025; TechCrunch, 2025; Google DeepMind, 2026), startups building autonomous laboratories (Flagship Pioneering, 2025; FutureHouse, 2025). Across these programmes, there is a hope in particular that AI could substantially improve analysis of large-scale data (The Royal Society, 2024; Department for Science, Innovation and Technology, 2025), and that advances in formalisation could certify results (Cornelio et al., 2026; Advanced Research and Invention Agency, 2026; Defense Advanced Research Projects Agency, 2026).

The difficulty of certifying a statistical analysis by AI is that this is not very well-defined, and many statisticians would view an attempt to formalise statistical inference as a losing proposition, because of the high degree of ambiguity, judgement and compromise involved in any statistical analysis.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

To better define the objective, it may be productive to analyse what proof verification actually comprises. Lean's logic is a dependent type theory (Martin-Löf, 1982; Coquand & Huet, 1988; Coquand & Paulin, 1990; de Moura & Ullrich, 2021), in which a proposition is a type, and a proof is a term of this type. Checking a proof is checking that the constructed term is of the declared type. By Gödel's incompleteness theorems (Gödel, 1931) the system can never prove itself to be consistent, and what has emerged is a product of many iterations deliberately circumventing identified paradoxes (Russell's (Russell, 1908), Girard's (Hurkens, 1995)).

There is already a substantial literature in Statistics on the foundations of inference, which has also been shaped by the discovery of paradoxes and impossibility theorems, including Hume's problem of induction (Hume, 1748), the Bahadur–Savage theorem (Bahadur & Savage, 1956), Lindley's paradox (Lindley, 1957), Simpson's paradox (Simpson, 1951), and more. Other tensions which have influenced the field include the notion that 'all models are wrong' (Box, 1976), motivating model-robust theory (Buja et al., 2019a), the use of asymptotic statements for finite sample inference, and the reproducibility crisis (Open Science Collaboration, 2015), motivating e-values (Shafer, 2021; Grünwald et al., 2024). Despite this comprehensive theory there is no computational system in which these tensions and paradoxes are satisfactorily addressed. We think many readers will have implicitly recognised that a mathematically-capable AI may essentially be *necessary* for such a system to work in practice.

In this paper, we present a protocol for hypothesis testing using AI. The central novelty is *proof carrying*: the AI supplies a proof of the validity of its proposal relative to the specific hypothesis testing problem given. A second distinctive feature is that we build the protocol in a way which directly confronts major statistical paradoxes and tensions, mirroring, in a limited way, how modern formalisation technology itself is designed as an answer to logical paradoxes. We focus on the widely encountered problem of testing a parameter under the linear model.

2 Related work

The key novelty of our proposal is that an AI develops and formalises a mathematical statement *within* the statistical protocol and we explain below why this difference over existing proposals is crucial.

There are many existing proposals for using AI for statistical inference on data. These include: agents that choose hypotheses and report p-values, calibration to e-values before combining, validity of each p-value assumed (Huang et al., 2025); a human-proved, machine-checked multiple-testing theorem wrapped around LLM-generated analyses, with each test called by a 'harness' (Sargsyan, 2025); an agent that proposes candidates while a frozen statistical referee judges them by betting (Qu et al., 2026); e-processes computed by a fixed procedure over LLM outputs (Sadhuka et al., 2025; Cho & Sun, 2026); betting strategies learned from the data as they arrive, within a function class (Podkopaev & Ramdas, 2023; Shekhar & Ramdas, 2024); and agents that report findings with no statistical guarantee (Miao et al., 2026; Lu et al., 2024).

Several other systems aim to certify statistical inference on data more generally. StatWhy and Why3-py verify that a program applies a named hypothesis test in accordance with annotated preconditions, such as the data being normally distributed (Kawamoto et al., 2025; Tanaka & Kawamoto, 2026). Tacet, targeting machine-learning leaderboards, is a formalised multiple testing accounting layer over given p-values (Abuah, 2026). Tea and Tisane choose a test on the analyst's behalf, by checking the preconditions of a fixed collection of classical tests against the data and substituting a rank-based test where a parametric approach is not supported (Jun et al., 2019; 2022). The software accompanying the e-value and anytime-valid literature provides sequential tests and confidence sequences whose exact validity is established under different conditions (Ly et al., 2026; Howard, 2021; Lindon et al., 2026). Tools such as statcheck recompute reported p-values from the accompanying test statistics and so verify the arithmetic of a published analysis, not its assumptions (Nuijten et al., 2016). More widely there are processes in science which aim to improve reproducibility of studies, such as replication studies (Open Science Collaboration, 2015), reporting guidelines (Schulz et al., 2010), and the pre-registration of hypotheses (Nosek et al., 2018), which is highly compatible with our proposal.

In all the above, where there is a theorem (formalised or not) proving validity, it is derived once, before any specific statistical problem is given. This means that the mathematical analysis is not specialised to the problem. A scientist cannot say, "my errors are a mixture of two normals with unknown weights" and expect this assumption to have been anticipated by the theory. This is why our protocol allows the AI to propose the test and prove its validity *after* the scientist has presented the problem.

The protocol recommends e-values, a measure of evidence which can be connected to betting. Our only novelty claim around this concept is its justification through Lindley's paradox, which is mathematically elementary but fundamental to our proposal. For other results and ideas surrounding e-values, see e.g. Shafer (2021); Vovk & Wang (2021); Grünwald et al. (2024); Ramdas et al. (2023); Ramdas & Wang (2025) and further references.

The protocol also recommends model-robust linear regression, which we have implemented in Lean following White (1980); MacKinnon & White (1985); Fahrmeir (1990); Abadie et al. (2014); Kuchibhotla et al. (2018; 2019). In the course of running the protocol in real data examples, several theorems were derived by AI, most of them instances of known results. We contrast these against existing results (Hoeffding, 1963; Bennett, 1962; Cui, 2026; Bentkus, 2004; Pinelis, 2006; Howard et al., 2020; 2021; Derumigny et al., 2025; Kuchibhotla et al., 2018; Abadie et al., 2014) in Section 5.1.

The impetus for this project was news, in the Lean community forum, of the development of the Central Limit Theorem (CLT), eventually merged into the core mathematical library Mathlib in March 2026 (Marion et al., 2026). There are several broader initiatives to formalise statistics, including Statlib (Statlib, 2026), StatsMLlib (StatsMLlib, 2026; Zhang et al., 2026; Sonoda et al., 2026) and an asymptotic theory formalised by LLM agents (Wei et al., 2026), as well as individual results (Degenne, 2023; Avigad et al., 2017; Tassarotti et al., 2021; Bagnall & Stewart, 2019). The focus of the present paper is less formalisation of theory, than the *use* of formalisation to certify statistical inference by AI.

The proposal is based on Lean (de Moura & Ullrich, 2021) and its mathematical library Mathlib (The mathlib Community, 2020), and is premised on the availability of an AI system which can both derive and formalise mathematics. Several models now exist to do this (Hubert et al., 2025; Ren et al., 2025; Lin et al., 2025; Wang et al., 2025; Achim et al., 2025; Anthropic, 2026).

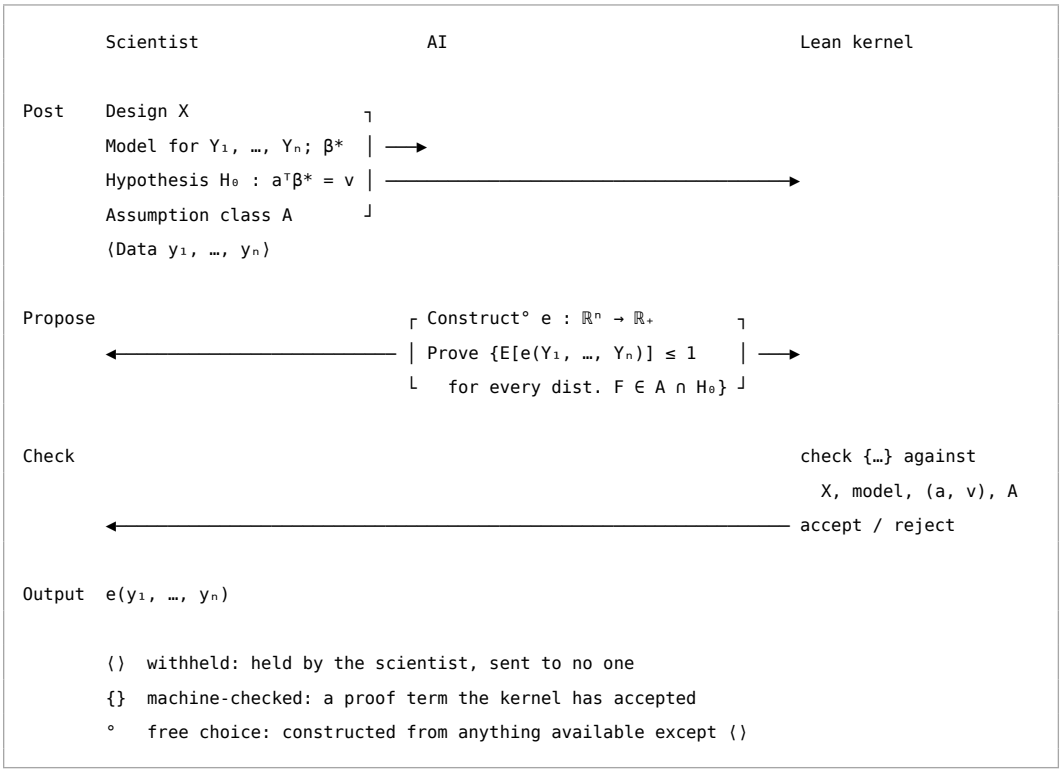
3 THE PROOF-CARRYING HYPOTHESIS TESTING PROTOCOL

Our proposed protocol is described in Box 1, assuming a design (or "feature") matrix $\mathbf{X}$, responses $Y_1, \ldots, Y_n$, null hypothesis H_0 and parameter β^*, all detailed in Section 4. The protocol comprises three parties: a scientist, an AI, and a proof checker, which we will take to be the Lean kernel. The scientist provides the null hypothesis, a (numeric) design matrix and assumption class (e.g. independence, and numeric error and standard deviation bounds). The observed values $y_1, \ldots, y_n$ are withheld. The AI returns an e-variable, a non-negative function e of n real arguments, together with a Lean proof that $\mathbb{E}_F[e(Y_1, \ldots, Y_n)] \leq 1$ for every distribution F satisfying $\mathcal{A}$ and H_0 (written $F \in \mathcal{A} \cap H_0$ hereafter); in constructing e the AI may use the design, the hypothesis, the assumption class, domain knowledge, and any mathematics it can prove, but not the data. The kernel checks the proof against the scientist's posting. The data are then entered and the number $e(y_1, \ldots, y_n)$ is output, an e-value for H_0.

The proof in the Propose step ensures that $e(y_1, \ldots, y_n)$ is a realisation of a valid e-variable: a non-negative function $E = e(Y_1, \ldots, Y_n)$ of the data whose expectation is at most one under every distribution F satisfying $\mathcal{A}$ and H_0 (Vovk & Wang, 2021). A high e-value is regarded as intrinsically strong evidence against H_0. If desired, it can be converted into a valid p-value via $p = 1/e$, as by Markov's inequality $\mathbb{P}_F(1/E \leq \alpha) \leq \alpha$ for any $\alpha \in [0, 1]$ for any distribution $F \in \mathcal{A} \cap H_0$.

Evidently a similar protocol can ask instead for a p-value $P = p(Y_1, \ldots, Y_n)$ with the guarantee $\mathbb{P}_F(P \leq \alpha) \leq \alpha$ for any $\alpha \in [0, 1]$ and any distribution $F \in \mathcal{A} \cap H_0$. Our preference for e-values is explained in Section 6.

What distinguishes the protocol from a statistical package is that the mathematics is not fixed in advance. Each posting defines a mathematical problem, and the AI's e-variable is a theorem



Box 1: Proof-carrying hypothesis testing protocol: message sequence

($\mathbb{E}\{e(Y_1, \ldots, Y_n)\} \leq 1$ for all $F \in \mathcal{A} \cap H_0$) and a proof. As an example of why the AI might adapt its strategy to the design, certain AI-derived bounds in Section 5 can be inapplicable when the leverage is too large.

We have named this protocol by analogy to proof-carrying code (Necula, 1997), a mechanism in which an "untrusted code producer must supply with the code a safety proof that attests to the code's adherence to a previously defined safety policy".

3.1 Risks

We perceive the main risks of this protocol to be as follows. The protocol can only be run once on a dataset, which means the assumption class has to be final. If it is too large, no evidence will be found, and if it is too small, the evidence is unreliable. In practice, prior iterations on synthetic data with the AI may be sensible, but the scientist should be wary of divulging properties of the data through this interaction (but the design can be revealed). The protocol holds for a proposal generated independently from Y, which may be hard to guarantee for historical data, some derivative of which the AI may have seen in training. However, the proposals are not 'black-boxes', they are mathematical statements which can be checked for suspicious targeting. The power of the protocol is a function of the capability of the AI, an inferior model is limited by what it can prove and formalise; also, nothing in the protocol obliges the AI to construct a strong proposal. The posting step in the protocol has a two-level "specification risk": the real problem being correctly transcribed into a mathematical hypothesis test; and then the latter being correctly transcribed into Lean. Finally, it may be debatable what is 'model' versus 'assumption': an important design choice in our implementation of the linear model was to define the model and projection parameter *without* assuming independence (equation (1), formalised without independence), so that this assumption is clearly signposted in $\mathcal{A}$. We discuss potential extensions of the protocol to allow multiple proposals and optional continuation in the supplementary material.

Table 1: Theorems behind Figure 1, derived and machine-checked as part of the protocol demonstration.

method	assumption class[‡]	p-value theorem	e-value theorem
Hoeffding	error upper bounds	`hoeffding_tail`	`hoeffding_evalue`
Bernstein	error and variance upper bounds	`bernstein_tail`	`bernstein_evalue`
two-point	error and variance upper bounds	`two_point_tail`	`two_point_evalue`
normal approx.[*]	variance range; third and fourth moment upper bounds	`finiteN_tail_twoSided_cantelli`	`calibrate_pvalue` (calibrated from p-value)
asymptotic[†]	moment conditions; $n \to \infty$	`hc0_pvalue_conservative`	N/A

[*]Berry–Esseen formalised elsewhere (Yan, 2026); constants not formalised (see Section 5.2). [†]Asymptotics not allowed within the protocol. [‡]Independence throughout.

4 Model-robust regression

Linear regression refers to a collection of statistical tools, designed to understand the mechanism by which variables predict an outcome, built around the characteristic equation

$$Y = \mathbf{X}\beta + e,$$

where $Y \in \mathbb{R}^n$ represents a collection of outcomes, $\mathbf{X} \in \mathbb{R}^{n \times p}$ holds the values of the corresponding predictor variables, $e \in \mathbb{R}^n$ represents the prediction errors, and $\beta \in \mathbb{R}^p$ is a vector of unknown coefficients to be estimated. We consider the ordinary least squares (OLS) estimate,

$$\hat{\beta} = \underset{\beta \in \mathbb{R}^p}{\arg\min} \|Y - \mathbf{X}\beta\|^2 = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top Y,$$

when the design matrix $\mathbf{X}$ is *deterministic*, and the outcomes in Y are random.

We treat the 'misspecified' case in which the mean of Y is not necessarily $\mathbf{X}\beta$ for some β, i.e., the true underlying relationship between outcome and predictor variables might be nonlinear. This general setting is known in the literature as assumption-lean (Berk et al., 2021; Vansteelandt & Dukes, 2022) or model-robust (White, 1980), although within this setting treating $\mathbf{X}$ as deterministic is perhaps atypical (Buja et al., 2019a; Kuchibhotla et al., 2021). We have opted for this treatment to align with mainstream use cases, e.g. two-group testing or trend estimation, where this is often the more realistic regime.

In the model-robust setting, $\hat{\beta}$ estimates an optimal linear approximation,

$$\beta^\star = \underset{\beta \in \mathbb{R}^p}{\arg\min}\, \mathbb{E}\|Y - \mathbf{X}\beta\|^2 = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mu, \qquad \mu_i = \mathbb{E}[Y_i], \tag{1}$$

which we take as the target of inference, whether or not the model is well-specified.

We consider hypothesis tests of the general form $H_0 : a^\top \beta^* = v$, where the contrast a and value v are given. For example, if β^* contains two coefficients, we could pick $v = 0$, and $a = (1, 0)^\top$ or $a = (1, -1)^\top$ to test respectively whether the first coefficient is zero or the two coefficients are equal.

Estimation of the mean is a special case of the (misspecified) linear model $p = 1$ with $\mathbf{X} = (1, \ldots, 1)^\top$, in which $\hat{\beta}$ is the sample mean and $\beta^\star$ the average of the μ_i. The Welch t-test for the difference in the mean outcomes of two groups, certainly the most frequently performed statistical procedure in science, also arises as a special case (Appendix C).

5 Protocol demonstration

5.1 Theorems generated

In implementing the protocol on 13 real data examples, several theorems were derived and formalised using AI, as summarised in Table 1.

The asymptotic p-value, implemented for heteroscedasticity-consistent standard errors HC0–HC3 (White, 1980; MacKinnon & White, 1985), closely followed the analysis in Kuchibhotla et al. (2018); the full treatment of HC0–HC3 under misspecification and deterministic design is not in the literature. The normal approximation is a finite-sample version of this argument which replaces the Lindeberg CLT (implemented here) with a Berry-Esseen CLT (implemented elsewhere, Yan, 2026). The closest existing result is Derumigny et al. (2025). Several human-AI iterations were necessary to remove the AI's insisted bound on misspecification, ultimately leading to a new result. The p-value is converted to an e-value using the standard calibration (formalised) $e = 1/(2\sqrt{p})$ (Vovk & Wang, 2021).

The Hoeffding and Bernstein p-values apply the classical exponential-moment bounds (Hoeffding, 1963; Bennett, 1962) to the error in the contrast $a^\top \hat{\beta}$, a weighted sum of independent bounded errors appearing in Cui (2026, Example 9.4.3) for a correctly specified model. The AI derived and formalised both, proving Bennett's moment-generating-function lemma from first principles, since Mathlib provides Hoeffding's lemma but not Bennett's. The 'two-point' route, relying on a two-point moment generating function bound, see e.g. Pinelis (2006, Lemma 3.2), was suggested to us by an expert. The statement in terms of a linear regression contrast appears to be new, but is a direct application of Bentkus (2004, Theorem 2.1).

The e-values of all three routes are the fixed-exponent e-variables of the same inequalities (Howard et al., 2020), averaged over a grid of exponents declared before the data are seen (Howard et al., 2021, Theorem 2).

Some (human) insight is given about the proof techniques in the supplementary material, Appendix B.

5.2 DATA

In this section we demonstrate the protocol on 13 datasets. In each case, we imagine we were the scientist looking at the data for the first time, with a relevant null hypothesis and contextual knowledge. Our assumption class always contains independence, and assumptions are selected extremely conservatively (with an exception explained below). For example, for Galton's data on children's and parents' heights (the discovery of regression to the mean), we assume errors of at most 18 inches with standard deviation at most 4.5 inches. The assumptions chosen for each dataset are given in the supplementary material.

The protocol is stated assuming one AI proposal and we discuss briefly how this could be extended to multiple agents in the supplementary material (Appendix A). In the following experiments, multiple proposals are presented simply to allow a method comparison.

The datasets were selected to cover consequential linear regressions across science, policy, medicine and AI, alongside standard pedagogical examples, with a public source for each analysis. We take a p-value to have been reported only if it is printed in some public source (sources given in the supplementary material). If an interval was printed, we select the upper bound. The datasets and assumptions-per-dataset are described in Appendix E.

The proof-carrying e-values and p-values derived by AI are shown in Figure 1, p-values on the left, and e-values on the right, on a log-scale, alongside the reported values, the Monte Carlo search, and the asymptotic p-values. The asymptotic validity of the asymptotic p-values is formalised, but these are not allowable within the protocol. An arrow indicates that the corresponding value lies beyond the plotting region and is extreme.

The exception mentioned above is the normal-approximation route and the accompanying Monte Carlo search, which are instead evaluated under stringent assumptions (equal variances, Gaussian moment ratios). This has two purposes. The first is to show that the normal-approximation route adopted (numerator/denominator split, Berry-Esseen above, Cantelli below) is not competitive. This also explains why we have calculated the bounds using the best constants available in the literature (Shevtsova, 2014; 2017) rather than what has currently been formalised (a constant of 30, Yan, 2026). The second is to demonstrate that several of the (extreme) asymptotic p-values cannot be supported based on *any* central limit theorem without some further assumption. We have conducted a search over the assumption class, and report the lowest p-value found, computed by Monte Carlo with precise error bounds; the class limit is at least this value.

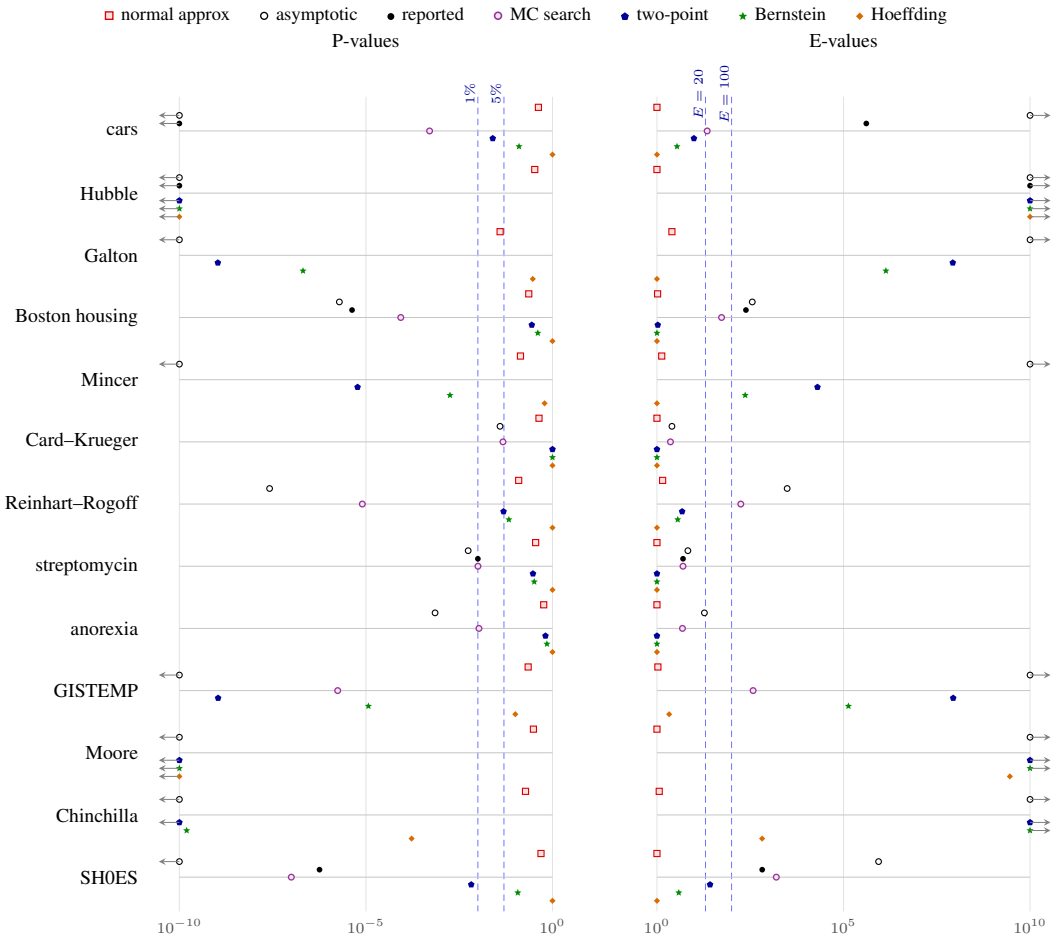


Figure 1: P-values (left, log-scale, further left is more significant; axis censored at 10^{-10}, an arrow indicating the value lies beyond) and e-values (right, log-scale, further right is more significant; axis censored at 10^{10}, an arrow indicating the corresponding value lies beyond). For the reported, asymptotic and normal-approximation routes the e-value shown is the calibrated value $1/(2\sqrt{p})$ (Vovk & Wang, 2021); for Hoeffding, Bernstein and two-point it is the mixture e-value of the route itself. The MC search marker is absent where the search found no member of the class reaching the observed statistic, providing no useful upper bound on the corresponding p-value. Dashed lines: significance levels 5% and 1%, and the corresponding Markov thresholds $E = 20$ and $E = 100$.

Conclusions which can be drawn from these experiments are: the use of asymptotic p-values for any of the (finite) sample sizes encountered in our examples is not supported by a central limit theorem, unless assumptions beyond moments, such as Gaussian shape, are made; machine-checked routes report strong e-value and p-value evidence in several cases; the best performing method however was suggested to us by an expert and not AI, and indeed this expert judgement also brought material improvements to the normal-approximation route allowing us to lose a misspecification tolerance; the choice of assumption class (e.g. moment versus error bounds) has large implications for the result and must be declared, well-motivated and conservative.

Finally, the demonstration departs from the protocol in considering datasets which are likely to have either entered, or at least not to be independent from, the data used to train AI. The bounds derived are generic, which limits what can have been gained, but in general this would compromise the validity of the hypothesis test.

6 DESIGN PRINCIPLES

The proof-carrying protocol has three design choices, model-robust versus well-specified, assumption class versus asymptotics, e-values versus p-values, direct answers to existing tensions and paradoxes in Statistics which we now discuss.

The implementation of model-robust theory is an answer to Box's famous "all models are wrong" (Box, 1976), as β^* is formally defined under only the assumption that the $Y_1, \ldots, Y_n$ follow *some* joint distribution with finite mean. A disadvantage of this approach is that β^* loses some interpretability, the risk of which is perhaps well-illustrated with Simpson's paradox. The conclusion that boys perform better than girls, when in fact they pick easier subjects, as the hidden subject covariate is left out of the design, gets labelled as 'robustness' to missing covariates, rather than a modelling error (Buja et al., 2019b).

The Bahadur–Savage theorem (Bahadur & Savage, 1956) (Lean statement below) motivates our preference for an explicit assumption class, over asymptotic p- or e-values. This result demonstrates that finite-sample inference about a mean (and by extension the parameters of a linear regression), is futile without assumptions. A p-value which is valid for any distribution with mean μ_0 has no power to detect any deviation from μ_0.

```
theorem bahadur_savage (P : ValidPValue n μ₀) (F : Distribution)
    (α : ℝ) (hα0 : 0 ≤ α) (hα1 : α ≤ 1) : F.prob {Y | P.value Y ≤ α} ≤ α
```

A corresponding result holds for e-values since any valid e-value provides a valid p-value. In view of this theorem, we have failed to make asymptotic p- or e-values possible within a formal framework, each of our attempts somehow exposing an assumption either unrealistic (e.g. data exactly or almost exactly Gaussian) or providing unusable bounds (see Section 5).

Lindley's paradox (Lindley, 1957) (Lean statement below) motivates our preference for e-values. Given i.i.d. normal(μ, σ^2) data in which we are testing $H_0 : \mu = \mu_0$, it is possible for data to simultaneously exhibit an arbitrarily small p-value (p) for H_0 and an arbitrarily high posterior (π) for H_0, even though H_0 and H_1 were equally likely a priori ($c = 1/2$ below) and μ is uniformly distributed over some interval around μ_0 under H_1 ($[\mu_0 - l, \mu_0 + l]$, below).

```
theorem lindley_paradox (hσ : 0 < σ) (hl : 0 < l) (hc0 : 0 < c) (hc1 : c < 1)
    (ε : ℝ) (hε0 : 0 < ε) (hε1 : ε < 1) :
    ∃ (n : ℕ) (Y : Fin n → ℝ), pValue σ μ₀ Y ≤ ε ∧ posterior σ μ₀ l c Y ≥ 1 - ε
```

The e-value $e = p_1(y_1, \ldots, y_n)/p_0(y_1, \ldots, y_n)$, where p_1 and p_0 are the marginal likelihoods of the data under each hypothesis 1) is valid (`eValue_isValidEValue`), 2) is equivalent in the sense $p = \mathbb{P}(E \geq e \mid H_0)$ (`pValue_eq_survival`), and 3) satisfies $e = (1 - \pi)/\pi$ (`eValue_equiv_posterior`). We cannot make both π and e arbitrarily large, so this e-value does not yield the paradox. Points 2) and 3) are mathematically straightforward, but we were surprised not to find them in the literature. E-values may therefore provide a reasonable compromise between frequentist and Bayesian approaches: evidence against H_0 can be established without an explicit model for H_1, but is in less direct conflict with the Bayesian conclusion when H_1 is made explicit. For further discussion see Shafer et al. (2011), Grünwald et al. (2024, §9) and Bickel (2026).

7 CONCLUSION

We have presented a statistical protocol, proof-carrying hypothesis testing, in which an AI produces a test and a formal proof of validity against a statistical problem set by the scientist, who withholds the data. In a demonstration on thirteen datasets, the AI derived and formalised finite-sample p-values and e-values for regression contrasts under declared error and variance bounds, finding strong evidence in several cases. At those sample sizes, asymptotic p-values based on the CLT cannot be supported without further assumptions.

Two qualifications should be made. The best-performing route was suggested to us by an expert rather than by the AI, and although validity is guaranteed, the *power* of the protocol is in general a function of the mathematical capability of the AI. Second, the demonstration departs from the protocol in considering datasets which are likely to have either entered, or at least not to be independent from, the data used to train AI.

A problem we have not solved is how asymptotic theory, central to statistical reasoning, could be used within a practical system of formal inference. Attempts to extract when 'n is large enough' from an asymptotic proof may fail, for example, when a step invokes a relationship between two entire sequences, $(a_m \to 0) \Rightarrow (b_m \to 0)$ so that a_n being small for the specific sample size n says nothing about b_n being small. Proof mining (Kohlenbach, 2008) and metastability (Tao, 2007) provide more sophisticated approaches for extracting finite-sample statements from asymptotic proofs, and whether these could bring acceptable guarantees on real data is an open question.

AI USE STATEMENT

The authors take full responsibility for the content of the paper and code.

We have made extensive use of AI in this project, but *no prose* in the main document is written by AI, and in the supplementary material only one section is written by AI, containing the derivations for an AI-calculated bound, and is clearly marked. AI was used for editing, feedback, data analysis including R scripts and Monte Carlo search, ideation, the figures and literature search, with every reference checked against its source.

Significant errors were found in AI outputs, some mentioned in the main document. The AI was certain a bound on misspecification could not be removed in the normal approximation route. An assumption of homoscedasticity also slipped in. The route does not need either, as the formalised statement now proves. The AI proposed very unrealistic moment bound assumptions for the datasets, which each had to be corrected. Arguably the worst errors were those of attribution, for example, AI was confident the entirety of asymptotic validity of p-value under misspecification was new, when it had appeared in essentially the same form in Kuchibhotla et al. (2018, §7), and in references therein.

Apart from formalising the proposals, obviously done by AI (that is the point of the paper), formalising supporting theory around the protocol was also substantially assisted by AI. This was alongside months of human coding and editing. The models Claude Opus 4.8, Fable 5, Fable 5.1, and Aristotle were used to write mathematical documents, followed by instruction files, then implemented in Lean, and then machine-checked.

To give a sense of the labour involved in human auditing, iterations over our code from 12 June to 20 July 2026 gave: Clt $1{,}323 \to 1{,}040$ lines (-21%); Ols $7{,}786 \to 4{,}170$ (-46%); ClassicalOls $1{,}196 \to 1{,}092$ (-9%); total $10{,}305 \to 6{,}302$ (-39%); theorems and lemmas $209 \to 225$ (density about $49 \to 28$ lines per theorem). Well after the first compilation, these are examples of AI coding behaviour which we sought to amend:

- A proof used bound B at two points; we later obtain a sharper bound B', but the AI only fits this into the proof at one of the two spots (the proof is still correct as we only need the sharper bound at one point, but this introduces two separate bound lemmas, which seems inefficient).
- The AI proves intermediate results that then are not actually used anywhere (this could be down to a revision which replaces the result with something new to fit a different proof approach).
- General messiness with commenting: comments frequently refer to the change the AI is making at the time, but if it is replacing a previous result then we have phantom references to things that no longer exist. Stylistically, the AI's comments are inconsistent: inconsistent numbering, inconsistent use of headers, inconsistent use of mathematical notation, etc. None of these affect the validity of results but make things more difficult for human comprehension.
- The AI will often do exactly as told, which sometimes misses some 'free wins'. For example, at one point we establish a sharp bound on the second-order Taylor expansion remainder for $\exp(iy)$, when in fact the proof structure can be used to get a bound for the

kth-order expansion. We only need the second order for our results, but it seems like a nice result to just have the general bound present, since it is only a line or two of extra work.

Reproducibility statement

The supplementary material contains the Lean 4 project, the analysis scripts and the data. Every theorem named in the paper is in the project's `Ols/` and `Clt/` directories; the project builds with `lake build` against the pinned Mathlib version, and the axioms of any theorem can be listed with `#print axioms`. The only `sorry` is the Berry–Esseen constant placeholder in `Clt/BerryEsseen.lean`. The R scripts construct the thirteen designs, evaluate the certified bounds under the declared assumptions of Appendix E, and draw Figure 1; the Monte Carlo search is in Python, with the exported designs, results and candidate distributions included. The values reported are the certified formulas evaluated by these scripts on the data, an ordinary computation that is not itself machine-checked; the Berry–Esseen constants used are literature values. Datasets shipped with R packages are loaded from those packages; the remaining public data files are included, or, for SH0ES, accompanied by download instructions, with every source listed in `DATA.md`.

Ethics statement

All datasets used are public, and the demonstration involves no new data collection. Our use of AI is disclosed above, and responsibility for the correctness of every statement, proof and citation rests with the authors. We have tried to follow the calls of the Leiden Declaration on Artificial Intelligence and Mathematics (Leiden Declaration Working Group, 2026): results obtained with AI are held to a higher standard, here machine-checking, and every reference was checked against its source, since models are known to omit or misattribute human work. Two further choices were made with the mathematical community in mind. Our formalisation strategy has been to build on existing community projects and not to undertake, with AI, work that others are doing by hand. Before submitting a paper with bounds derived by AI, we discussed these with researchers in the area; we are not aware of anyone pursuing those results.

References

Alberto Abadie, Guido W. Imbens, and Fanyin Zheng. Inference for misspecified models with fixed regressors. *Journal of the American Statistical Association*, 109(508):1601–1614, 2014.

Alberto Abadie, Susan Athey, Guido W. Imbens, and Jeffrey M. Wooldridge. Sampling-based versus design-based uncertainty in regression analysis. *Econometrica*, 88(1):265–296, 2020.

Chiké Abuah. Tacet: A language and type system for automatic statistical validity accounting. *arXiv preprint arXiv:2608.27451*, 2026.

Tudor Achim, Alex Best, Alberto Bietti, et al. Aristotle: IMO-level automated theorem proving, 2025.

Advanced Research and Invention Agency. Safeguarded AI. Programme page, `https://www.aria.org.uk/opportunity-spaces/mathematics-for-safe-ai/safeguarded-ai/`, 2026. Accessed 22 September 2026.

Anthropic. AI for Science program. `https://www.anthropic.com/news/ai-for-science-program`, 2025. 5 May 2025.

Anthropic. Formalizing Fermat's Last Theorem. `https://www.anthropic.com/research/formalizing-fermats-last-theorem`, 2026. 4 September 2026.

Jeremy Avigad, Johannes Hölzl, and Luke Serafin. A formally verified proof of the central limit theorem. *Journal of Automated Reasoning*, 59(4):389–423, 2017. doi: 10.1007/s10817-017-9404-x.

Alexander Bagnall and Gordon Stewart. Certifying the true error: Machine learning in Coq with verified generalization guarantees. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI-19)*, volume 33, pp. 2662–2669, 2019. doi: 10.1609/aaai.v33i01.33012662.

R. R. Bahadur and L. J. Savage. The nonexistence of certain statistical procedures in nonparametric problems. *The Annals of Mathematical Statistics*, 27(4):1115–1122, 1956. doi: 10.1214/aoms/1177728077.

George Bennett. Probability inequalities for the sum of independent random variables. *Journal of the American Statistical Association*, 57(297):33–45, 1962.

Vidmantas Bentkus. On Hoeffding's inequalities. *The Annals of Probability*, 32(2):1650–1673, 2004. doi: 10.1214/009117904000000360. arXiv:math/0410159.

Richard Berk, Andreas Buja, Lawrence Brown, Edward George, Arun Kumar Kuchibhotla, Weijie Su, and Linda Zhao. Assumption lean regression. *The American Statistician*, 75(1):76–84, 2021.

Tamay Besiroglu, Ege Erdil, Matthew Barnett, and Josh You. Chinchilla scaling: A replication attempt. arXiv:2404.10102, 2024.

David R. Bickel. The Bayes e-value: Testing by betting given a prior probability of the null hypothesis. *Journal of Statistical Theory and Practice*, 20(4):110, 2026. doi: 10.1007/s42519-026-00600-4.

George E. P. Box. Science and statistics. *Journal of the American Statistical Association*, 71(356):791–799, 1976. doi: 10.1080/01621459.1976.10480949.

Andreas Buja, Lawrence Brown, Richard Berk, Edward George, Emil Pitkin, Mikhail Traskin, Kai Zhang, and Linda Zhao. Models as approximations I: Consequences illustrated with linear regression. *Statistical Science*, 34(4):523–544, 2019a.

Andreas Buja, Lawrence Brown, Arun Kumar Kuchibhotla, Richard Berk, Edward George, and Linda Zhao. Models as approximations II: A model-free theory of parametric regression. *Statistical Science*, 34(4):545–565, 2019b. doi: 10.1214/18-STS694.

David Card and Alan B. Krueger. Minimum wages and employment: A case study of the fast-food industry in New Jersey and Pennsylvania. *American Economic Review*, 84(4):772–793, 1994.

Young Hyun Cho and Will Wei Sun. When should an AI workflow release? always-valid inference for black-box generate-verify systems. *arXiv preprint arXiv:2605.12947*, 2026.

Thierry Coquand and Gérard Huet. The calculus of constructions. *Information and Computation*, 76(2–3):95–120, 1988. doi: 10.1016/0890-5401(88)90005-3.

Thierry Coquand and Christine Paulin. Inductively defined types. In *COLOG-88*, Lecture Notes in Computer Science, pp. 50–66. Springer, 1990. doi: 10.1007/3-540-52335-9_47.

Cristina Cornelio, Takuya Ito, Ryan Cory-Wright, Sanjeeb Dash, and Lior Horesh. The need for verification in artificial intelligence-driven scientific discovery. *Philosophical Transactions of the Royal Society A*, 384(2317):20240591, 2026. doi: 10.1098/rsta.2024.0591.

Eliuvish Han Cui. The statistical compass: From data to decision, from measure to meaning. arXiv:2606.11282, 2026. Textbook/monograph working manuscript, 669 pages, submitted 9 June 2026; Example 9.4.3.

Leonardo de Moura and Sebastian Ullrich. The Lean 4 theorem prover and programming language. In *Automated Deduction – CADE 28*, Lecture Notes in Computer Science, pp. 625–635. Springer, 2021. doi: 10.1007/978-3-030-79876-5_37.

Defense Advanced Research Projects Agency. expMath: Exponentiating mathematics. Program page, `https://www.darpa.mil/research/programs/expmath-exponential-mathematics`, 2026. Accessed 22 September 2026.

Rémy Degenne. CLT: Proof of the central limit theorem in Lean. GitHub repository, `https://github.com/RemyDegenne/CLT`, 2023. README: the CLT for i.i.d. real-valued random variables is complete and upstreamed to Mathlib. Accessed 21 September 2026.

Department for Science, Innovation and Technology. AI for Science Strategy. Technical report, UK Government, 2025. 20 November 2025; deposited paper DEP2025-0762.

Alexis Derumigny, Lucas Girard, and Yannick Guyonvarch. Can we have it all? Non-asymptotically valid and asymptotically exact confidence intervals for expectations and linear regressions. arXiv:2507.16776, 2025. v2, 23 July 2025.

European Commission. A European strategy for artificial intelligence in science. COM(2025) 724, 8 October 2025; https://research-and-innovation.ec.europa.eu/strategy/strategy-research-and-innovation/our-digital-future/european-ai-science-strategy_en, 2025. RAISE pilot launched 3 November 2025 with EUR 107 million under Horizon Europe.

Ludwig Fahrmeir. Maximum likelihood estimation in misspecified generalized linear models. *Statistics*, 21(4):487–502, 1990.

Flagship Pioneering. Flagship Pioneering unveils Lila Sciences to build superintelligence in science. Press release, https://www.flagshippioneering.com/news/press-release/flagship-pioneering-unveils-lila-sciences-to-build-superintelligence-in-science, 2025. 10 March 2025.

Wendy L. Freedman et al. Final results from the Hubble Space Telescope Key Project to measure the Hubble constant. *The Astrophysical Journal*, 553:47–72, 2001.

FutureHouse. Announcing Edison Scientific. https://www.futurehouse.org/research-announcements/announcing-edison-scientific, 2025. 5 November 2025.

Francis Galton. Regression towards mediocrity in hereditary stature. *Journal of the Anthropological Institute of Great Britain and Ireland*, 15:246–263, 1886.

GISTEMP Team. GISS surface temperature analysis (GISTEMP), version 4. NASA Goddard Institute for Space Studies, https://data.giss.nasa.gov/gistemp/, 2026. Accessed September 2026.

Kurt Gödel. Über formal unentscheidbare Sätze der Principia Mathematica und verwandter Systeme I. *Monatshefte für Mathematik und Physik*, 38:173–198, 1931. doi: 10.1007/BF01700692.

Google DeepMind. Science: unlocking a new era of discovery with AI. https://deepmind.google/science/, 2026. Accessed 25 September 2026.

Peter Grünwald, Rianne de Heide, and Wouter Koolen. Safe testing. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 86(5):1091–1128, 2024. doi: 10.1093/jrsssb/qkae011.

D. J. Hand, F. Daly, A. D. Lunn, K. J. McConway, and E. Ostrowski. *A Handbook of Small Data Sets*. Chapman & Hall, 1993.

David Harrison and Daniel L. Rubinfeld. Hedonic housing prices and the demand for clean air. *Journal of Environmental Economics and Management*, 5:81–102, 1978.

Thomas Herndon, Michael Ash, and Robert Pollin. Does high public debt consistently stifle economic growth? A critique of Reinhart and Rogoff. *Cambridge Journal of Economics*, 38(2):257–279, 2014.

Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963.

Jordan Hoffmann et al. Training compute-optimal large language models. arXiv:2203.15556, 2022.

Steven R. Howard. confseq: Confidence sequences and uniform boundaries. Python and R library, version 0.0.11, https://github.com/gostevehoward/confseq, 2021. Accessed 21 September 2026.

Steven R. Howard, Aaditya Ramdas, Jon McAuliffe, and Jasjeet Sekhon. Time-uniform Chernoff bounds via nonnegative supermartingales. *Probability Surveys*, 17:257–317, 2020. doi: 10.1214/18-PS321.

Steven R. Howard, Aaditya Ramdas, Jon McAuliffe, and Jasjeet Sekhon. Time-uniform, nonparametric, nonasymptotic confidence sequences. *The Annals of Statistics*, 49(2):1055–1080, 2021. doi: 10.1214/20-AOS1991.

Kexin Huang, Ying Jin, Ryan Li, Michael Y. Li, Emmanuel Candès, and Jure Leskovec. Automated hypothesis validation with agentic sequential falsifications. In *Proceedings of the 42nd International Conference on Machine Learning (ICML 2025)*, volume 267 of *Proceedings of Machine Learning Research*, pp. 25372–25437. PMLR, 2025.

Edwin Hubble. A relation between distance and radial velocity among extra-galactic nebulae. *Proceedings of the National Academy of Sciences*, 15(3):168–173, 1929.

Thomas Hubert, Rishi Mehta, Laurent Sartran, et al. Olympiad-level formal mathematical reasoning with reinforcement learning. *Nature*, 651(8106):607–613, 2025. doi: 10.1038/s41586-025-09833-y.

David Hume. An enquiry concerning human understanding. In *The Clarendon Edition of the Works of David Hume*. Oxford University Press, 1748. doi: 10.1093/oseo/instance.00032980. Clarendon edition; first published 1748.

Antonius J. C. Hurkens. A simplification of Girard's paradox. In *Typed Lambda Calculi and Applications (TLCA '95)*, Lecture Notes in Computer Science, pp. 266–278. Springer, 1995. doi: 10.1007/BFb0014058.

Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. *An Introduction to Statistical Learning*. Springer, 2013.

Eunice Jun, Maureen Daum, Jared Roesch, Sarah Chasins, Emery Berger, René Just, and Katharina Reinecke. Tea: A high-level language and runtime system for automating statistical analysis. In *Proceedings of the 32nd Annual ACM Symposium on User Interface Software and Technology (UIST '19)*, pp. 591–603. ACM, 2019. doi: 10.1145/3332165.3347940.

Eunice Jun, Audrey Seo, Jeffrey Heer, and René Just. Tisane: Authoring statistical models via formal reasoning from conceptual and data relationships. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems (CHI '22)*, pp. 1–16. ACM, 2022. doi: 10.1145/3491102.3501888.

Yusuke Kawamoto, Kentaro Kobayashi, and Kohei Suenaga. StatWhy: Formal verification tool for statistical hypothesis testing programs. In *Computer Aided Verification (CAV 2025), Part II*, volume 15932 of *Lecture Notes in Computer Science*, pp. 216–230. Springer, 2025. doi: 10.1007/978-3-031-98679-6_10.

Ulrich Kohlenbach. *Applied Proof Theory: Proof Interpretations and their Use in Mathematics*. Springer Monographs in Mathematics. Springer, 2008. doi: 10.1007/978-3-540-77533-1.

Arun K. Kuchibhotla, Lawrence D. Brown, and Andreas Buja. Model-free study of ordinary least squares linear regression. *arXiv preprint arXiv:1809.10538*, 2018.

Arun K. Kuchibhotla, Lawrence D. Brown, Andreas Buja, and Junhui Cai. All of linear regression. *arXiv preprint arXiv:1910.06386*, 2019.

Arun Kumar Kuchibhotla, Alessandro Rinaldo, and Larry Wasserman. Berry–Esseen bounds for projection parameters and partial correlations with increasing dimension. *arXiv preprint arXiv:2007.09751*, 2021.

Leiden Declaration Working Group. Leiden declaration on artificial intelligence and mathematics. `https://leidendeclaration.ai/`, 2026. 2 June 2026; endorsed by the International Mathematical Union.

Yong Lin, Shange Tang, Bohan Lyu, et al. Goedel-Prover-V2: Scaling formal theorem proving with scaffolded data synthesis and self-correction, 2025.

D. V. Lindley. A statistical paradox. *Biometrika*, 44(1–2):187–192, 1957.

Michael Lindon, Dae Woong Ham, Martin Tingley, and Iavor Bojinov. Anytime-valid inference in linear models with applications to regression-adjusted causal inference. *Journal of the American Statistical Association*, 2026. doi: 10.1080/01621459.2026.2692052. arXiv:2210.08589.

Regina Y. Liu and Kesar Singh. Using i.i.d. bootstrap inference for general non-i.i.d. models. *Journal of Statistical Planning and Inference*, 43(1–2):67–75, 1995.

Chris Lu, Cong Lu, Robert Tjarko Lange, Jakob Foerster, Jeff Clune, and David Ha. The AI scientist: Towards fully automated open-ended scientific discovery, 2024. arXiv preprint.

Alexander Ly, Rosanne Turner, and Yongqi Wang. *safestats: Safe Anytime-Valid Inference*, 2026. URL `https://CRAN.R-project.org/package=safestats`. R package version 0.8.8.

James G. MacKinnon and Halbert White. Some heteroskedasticity-consistent covariance matrix estimators with improved finite sample properties. *Journal of Econometrics*, 29(3):305–325, 1985.

Etienne Marion, Thomas Zhu, and Rémy Degenne. feat: central limit theorem. `https://github.com/leanprover-community/mathlib4/pull/36208`, 2026. Merged into Mathlib, March 2026.

Per Martin-Löf. Constructive mathematics and computer programming. In *Logic, Methodology and Philosophy of Science VI*, volume 104 of *Studies in Logic and the Foundations of Mathematics*, pp. 153–175. North-Holland, 1982. doi: 10.1016/S0049-237X(09)70189-2.

Medical Research Council. Streptomycin treatment of pulmonary tuberculosis. *British Medical Journal*, 2:769–782, 1948.

MEXT. Basic strategic policy for promoting AI for Science. `https://www.mext.go.jp/b_menu/shingi/chousa/shinkou/077/aifors_strategy.html`, 2026. Formulated 31 March 2026; 2026–2030 designated an intensive reform period.

Jiacheng Miao, Jin Mu, Guanhua Chen, and James Zou. Fisher-R1: Training LLM agents for reliable hypothesis testing, 2026. arXiv preprint, August 2026.

Jacob Mincer. *Schooling, Experience, and Earnings*. National Bureau of Economic Research, 1974.

Ministry of Science and Technology of China. China launches project to promote use of AI in sci-tech research. Xinhua, via `https://english.www.gov.cn/statecouncil/ministries/202303/28/content_WS642224f1c6d0f528699dc4bd.html`, 2023. Deployment of "AI for Science" announced 28 March 2023.

MIT Technology Review. Inside OpenAI's big play for science. `https://www.technologyreview.com/2026/01/26/1131728/inside-openais-big-play-for-science/`, 2026. 26 January 2026.

George C. Necula. Proof-carrying code. In *Proceedings of the 24th ACM SIGPLAN-SIGACT Symposium on Principles of Programming Languages (POPL '97)*, pp. 106–119. ACM, 1997. doi: 10.1145/263699.263712.

Jerzy Neyman. On the application of probability theory to agricultural experiments. *Statistical Science*, 5(4):465–472, 1990. Translated and edited by D. M. Dabrowska and T. P. Speed from the 1923 Polish original.

Brian A. Nosek, Charles R. Ebersole, Alexander C. DeHaven, and David T. Mellor. The preregistration revolution. *Proceedings of the National Academy of Sciences*, 115(11):2600–2606, 2018. doi: 10.1073/pnas.1708274114.

Michèle B. Nuijten, Chris H. J. Hartgerink, Marcel A. L. M. van Assen, Sacha Epskamp, and Jelte M. Wicherts. The prevalence of statistical reporting errors in psychology (1985–2013). *Behavior Research Methods*, 48(4):1205–1226, 2016. doi: 10.3758/s13428-015-0664-2.

Open Science Collaboration. Estimating the reproducibility of psychological science. *Science*, 349 (6251):aac4716, 2015. doi: 10.1126/science.aac4716.

OpenAI. On the Navier–Stokes Millennium Prize problem. https://openai.com/index/navier-stokes-solution/, 2026. 8 September 2026.

Our World in Data. Transistors per microprocessor. https://ourworldindata.org/moores-law, 2026. Accessed September 2026.

Iosif Pinelis. Binomial upper bounds on generalized moments and tail probabilities of (super)martingales with differences bounded from above. In *High Dimensional Probability*, volume 51 of *IMS Lecture Notes–Monograph Series*, pp. 33–52. Institute of Mathematical Statistics, 2006. doi: 10.1214/074921706000000743. arXiv:math/0512301.

Aleksandr Podkopaev and Aaditya Ramdas. Sequential predictive two-sample and independence testing. In *Advances in Neural Information Processing Systems 36 (NeurIPS 2023)*, pp. 53275–53307, 2023. doi: 10.52202/075280-2319.

Bo Qu, Mingguang Chen, and Licheng Wang. Propose, don't judge: An anytime-valid referee for LLM agents that mine investment factors. *arXiv preprint arXiv:2609.27051*, 2026.

Aaditya Ramdas and Ruodu Wang. Hypothesis testing with e-values. *Foundations and Trends in Statistics*, 1(1–2):1–390, 2025. doi: 10.1561/3600000002.

Aaditya Ramdas, Peter Grünwald, Vladimir Vovk, and Glenn Shafer. Game-theoretic statistics and safe anytime-valid inference. *Statistical Science*, 38(4):576–601, 2023. doi: 10.1214/23-STS894.

Z. Z. Ren, Zhihong Shao, Junxiao Song, Huajian Xin, et al. DeepSeek-Prover-V2: Advancing formal mathematical reasoning via reinforcement learning for subgoal decomposition, 2025.

Adam G. Riess et al. A comprehensive measurement of the local value of the Hubble constant. *The Astrophysical Journal Letters*, 934:L7, 2022.

Bertrand Russell. Mathematical logic as based on the theory of types. *American Journal of Mathematics*, 30(3):222–262, 1908. doi: 10.2307/2369948.

Shuvom Sadhuka, Drew Prinster, Clara Fannjiang, Gabriele Scalia, Bonnie Berger, Aviv Regev, and Hanchen Wang. E-valuator: Reliable agent verifiers with sequential hypothesis testing, 2025. arXiv preprint, v2, May 2026.

Karen Sargsyan. Structural enforcement of statistical rigor in AI-driven discovery: A functional architecture, 2025. arXiv preprint, v3, July 2026.

Kenneth F. Schulz, Douglas G. Altman, David Moher, and for the CONSORT Group. CONSORT 2010 statement: updated guidelines for reporting parallel group randomised trials. *BMJ*, 340: c332, 2010. doi: 10.1136/bmj.c332.

Glenn Shafer. Testing by betting: A strategy for statistical and scientific communication. *Journal of the Royal Statistical Society Series A: Statistics in Society*, 184(2):407–431, 2021. doi: 10.1111/rssa.12647.

Glenn Shafer, Alexander Shen, Nikolai Vereshchagin, and Vladimir Vovk. Test martingales, Bayes factors and p-values. *Statistical Science*, 26(1):84–101, 2011. doi: 10.1214/10-STS347.

Shubhanshu Shekhar and Aaditya Ramdas. Nonparametric two-sample testing by betting. *IEEE Transactions on Information Theory*, 70(2):1178–1203, 2024. doi: 10.1109/TIT.2023.3305867.

Irina G. Shevtsova. On the absolute constants in the Berry–Esseen-type inequalities. *Doklady Mathematics*, 89(3):378–381, 2014. doi: 10.1134/S1064562414030338.

Irina G. Shevtsova. On the absolute constants in Nagaev–Bikelis-type inequalities. In Iosif Pinelis (ed.), *Inequalities and Extremal Problems in Probability and Statistics*, pp. 47–102. Elsevier, London, 2017. doi: 10.1016/B978-0-12-809818-9.00003-3.

E. H. Simpson. The interpretation of interaction in contingency tables. *Journal of the Royal Statistical Society Series B*, 13(2):238–241, 1951. doi: 10.1111/j.2517-6161.1951.tb00088.x.

Sho Sonoda, Kazumi Kasaura, Yuma Mizuno, Kei Tsukamoto, and Naoto Onda. Lean formalization of generalization error bound by Rademacher complexity and Dudley's entropy integral. In *17th International Conference on Interactive Theorem Proving (ITP 2026)*, volume 382 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pp. 8:1–8:17. Schloss Dagstuhl – Leibniz-Zentrum für Informatik, 2026. doi: 10.4230/LIPIcs.ITP.2026.8.

Statlib. Statlib: Lean formalization project for probability and statistics. `https://github.com/stat-lib/statlib`, 2026.

StatsMLlib. StatsMLlib: a Lean 4 library for probability, statistics and learning theory. `https://github.com/Lean-MoDS/StatsMLlib`, 2026. Announced 7 August 2026 on the Lean community forum; website `https://statsmllib.github.io/`.

Akira Tanaka and Yusuke Kawamoto. Why3-py: A tool for formal verification of hypothesis testing and meta-analysis in Python, 2026. arXiv preprint, July 2026.

Terence Tao. Soft analysis, hard analysis, and the finite convergence principle. Blog post, 23 May 2007, `https://terrytao.wordpress.com/2007/05/23/soft-analysis-hard-analysis-and-the-finite-convergence-principle/`, 2007. Reprinted in *Structure and Randomness*, American Mathematical Society, 2008.

Joseph Tassarotti, Koundinya Vajjha, Anindya Banerjee, and Jean-Baptiste Tristan. A formal proof of PAC learnability for decision stumps. In *Proceedings of the 10th ACM SIGPLAN International Conference on Certified Programs and Proofs (CPP 2021)*, pp. 5–17. ACM, 2021. doi: 10.1145/3437992.3439917.

TechCrunch. Microsoft wants to tap AI to accelerate scientific discovery. `https://techcrunch.com/2025/05/19/microsoft-wants-to-tap-ai-to-accelerate-scientific-discovery/`, 2025. 19 May 2025.

The mathlib Community. The Lean mathematical library. In *Proceedings of the 9th ACM SIGPLAN International Conference on Certified Programs and Proofs (CPP 2020)*, pp. 367–381. ACM, 2020. doi: 10.1145/3372885.3373824.

The Royal Society. Science in the age of AI: How artificial intelligence is changing the nature and method of scientific research. Technical report, The Royal Society, London, 2024. URL `https://royalsociety.org/news-resources/projects/science-in-the-age-of-ai/`. May 2024.

The White House. Launching the Genesis Mission. Executive Order 14363, `https://www.whitehouse.gov/presidential-actions/2025/11/launching-the-genesis-mission/`, 2025. 24 November 2025.

Stijn Vansteelandt and Oliver Dukes. Assumption-lean inference for generalised linear model parameters. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 84(3):657–685, 2022.

Vladimir Vovk and Ruodu Wang. E-values: Calibration, combination and applications. *The Annals of Statistics*, 49(3):1736–1754, 2021. doi: 10.1214/20-AOS2020.

Haiming Wang, Mert Unsal, Xiaohan Lin, et al. Kimina-Prover Preview: Towards large formal reasoning models with reinforcement learning, 2025.

Tingzhou Wei, Zeyu Zheng, Ethan X. Fang, and Junwei Lu. Hypothesis-disciplined multi-agent automated formalization of asymptotic statistical theory, 2026. arXiv preprint, June 2026.

B. L. Welch. The generalization of Student's problem when several different population variances are involved. *Biometrika*, 34(1–2):28–35, 1947.

Halbert White. A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica*, 48(4):817–838, 1980.

Simon N. Wood. *Generalized Additive Models: An Introduction with R*. Chapman and Hall/CRC, 2nd edition, 2017.

Jeffrey M. Wooldridge. *Introductory Econometrics: A Modern Approach*. Cengage, 2019.

Asher Yan. ProbabilityApproximation: Berry–Esseen bounds for independent sums. Lean 4 library, version 0.9.6, released 20 July 2026, 2026. URL `https://github.com/Polarnova/ProbabilityApproximation`. Apache-2.0 licence.

Yuanhe Zhang, Jason D. Lee, and Fanghui Liu. AI4SLT: Empirical processes in Lean 4 for formal statistical learning theory. In *Proceedings of the 43rd International Conference on Machine Learning (ICML 2026)*, 2026. PMLR volume and pages not yet available.

A THE BETTING ANALOGY

The world of betting gives a useful interpretation of the e-value (Shafer et al., 2011; Shafer, 2021; Ramdas et al., 2023; Grünwald et al., 2024): a bookmaker is imagined to believe the null model for the data; a bet with a stake of 1 unit offers a payout $e(y)$ depending on the realised value of the data y; this bet is acceptable to the bookmaker if (in their eyes) its expectation does not exceed 1. A large payout is evidence against H_0 in the plain sense that losing a great deal of money points to a flaw in the bookmaker's model.

The size of the assumption class can be interpreted as the bookmaker's advantage: The larger the class, the weaker the assumptions, the less any bet can pay. The Bahadur–Savage theorem (Bahadur & Savage, 1956) (Section 6) shows a case in which the assumption class is so large than no bet can have positive expected return even when the null hypothesis is false.

The betting picture also makes it intuitive how evidence can accumulate over time and across studies without hitting the difficulties that p-values meet under multiple comparisons and optional stopping. E-values of arbitrary dependence may be combined by averaging, just as a bookmaker expects to win, on average, against multiple punters. It is perfectly valid to report the present e-value, or choose ask for more data and continue, just as a bookmaker expects to win whether the punter stops or decides to continue betting based on past outcomes. Obviously we do not view these analogies as formal guarantees, but they evoke the possibility of extending the protocol to a full "betting market" involving multiple agents, staged release of data, and continuation of studies.

B DESCRIPTION OF PROOFS

With $\varepsilon_i = y_i - \mu_i$ denoting the true errors, we have

$$T_{\texttt{raw}} = a^\top(\hat{\beta} - \beta^\star) = \sum \gamma_i \varepsilon_i, \tag{2}$$

where the weights $\gamma_i = a^\top(\mathbf{X}^\top\mathbf{X})^{-1}x_i$, with x_i the ith row of $\mathbf{X}$, are known (the subscript `raw` standing for unnormalised).

The RHS is a weighted sum of independent zero mean random variables, and is completely unaffected by any model misspecification.

The following bounds are now used; for $\lambda \geq 0$

$$\mathbb{E}\big(e^{\lambda\gamma_i\varepsilon_i}\big) \;\leq\; \exp\big(\lambda^2\gamma_i^2 B_i^2/2\big) \qquad \text{if } \varepsilon_i \in [-B_i, B_i] \tag{3}$$

$$\mathbb{E}\big(e^{\lambda\gamma_i\varepsilon_i}\big) \;\leq\; \exp\Big(\tfrac{\lambda^2\gamma_i^2 s_i^2}{2(1-\lambda M/3)}\Big) \qquad \text{if } \varepsilon_i \in [-B_i, B_i], \mathrm{sd}(\varepsilon_i) \leq s_i;\, 0 \leq \lambda < 3/M \tag{4}$$

$$\mathbb{E}\big(e^{\lambda\gamma_i\varepsilon_i}\big) \;\leq\; p_i e^{\lambda b_i} + (1-p_i)e^{-\lambda c_i} \qquad \text{if } \varepsilon_i \in [-B_i, B_i], \mathrm{sd}(\varepsilon_i) \leq s_i \tag{5}$$

where $M = \max_i |\gamma_i| B_i$, $b_i = |\gamma_i| B_i$, $c_i = |\gamma_i| s_i^2/B_i$ and $p_i = s_i^2/(B_i^2 + s_i^2)$. The first is Hoeffding's lemma (Hoeffding, 1963); the second is Bennett's lemma (Bennett, 1962); the third is the two-point moment generating function bound, see e.g. Pinelis (2006, Lemma 3.2).

Whichever line above we use, denoting the RHS by u_i, the variable $E_\lambda := e^{\lambda T_{\mathtt{raw}}} / \prod_i u_i$ is a valid e-value since:

$$\mathbb{E}\Big(\frac{e^{\lambda T_{\mathtt{raw}}}}{\prod_i u_i}\Big) \;=\; \prod_i \frac{\mathbb{E}\big(e^{\lambda \gamma_i \varepsilon_i}\big)}{u_i} \;\le\; 1,$$

using independence of the errors. Corresponding p-values are obtained through Markov's inequality (for nonnegative X and $t > 0$, $\mathbb{P}(X \ge t) \le \mathbb{E}(X)/t$): for $\lambda > 0$,

$$\mathbb{P}(T_{\mathtt{raw}} \ge t) \;=\; \mathbb{P}\big(e^{\lambda T_{\mathtt{raw}}} \ge e^{\lambda t}\big) \;\le\; e^{-\lambda t}\,\mathbb{E}\big(e^{\lambda T_{\mathtt{raw}}}\big) \;\le\; e^{-\lambda t} \prod_i u_i.$$

Since the bound holds for any $\lambda > 0$, the value λ can be chosen to make the bound small, after observing the data. This produces the standard Hoeffding and Bernstein bounds, and for the two-point case we optimise the bound numerically (which is also valid).

Observe that optimising λ after observing the data is *not valid* in the case of an e-value. Instead, we report the (valid) average of $E_{\lambda_1}, E_{-\lambda_1}, \ldots, E_{\lambda_k}, E_{-\lambda_k}$ (in +/- pairs because we are conducting a two-sided test) over a fixed-in-advance grid $\lambda_1, \ldots, \lambda_k$.

With $\hat\varepsilon_i = y_i - x_i^\top \hat\beta$ denoting the residuals, let

$$T_{\mathtt{norm}} \;=\; \frac{T_{\mathtt{raw}}}{D}, \qquad D^2 \;=\; \sum_i \gamma_i^2\, d_i\, \hat\varepsilon_i^2.$$

D is a heteroscedasticity-consistent (HC) standard error (White, 1980; MacKinnon & White, 1985), with 4 variants

$$d_i^{\mathrm{HC0}} = 1, \qquad d_i^{\mathrm{HC1}} = \frac{n}{n-p}, \qquad d_i^{\mathrm{HC2}} = \frac{1}{1-h_i}, \qquad d_i^{\mathrm{HC3}} = \frac{1}{(1-h_i)^2},$$

where $h_i = x_i^\top (\mathbf{X}^\top \mathbf{X})^{-1} x_i$ are the leverages.

Under regularity conditions $T_{\mathtt{raw}}/\mathrm{sd}(T_{\mathtt{raw}}) \rightsquigarrow \mathrm{normal}(0,1)$ and $D^2 / \sum_i \gamma_i^2(\sigma_i^2 + r_i^2) \xrightarrow{p} 1$, where $\sum_i \gamma_i^2(\sigma_i^2 + r_i^2) \ge \mathrm{Var}(T_{\mathtt{raw}})$ (convergence in distribution and in probability, respectively).

The resulting theoretical picture might come as a 'pleasant surprise': the reported uncertainty is conservative, over-estimating the variability, catering for a *worst-case* scenario in which model is *correct*. This asymptotic over-estimation of the variance under misspecification is known in several places (Fahrmeir, 1990; Liu & Singh, 1995; Abadie et al., 2014; Kuchibhotla et al., 2019; Neyman, 1990; Abadie et al., 2020), the analysis by (Kuchibhotla et al., 2018, Section 3) being the closest to what we have formalised. To our knowledge, the full treatment of HC0–HC3 under misspecification and deterministic design is not in the literature.

Consequently, the p-value $2\{1 - \Phi(|T_{\mathtt{norm}}|)\}$, where Φ is the normal distribution function, is asymptotically conservative:

$$\limsup_{n\to\infty}\; \mathbb{P}\big(2\{1 - \Phi(|T_{\mathtt{norm}}|)\} \le \alpha\big) \;\le\; \alpha \quad \text{for every } \alpha,$$

A finite-sample version of this p-value is obtained through the following inclusion, holding for any $\theta \in (0,1)$, with $\nu := \mathrm{sd}(T_{\mathtt{raw}})$, $\nu^2 = \sum_i \gamma_i^2 \sigma_i^2$, $\sigma_i^2 = \mathrm{Var}(\varepsilon_i)$:

$$\big\{|T_{\mathtt{norm}}| \ge t\big\} \;\subseteq\; \big\{|T_{\mathtt{raw}}| \ge t\sqrt{\theta}\,\nu\big\} \;\cup\; \big\{D^2 \le \theta\nu^2\big\}$$

(on the event $|T_{\mathtt{norm}}| \ge t$, either a) $D^2 > \theta\nu^2$, implying $|T_{\mathtt{raw}}| = |T_{\mathtt{norm}}|\, D \ge t\sqrt{\theta}\,\nu$, or b) $D^2 \le \theta\nu^2$), so that

$$\mathbb{P}\big(|T_{\mathtt{norm}}| \ge t\big) \;\le\; \mathbb{P}\big(|T_{\mathtt{raw}}| \ge t\sqrt{\theta}\,\nu\big) \;+\; \mathbb{P}\big(D^2 \le \theta\nu^2\big).$$

The value θ is chosen at the end.

The first probability is bounded by the Berry–Esseen theorem for independent, non-identical summands, in its combined uniform and nonuniform form: for every z,

$$\Big|\mathbb{P}\Big(\tfrac{T_{\mathtt{raw}}}{\nu} \le z\Big) - \Phi(z)\Big| \;\le\; \varrho\, \min\Big\{0.5583,\; \frac{21.82}{(1+|z|)^3}\Big\}, \qquad \varrho = \frac{\sum_i |\gamma_i|^3\, \mathbb{E}|\varepsilon_i|^3}{\nu^3},$$

the two constants being literature values, not machine-checked (Shevtsova, 2014; 2017). The ratio ϱ is not computable, because the moments of the errors are unknown, but assuming $\mathbb{E}|\varepsilon_i|^3 \le b_3\sigma_i^3$ and $\sigma_i \in [s_{\min}, s_{\max}]$, the numerator is at most $b_3 s_{\max}^3 \sum_i |\gamma_i|^3$ and $\nu^3 \ge s_{\min}^3 (\sum_i \gamma_i^2)^{3/2}$, so

$$\varrho \;\le\; \bar{\varrho} \;\coloneqq\; b_3 \Big(\frac{s_{\max}}{s_{\min}}\Big)^3 \frac{\sum_i |\gamma_i|^3}{\big(\sum_i \gamma_i^2\big)^{3/2}},$$

and hence

$$\mathbb{P}\big(|T_{\texttt{raw}}| \ge t\sqrt{\theta}\,\nu\big) \;\le\; 2\big\{1-\Phi(t\sqrt{\theta})\big\} \;+\; 2\,\bar{\varrho}\,\min\Big\{0.5583,\; \frac{21.82}{(1+t\sqrt{\theta})^3}\Big\}.$$

The second probability is bounded by Cantelli's inequality, the one-sided form of Chebyshev's: for a random variable X with finite variance and any $a > 0$,

$$\mathbb{P}\big(X - \mathbb{E}X \le -a\big) \;\le\; \frac{\mathrm{Var}(X)}{\mathrm{Var}(X) + a^2}.$$

Applied to $X = D^2$ with $a = \Delta \coloneqq \mathbb{E}D^2 - \theta\nu^2$, the *margin* of the mean of D^2 over the threshold, this gives, whenever $\Delta > 0$ (a condition on θ),

$$\mathbb{P}\big(D^2 \le \theta\nu^2\big) \;\le\; \frac{\mathrm{Var}(D^2)}{\mathrm{Var}(D^2) + \Delta^2}. \tag{6}$$

Write $\hat{\varepsilon} = r + \tilde{\varepsilon}$, where $r = \mu - \mathbf{X}\beta^\star$ is the model error and $\tilde{\varepsilon} = (\mathbf{I} - \mathbf{H})\varepsilon$, and $w_i = \gamma_i^2 d_i$. Then

$$D^2 \;=\; \sum_i w_i (r_i + \tilde{\varepsilon}_i)^2 \;=\; \sum_i w_i r_i^2 \;+\; 2\sum_i w_i r_i \tilde{\varepsilon}_i \;+\; \sum_i w_i \tilde{\varepsilon}_i^2.$$

The middle term has mean zero, so the margin $\Delta = \mathbb{E}D^2 - \theta\nu^2$ grows with misspecification, through $\rho^2 \coloneqq \sum_i w_i r_i^2$ (bringing the bound (6) down); but so does $\mathrm{Var}(D^2)$, through the same middle term (pushing it up). However, Δ^2 grows like ρ^4 and $\mathrm{Var}(D^2)$ at most like ρ^2, so for large ρ the gain in (6) through Δ^2 outweighs the loss through $\mathrm{Var}(D^2)$, and taking the worst case over all r gives a misspecification-free bound. Concretely one obtains

$$\frac{\mathrm{Var}(D^2)}{\Delta^2} \;\le\; \bar{q}(\theta) \qquad \text{for every error distribution in the class and every } r,$$

and since $x/(x+y)$ increases in x and decreases in y, Cantelli's bound becomes

$$\mathbb{P}\big(D^2 \le \theta\nu^2\big) \;\le\; \frac{\mathrm{Var}(D^2)}{\mathrm{Var}(D^2) + \Delta^2} \;\le\; \frac{\bar{q}(\theta)}{1 + \bar{q}(\theta)}.$$

The expression for $\bar{q}(\theta)$ and its derivation are in Appendix D; it is valid for every θ at which the margin is positive over the declared class, which under HC3 is every $\theta \in (0, 1)$, and no other condition on the leverages is imposed. The derivation was done by AI and is machine-checked for all of HC0–HC3 (`finiteN_tail_twoSided_cantelli`).

Assembling the two parts, for every admissible θ,

$$\mathbb{P}\big(|T_{\texttt{norm}}| \ge t\big) \;\le\; 2\big\{1-\Phi(t\sqrt{\theta})\big\} \;+\; 2\,\bar{\varrho}\,\min\Big\{0.5583,\; \frac{21.82}{(1+t\sqrt{\theta})^3}\Big\} \;+\; \frac{\bar{q}(\theta)}{1+\bar{q}(\theta)},$$

and since the bound is valid for each such θ, the reported p-value is the minimum over θ.

C Welch t-test

The Welch t-test for the difference in the mean outcomes of two groups, certainly the most frequently performed statistical procedure in science, arises as a special case. The outcomes are recorded in y and in this case the rows of $\mathbf{X}$ are set either to $(1, 0)$ or $(0, 1)$, encoding group membership. The Welch statistic (Welch, 1947) is

$$\frac{\bar{y}_A - \bar{y}_B}{\sqrt{s_A^2/n_A + s_B^2/n_B}} = \frac{a^\top \hat{\beta}}{\sqrt{a^\top \hat{\mathbf{V}}^{\mathrm{HC2}}\, a}},$$

where,

- on the left: $\bar{y}_g$, s_g^2 and n_g are the sample mean, variance and size of group g;
- on the right: $\hat{\beta}$ is the ordinary least squares estimate, $\hat{\mathbf{V}}^{\mathrm{HC2}}$ is the "HC2" covariance estimate for $\hat{\beta}$ (both defined below), and $a = [1, -1]^\top$.

The identity above is proved in theorem `hc2_studentized_twoSample_eq_welch` of the file `Ols/Welch.lean`.

D THE DENOMINATOR TERM OF THE NORMAL-APPROXIMATION BOUND [AI]

Notation as in Section 4: $w_i = \gamma_i^2 d_i$, $\mathbf{W} = \mathrm{diag}(w)$, $\hat{\varepsilon} = r + \tilde{\varepsilon}$ with $\tilde{\varepsilon} = (\mathbf{I} - \mathbf{H})\varepsilon$, and

$$D^2 \;=\; \rho^2 + A + B, \qquad \rho^2 = \sum_i w_i r_i^2, \quad A = 2\sum_i w_i r_i \tilde{\varepsilon}_i = 2u^\top \varepsilon, \quad B = \sum_i w_i \tilde{\varepsilon}_i^2 = \varepsilon^\top \mathbf{Q} \varepsilon,$$

with $u = (\mathbf{I} - \mathbf{H})\mathbf{W} r$ and $\mathbf{Q} = (\mathbf{I} - \mathbf{H})\mathbf{W}(\mathbf{I} - \mathbf{H})$. Since $\mathbb{E}A = 0$ and $\mathbb{E}B = \sum_j Q_{jj}\sigma_j^2$,

$$\Delta \;=\; \mathbb{E}D^2 - \theta\nu^2 \;=\; \rho^2 + \sum_j c_j \sigma_j^2, \qquad c_j \coloneqq Q_{jj} - \theta\gamma_j^2,$$

and, the first term of D^2 being a constant, $\mathrm{Var}(D^2) = \mathrm{Var}(B) + \mathrm{Var}(A) + 2\,\mathrm{Cov}(A, B)$. Let $g = \max_i w_i$ and $S = (\sum_i Q_{ii}^2)^{1/2}$. Under the declared constants, using independence of the errors, $\mathbb{E}\varepsilon_i^4 - \sigma_i^4 \le (\kappa - 1)\sigma_i^4$, the Cauchy–Schwarz inequality, and $\|u\| \le \|\mathbf{W} r\| \le \sqrt{g}\,\rho$ (the first because $\mathbf{I} - \mathbf{H}$ is an orthogonal projection),

$$\begin{aligned}
\mathrm{Var}(B) &= \sum_i Q_{ii}^2 (\mathbb{E}\varepsilon_i^4 - \sigma_i^4) + 2\sum_{i \ne j} Q_{ij}^2 \sigma_i^2 \sigma_j^2 \le V_0 \coloneqq s_{\max}^4 \Big\{ (\kappa - 1) \sum_i Q_{ii}^2 + 2 \sum_{i \ne j} Q_{ij}^2 \Big\}, \\
\mathrm{Var}(A) &= 4 \sum_i u_i^2 \sigma_i^2 \;\le\; 4 s_{\max}^2 g\, \rho^2, \\
|\mathrm{Cov}(A, B)| &= 2 \Big| \sum_i Q_{ii} u_i\, \mathbb{E}\varepsilon_i^3 \Big| \;\le\; 2 b_3 s_{\max}^3 S \|u\| \;\le\; 2 b_3 s_{\max}^3 S \sqrt{g}\, \rho, \\
\Delta &= \rho^2 + \sum_j c_j \sigma_j^2 \;\ge\; \rho^2 + a_-, \qquad a_- \;\coloneqq\; s_{\min}^2 \sum_{j:\, c_j \ge 0} c_j + s_{\max}^2 \sum_{j:\, c_j < 0} c_j,
\end{aligned}$$

the last line placing each σ_j^2 at the end of $[s_{\min}^2, s_{\max}^2]$ that minimises $c_j \sigma_j^2$, so that a_- is the smallest margin over the declared class (attained at $r = 0$). The Cantelli ratio $x/(x + y)$ increases in x and decreases in y, so, for any r,

$$\frac{\mathrm{Var}(D^2)}{\mathrm{Var}(D^2) + \Delta^2} \;\le\; \frac{V(\rho)}{V(\rho) + (\rho^2 + a_-)^2}, \qquad V(\rho) \coloneqq V_0 + 4 s_{\max}^2 g\, \rho^2 + 4 b_3 s_{\max}^3 S \sqrt{g}\, \rho.$$

The right-hand side is a function of ρ alone; it is less than 1 for every ρ (when $a_- > 0$) and tends to 0 as $\rho \to \infty$, so its supremum over $\rho \ge 0$ is attained and is less than 1: this is the misspecification-free bound. A closed form that avoids the maximisation: writing $q(\rho) = V(\rho)/(\rho^2 + a_-)^2$, the bound is $q/(1 + q)$, increasing in q, and the three terms of q are separately maximised at $\rho^2 = 0$, $\rho^2 = a_-$ and $\rho^2 = a_-/3$ (using $(x + a_-)^2 \ge 4 x a_-$ and $\sqrt{x}/(x + a_-)^2 \le \frac{3\sqrt{3}}{16} a_-^{-3/2}$), giving

$$q(\rho) \;\le\; \bar{q}(\theta) \;\coloneqq\; \frac{V_0}{a_-^2} + \frac{s_{\max}^2 g}{a_-} + \frac{3\sqrt{3}}{4} \cdot \frac{b_3 s_{\max}^3 S \sqrt{g}}{a_-^{3/2}}, \qquad \mathbb{P}\big(D^2 \le \theta\nu^2\big) \;\le\; \frac{\bar{q}(\theta)}{1 + \bar{q}(\theta)},$$

which overstates $\sup_\rho q(\rho)$ by at most a factor of three. The only condition is $a_-(\theta) > 0$. Keeping only the $i = j$ term of Q_{jj} gives $c_j \ge \gamma_j^2 \{ d_j (1 - h_j)^2 - \theta \}$, so under HC3 ($d_j (1 - h_j)^2 = 1$) every $c_j \ge (1 - \theta)\gamma_j^2 \ge 0$ and $a_- \ge (1 - \theta) s_{\min}^2 \sum_j \gamma_j^2 > 0$ for every $\theta \in (0, 1)$; under HC0–HC2 some c_j can be negative at high leverage and a_- has to be checked. All of this is machine-checked (`finiteN_tail_twoSided_cantelli` and the lemmas `aMinus`, `cantelliQBar`, `aMinus_pos_hc3` in `Ols/FiniteN.lean`).

E DATASETS

The union of assumptions needed (see Table 1) across all methods is: the variance lower bound ($\sigma_i \geq s_{\min}$), a variance upper bound ($\sigma_i \leq s_{\max}$, or $\mathrm{sd}(\varepsilon_i) \leq s_i$ per observation), an error upper bound ($|\varepsilon_i| \leq B_i$), a third moment upper bound ($\mathbb{E}|\varepsilon_i|^3 \leq b_3\sigma_i^3$), a fourth moment upper bound ($\mathbb{E}\varepsilon_i^4 \leq \kappa\sigma_i^4$), and independence throughout. We provide the values assumed for each, per dataset, below. In all cases, the normal approximation assumes $s_{\min} = s_{\max}$, $b_3 = \sqrt{8/\pi}$ and $\kappa = 3$.

`cars` (teaching): Speed (x) of cars and the distances (y) taken to stop ($n = 50$). Standard pedagogical example; testing whether slope is zero. Data from R package `datasets`. *Reported p-value:* from public teaching materials (the r-statistics.co linear regression tutorial and the University of Sheffield course text MAS61004, §4.3). *Assumptions:* independent errors; $|\varepsilon_i| \leq 175$ ft; $\mathrm{sd}(\varepsilon_i) \leq 35$ ft.

`Hubble` (science): Distances (x) and recession velocities (y) of galaxies with Cepheid-calibrated distance ($n = 22$). First empirical discovery of the expansion of the universe (Hubble, 1929) (data re-measured by Freedman et al. (2001)). Example from Wood (2017, §1.1.3), testing whether the universe is 6000 years old. Data from R package `gamair`. *Reported p-value:* from Wood (2017, §1.1.3). *Assumptions:* independent errors; $|\varepsilon_i| \leq 2000$ km/s; $\mathrm{sd}(\varepsilon_i) \leq 500$ km/s.

`Galton` (science): heights (y) of adult children and mid-parent heights (x) ($n = 928$); the dataset by which Galton (1886) discovered regression to the mean; testing whether slope is zero. Data from R package `HistData`. *Reported p-value:* from a public RPubs analysis of this dataset. *Assumptions:* independent errors; $|\varepsilon_i| \leq 18$ in; $\mathrm{sd}(\varepsilon_i) \leq 4.5$ in.

`Boston` (teaching): median home values (y) of Boston census tracts and thirteen covariates (x) including air pollution ($n = 506$); from the willingness-to-pay-for-clean-air study of Harrison & Rubinfeld (1978). Standard pedagogical example (James et al., 2013, §3.6), testing whether the pollution (NOx) coefficient is zero. Data from R package `MASS`. *Reported p-value:* from James et al. (2013, p. 114). *Assumptions:* independent errors; $|\varepsilon_i| \leq 45$ and $\mathrm{sd}(\varepsilon_i) \leq 10$, in thousands of 1970 dollars.

`Mincer` (teaching): log hourly wages (y) and education, experience and tenure (x) ($n = 526$, US CPS 1976); the return-to-education regression of labour economics (Mincer, 1974); standard pedagogical example (Wooldridge, 2019, Example 4.1), testing whether the education coefficient is zero. Data from R package `wooldridge`. *Reported p-value:* from the printed statistic in Wooldridge (2019, Example 4.1). *Assumptions:* independent errors; $|\varepsilon_i| \leq 3.6$ and $\mathrm{sd}(\varepsilon_i) \leq 0.9$, on the log-wage scale.

`Card--Krueger` (policy): changes in fast-food employment (y) and a New-Jersey-versus-Pennsylvania indicator (x) ($n = 384$) after New Jersey's 1992 minimum-wage rise; the natural experiment of Card & Krueger (1994), part of the 2021 Nobel citation; testing whether the difference is zero. Data from David Card's public archive. *Reported p-value:* from the printed statistic in Card & Krueger (1994, Table 4). *Assumptions:* independent errors; $|\varepsilon_i| \leq 70$ and $\mathrm{sd}(\varepsilon_i) \leq 18$ full-time-equivalent employees.

`Reinhart--Rogoff` (policy): real GDP growth (y) of advanced economies and public-debt categories (x) ($n = 1175$ country-years, 1946–2009); the corrected data behind the austerity debate; testing whether the $\geq 90\%$-debt coefficient is zero. Data from the replication files of Herndon et al. (2014). *Reported p-value:* from the printed statistic in Herndon et al. (2014, Table 4). *Assumptions:* independent errors; $|\varepsilon_i| \leq 30$ and $\mathrm{sd}(\varepsilon_i) \leq 6$ percentage points of annual growth.

`streptomycin` (medicine): deaths within six months (y) and a streptomycin-versus-bed-rest indicator (x) ($n = 107$); the first modern randomised controlled trial (Medical Research Council, 1948); testing whether the mortality difference is zero. Data from R package `medicaldata`. *Reported p-value:* from Medical Research Council (1948, p. 771). *Assumptions:* independent errors; $|\varepsilon_i| \leq 1$ and $\mathrm{sd}(\varepsilon_i) \leq \frac{1}{2}$, which hold for any $0/1$ response.

`anorexia` (medicine): weight changes (y) of anorexia patients and a family-therapy-versus-control indicator (x) ($n = 43$) (Hand et al., 1993); the textbook Welch two-sample setting; testing whether the difference is zero. Data from R package `MASS`. *Reported p-value:* from none printed; computed by us. *Assumptions:* independent errors; $|\varepsilon_i| \leq 45$ lb; $\mathrm{sd}(\varepsilon_i) \leq 15$ lb.

`GISTEMP` (science): annual global mean temperature anomalies (y) and year (x) ($n = 56$, 1970–2025); testing whether the warming trend is zero. Data from GISTEMP Team (2026). *Reported p-value:* from none printed; computed by us. *Assumptions:* independent errors; $|\varepsilon_i| \leq 1.0\,^{\circ}\mathrm{C}$; $\mathrm{sd}(\varepsilon_i) \leq 0.25\,^{\circ}\mathrm{C}$.

`Moore` (AI): $\log_2$ transistors per microprocessor (y) and year (x) ($n = 39$, 1971–2021); Moore's law; testing whether the doubling rate is zero. Data from Our World in Data (2026). *Reported p-value:* from none printed; computed by us. *Assumptions:* independent errors; $|\varepsilon_i| \leq 6$ and $\mathrm{sd}(\varepsilon_i) \leq 1.5$ doublings ($\log_2$ scale).

`Chinchilla` (AI): log excess losses (y) of language-model training runs and log model size with log training tokens (x) ($n = 245$), reconstructed from Hoffmann et al. (2022) by Besiroglu et al. (2024); the basis of compute-optimal scaling; testing whether the data exponent is zero. Data from the release of Besiroglu et al. (2024). *Reported p-value:* from none printed; computed by us. *Assumptions:* independent errors; $|\varepsilon_i| \leq 1.0$ and $\mathrm{sd}(\varepsilon_i) \leq 0.2$, on the log excess-loss scale.

`SH0ES` (science): the Riess et al. (2022) distance-calibration system for the Hubble constant ($n = 3492$); testing H_0 against the Planck value (the "Hubble tension"). Data from the public PantheonPlusSH0ES release. *Reported p-value:* from the five-sigma statement of Riess et al. (2022). *Assumptions:* independent errors; per observation, $|\varepsilon_i| \leq 6e_i$ and $\mathrm{sd}(\varepsilon_i) \leq 1.5e_i$, where e_i is the quoted measurement error; the normal approximation is not applicable ($h_{\max} = 0.53$).