

Strength of Evidence: Likelihood and How Much a Result Should Change a Belief

Juan Segurola

3Dresyns by Resyner Technologies S.L.

© 2026 Juan Segurola. All Rights Reserved.

Correspondence: info@3Dresyns.com · www.3Dresyns.com

INF-005 · Scientific Inference and Experimental Reasoning · Version v0.2

Abstract

Scientific evidence is not exhausted by whether a result crosses a significance threshold. Evidence concerns how observations discriminate among competing hypotheses or models, how surprising the observations are under each alternative, and how strongly the data should alter an inferential state given declared assumptions. This article develops a bounded treatment of evidential strength using likelihood and Bayes-factor ideas without turning INF into a general statistics course. It distinguishes evidence in the data from prior odds, posterior belief and long-run error control. It also explains why p values and confidence intervals answer questions different from the likelihood ratio between specified alternatives, and why a small p value does not by itself measure how strongly one scientific explanation is supported over another. The article does not prescribe a universal evidence scale. Its purpose is to force the analyst to state the rival claims, the data model under each, and the quantity being interpreted as evidential weight.

Keywords: strength of evidence; likelihood; likelihood ratio; Bayes factor; p value; scientific inference; competing hypotheses

1. Evidence is comparative

A datum becomes evidential only relative to alternatives. Royall's likelihood paradigm makes this explicit: the same observation can favour one specified hypothesis over another by a likelihood ratio even though the absolute probability of the observation is small under both [1]. Bayesian analysis similarly uses Bayes factors to compare how well data are predicted by competing models, while prior odds enter separately when posterior odds are calculated [2,3].

2. Data evidence, prior information and decision are different objects

- Data evidence concerns how the observed data discriminate among specified hypotheses under stated sampling or generative models.
- Prior information concerns constraints or probabilities assigned before the present data are incorporated.
- Posterior belief combines the data with prior information under a Bayesian model.
- Decision rules add utilities, losses, costs or error-control objectives and therefore cannot be equated with evidential strength alone.

INF-006 owns prior knowledge and base rates. INF-005 therefore keeps the data-evidence question conceptually separate even when a Bayesian calculation later combines them.

3. Why statistical significance is not evidential strength

A p value is defined under a null model and a test statistic; it is not the posterior probability that the null is true and does not compare the null with one specific alternative unless additional structure is introduced. The literature documents persistent misinterpretations of p values, confidence intervals and power [4]. This does not make those tools useless. It means that their interpretation must match the question they answer.

4. Likelihood ratios and Bayes factors

For two specified hypotheses H_1 and H_0 and observed data y , the likelihood ratio $L(H_1|y)/L(H_0|y)$ describes how much more compatible the observed data are with one hypothesis than the other under the models used to generate the likelihood. A Bayes factor integrates or averages over parameter uncertainty according to model-specific priors and compares the marginal predictive support of the models [2,3]. Neither quantity is universal: both depend on how the competing models are specified, and Bayes factors can depend materially on prior distributions within models.

5. Evidence can be weak even when an estimate is precise

Precision within one fitted model is not the same as discrimination among rival models. A narrow confidence or credible interval around a parameter can coexist with structural uncertainty about which model is appropriate. INF-008 owns that structural-versus-parametric distinction. INF-005 therefore treats evidential strength as conditional on the alternatives being compared and does not allow a precise estimate within one structure to stand in for model comparison.

6. No universal evidence threshold

Terms such as “strong”, “moderate” or “decisive” can be useful summaries only when the underlying scale and context are declared. INF does not impose a universal likelihood-ratio or Bayes-factor cutoff. The consequences of a mistaken

inference differ among applications, and a fixed numerical threshold can create false precision about what remains an assumption-dependent judgment. Mayo's severe-testing perspective provides a complementary warning: evidence depends not only on a statistic but also on how capable the procedure was of exposing relevant error [5].

7. Evidential weight and negative results

Failure to reject a null is not equivalent to evidence for the null. A negative result can favour one hypothesis over another only when the competing predictions are explicit and the design has adequate capability. Conversely, a very sensitive experiment can produce evidence against a large-effect hypothesis even when the observed estimate is close to zero. The distinction between non-detection and evidence of absence remains dependent on method capability.

8. Minimum reporting discipline

- Name the competing hypotheses or models.
- State whether the reported quantity is a p value, likelihood ratio, Bayes factor, posterior probability, confidence interval or another inferential object.
- Do not translate one quantity into another by rhetoric.
- State the model assumptions that determine the evidence measure.
- Separate evidence in the data from prior information and from decision costs.
- Avoid universal verbal labels unless the underlying quantitative convention and context are explicit.

References

- [1] Royall, R. *Statistical Evidence: A Likelihood Paradigm*. London: Chapman & Hall, 1997. DOI: 10.1201/9780203738665.
- [2] Goodman, S.N. Toward Evidence-Based Medical Statistics. 2: The Bayes Factor. *Annals of Internal Medicine* 130(12), 1005-1013 (1999). DOI: 10.7326/0003-4819-130-12-199906150-00019.
- [3] Kass, R.E.; Raftery, A.E. Bayes Factors. *Journal of the American Statistical Association* 90(430), 773-795 (1995). DOI: 10.1080/01621459.1995.10476572.
- [4] Greenland, S.; Senn, S.J.; Rothman, K.J.; Carlin, J.B.; Poole, C.; Goodman, S.N.; Altman, D.G. Statistical Tests, P Values, Confidence Intervals, and Power: A Guide to Misinterpretations. *European Journal of Epidemiology* 31, 337-350 (2016). DOI: 10.1007/s10654-016-0149-3.
- [5] Mayo, D.G. *Statistical Inference as Severe Testing: How to Get Beyond the Statistics Wars*. Cambridge University Press, 2018. DOI: 10.1017/9781107286184.

Data and materials statement

No new experimental dataset was generated for this article. The article is a methodological analysis based on the cited literature and the frozen ownership architecture of the INF series.