FROM THE RESEARCH BENCH

What is statistical conclusion validity? A significant result can still be wrong

Statistical conclusion validity asks whether the data and model warrant the stated result. Check effect, uncertainty, power, and testing before deciding.

Sinan Isoglu, MBA (Quantic)

Commercial growth leader, lecturer and doctoral researcher

PUBLISHED
6 September 2026

READING TIME
4 min

WORDS
910

https://isoglu.com/insights/journal/what-is-statistical-conclusion-validity/

CONTENTS

MANAGEMENT SUMMARY

Statistical conclusion validity asks whether the inference from observed data and a chosen model to the stated statistical conclusion is warranted under its assumptions and error controls. It keeps statistical significance separate from effect size, confidence intervals, power, type I and type II error, multiple testing, practical importance, causal identification, and external validity. Cohen's critique of mechanical significance testing and Ioannidis's conditional framework for prior plausibility, power, bias, and multiplicity provide the bounded research anchors. This article builds a synthetic inference table with effect estimates, intervals, testing families, and permitted wording. The table and decision rules are author synthesis. They do not provide a universal p-value threshold, false-positive rate, or verdict about a current company result.

Keywords: Statistical Conclusion Validity · Statistical Power · Effect Size · Confidence Interval · Multiple Testing · False Positive

A result can be statistically significant and still be too small to matter. A result can be non-significant because the study contains too little information to distinguish a useful effect from zero. A table can report an exact p value while hiding that twenty outcomes and six model specifications were tried first.

Statistical conclusion validity asks whether the data and model warrant the statistical conclusion as written. It is a question about the inference path, not a synonym for significance.

The common-method-bias article owns the gap between a target and its indicator. The measurement-invariance article owns transfer to a named target. This page owns the narrower move from an observed analysis to a statistical sentence.

What does statistical conclusion validity mean?

Keep these objects separate:

Table 1 What does statistical conclusion validity mean?

Object	Question	What it cannot carry alone
Estimate	How large is the observed difference or association?	Importance or causality
Confidence interval	Which values remain compatible with the stated procedure?	Probability that a particular value is true
p value	How discordant are these data under the specified null model?	Probability that the null is true
Power or information	Could the design detect the declared effect under its assumptions?	Proof that an undetected effect is absent
Testing family	How many hypotheses, outcomes, or specifications were examined?	A universal correction independent of the analysis plan
Conclusion	What sentence does the evidence permit?	Claims outside the population, model, or outcome

Table from this essay. Sources and interpretation are given in the article.

Cohen argued against mechanical null-hypothesis significance testing and recommended attention to effect sizes, confidence intervals, graphical methods, and replication. His warning is simple but often lost: a threshold does not tell a reader whether the effect is consequential.

Why can a positive result still be fragile?

Ioannidis's framework shows that the probability a positive research finding is true depends on factors including prior plausibility, power, bias, and the number of relationships examined. Low power, small prior effects, flexible analysis, and multiple testing can make positive findings less reliable under the model. The framework is conditional. It is not a false-positive rate for every discipline, company, or metric.

That distinction produces four separate review questions:

- Is the estimate large enough to matter for the decision?
- Is the interval narrow enough to distinguish the relevant alternatives?
- Was the testing or specification process declared and controlled?
- Does the conclusion stay inside the studied population, outcome, model, and time window?

A significant result can fail any of the last three. A non-significant result can fail because the interval is wide or the study had little information. Neither label should replace the evidence.

What does an inference review look like?

The six rows below are synthetic. They illustrate conclusion permissions, not results from a company or study.

Figure 1 The synthetic statistical-inference review

ID	Reported result	Uncertainty and information	Testing or model context	Conclusion permitted	Still not established
T-01	Estimate +2.0 points, p = 0.01	95% interval +0.5 to +3.5; adequate planned power	One declared outcome and model	Evidence of a positive association in the studied sample	Practical importance or causality
T-02	Estimate +0.2 points, p < 0.001	Narrow interval +0.15 to +0.25	Large sample; outcome scale is small	Precisely estimated small difference	Material business value
T-03	Estimate +8 points, p = 0.08	Wide interval -1 to +17	One outcome; limited information	Data are insufficient to rule in or out the declared effect	No-effect conclusion
T-04	Estimate +5 points, p = 0.03	Interval +0.4 to +9.6	Twelve outcomes tested; no family rule stated	At most a flagged result pending multiplicity review	Confirmed discovery
T-05	Estimate +10 points, p = 0.02	Interval +2 to +18	Model changed after inspecting outcome	Conditional post hoc result	Ex ante test interpretation
T-06	Estimate +6 points, p = 0.01	Interval +2 to +10	Studied in one segment and 14-day window	Positive result in that segment and window	Transfer to all customers or 90-day renewal

Author's synthetic inference review grounded in Cohen (1994) and Ioannidis (2005). Estimates, intervals, testing histories, and permitted conclusions are illustrative.

T-02 shows why a tiny interval can still surround a trivial decision effect. T-03 shows why non-significance is not evidence of no effect. T-04 and T-05 show why a correct calculation does not erase an undeclared testing family or post hoc specification. T-06 shows that statistical conclusion validity remains bounded by population, outcome, and horizon.

What are type I and type II error?

Under a declared testing procedure, a type I error is rejecting a true null hypothesis. A type II error is failing to reject a null when the declared alternative is true. They are not interchangeable, and a p value does not report either risk by itself. Their interpretation depends on the null, the alternative, the design, the testing family, the decision threshold, and the information available.

Multiple testing changes the question because a team may have examined several outcomes, subgroups, models, or time windows. The solution is not always one formula. It is to preserve the family, declare the control procedure, and report what was selected before the result was known.

How should a team write the conclusion?

1. State the unit, population, outcome, comparison, and observation window.
2. Report estimate and uncertainty in the outcome's units.
3. State the testing or estimation procedure and any multiplicity control.
4. Record power or information relative to the decision-relevant effect.
5. Write whether the result is descriptive, associational, causal, or predictive.
6. Keep practical importance and external transfer as separate questions.

Use insufficient evidence for the declared effect under this design when the interval or design cannot support a stronger conclusion. Use evidence consistent with a positive association in this population and window when that is what the analysis supports. Avoid "proved," "no effect," and "works everywhere" unless the design actually warrants those words.

A significant result is one field in an inference record. Statistical conclusion validity is the discipline of keeping every other field that limits the sentence visible.

Cohen, J. (1994). The earth is round (p < .05). American Psychologist, 49(12), 997-1003. https://doi.org/10.1037/0003-066X.49.12.997

Ioannidis, J. P. A. (2005). Why most published research findings are false. PLoS Medicine, 2(8), e124. https://doi.org/10.1371/journal.pmed.0020124

END OF ESSAY

HOW TO CITE THIS ESSAY

Isoglu, S. (2026, September 6). What is statistical conclusion validity? A significant result can still be wrong. Sinan Isoglu. https://isoglu.com/insights/journal/what-is-statistical-conclusion-validity/

KEEP READING

From the research bench

More advertising observations do not guarantee better decisions

https://isoglu.com/insights/journal/more-advertising-observations-do-not-guarantee-better-decisions/

From the research bench

What is external validity? A result needs a transfer boundary

https://isoglu.com/insights/journal/what-is-external-validity/

From the research bench

What is data lineage? The transformations behind a result

https://isoglu.com/insights/journal/what-is-data-lineage/

Sinan Isoglu, MBA (Quantic)
Commercial growth leader, lecturer and doctoral researcher

Online version: https://isoglu.com/insights/journal/what-is-statistical-conclusion-validity/