AUS DER FORSCHUNG

Was ist statistische Schlussfolgerungsvalidität? Ein signifikanter Befund kann falsch sein

Statistische Schlussfolgerungsvalidität fragt, ob Daten und Modell den Befund tragen. Prüfen Sie Effekt, Unsicherheit, Teststärke und Mehrfachtests.

Sinan Isoglu, MBA (Quantic)
Führungskraft für Umsatzwachstum, Dozent und Doktorand

VERÖFFENTLICHT
6. September 2026

LESEZEIT
4 Min.

WÖRTER
847

https://isoglu.com/de/insights/journal/was-ist-statistische-schlussfolgerungsvaliditaet/

INHALT

MANAGEMENT SUMMARY

Statistische Schlussfolgerungsvalidität fragt, ob die Schlussfolgerung aus beobachteten Daten und einem gewählten Modell unter seinen Annahmen und Fehlerkontrollen gerechtfertigt ist. Sie trennt statistische Signifikanz von Effektgröße, Konfidenzintervallen, Teststärke, Fehlern erster und zweiter Art, multiplem Testen, praktischer Bedeutung, Kausalidentifikation und externer Validität. Cohens Kritik an mechanischem Signifikanztesten und Ioannidis' bedingter Rahmen zu Vorplausibilität, Teststärke, Bias und Mehrfachtestung liefern die begrenzten Forschungsanker. Dieser Beitrag baut eine synthetische Inferenz-Tabelle mit Effektschätzungen, Intervallen, Testfamilien und zulässiger Formulierung. Tabelle und Entscheidungsregeln sind eigene Synthese. Sie liefern weder einen universellen p-Wert-Schwellenwert noch eine Falsch-Positiv-Quote oder ein Urteil über ein aktuelles Unternehmensergebnis.

Schlagwörter: Statistische Schlussfolgerungsvalidität · Teststärke · Effektgröße · Konfidenzintervall · Multiples Testen · Falsch-positiver Befund

Ein Ergebnis kann statistisch signifikant und trotzdem zu klein für die Entscheidung sein. Ein nicht signifikantes Ergebnis kann entstehen, weil die Studie zu wenig Information enthält, um einen nützlichen Effekt von null zu unterscheiden. Eine Tabelle kann einen exakten p-Wert zeigen und verbergen, dass vorher zwanzig Ergebnisse und sechs Modellspezifikationen geprüft wurden.

Statistische Schlussfolgerungsvalidität fragt, ob Daten und Modell die statistische Aussage in ihrer Form tragen. Sie ist kein Synonym für Signifikanz.

Der Beitrag zu Common-Method-Bias besitzt die Lücke zwischen Ziel und Indikator. Der Beitrag zur Messinvarianz besitzt den Transfer auf ein benanntes Ziel. Diese Seite besitzt den engeren Schritt von einer beobachteten Analyse zu einem statistischen Satz.

Was bedeutet statistische Schlussfolgerungsvalidität?

Halten Sie die Objekte getrennt:

Tabelle 1 Was bedeutet statistische Schlussfolgerungsvalidität?

Objekt	Frage	Was es allein nicht tragen kann
Schätzung	Wie groß ist der beobachtete Unterschied oder Zusammenhang?	Bedeutung oder Kausalität
Konfidenzintervall	Welche Werte sind mit dem Verfahren vereinbar?	Wahrscheinlichkeit, dass ein bestimmter Wert wahr ist
p-Wert	Wie unvereinbar sind die Daten mit dem erklärten Nullmodell?	Wahrscheinlichkeit, dass die Null wahr ist
Teststärke oder Information	Könnte das Design den erklärten Effekt erkennen?	Beweis für das Fehlen eines nicht erkannten Effekts
Testfamilie	Wie viele Hypothesen, Ergebnisse oder Spezifikationen wurden geprüft?	Universelle Korrektur unabhängig vom Analyseplan
Schlussfolgerung	Welcher Satz ist durch die Evidenz erlaubt?	Aussagen außerhalb von Population, Modell oder Ergebnis

Tabelle aus diesem Essay. Quellen und Einordnung stehen im Beitrag.

Cohen wandte sich gegen mechanisches Testen von Nullhypothesen und empfahl Aufmerksamkeit für Effektgrößen, Konfidenzintervalle, grafische Verfahren und Replikation. Die Warnung ist einfach und wird trotzdem oft verloren: Ein Schwellenwert sagt nicht, ob ein Effekt für die Entscheidung relevant ist.

Warum kann ein positiver Befund fragil sein?

Ioannidis' Rahmen zeigt, dass die Wahrscheinlichkeit eines wahren positiven Forschungsbefunds unter anderem von Vorplausibilität, Teststärke, Bias und der Zahl geprüfter Beziehungen abhängt. Geringe Teststärke, kleine Vorwirkungen, flexible Analyse und multiples Testen können den Anteil falscher positiver Befunde im Modell erhöhen. Der Rahmen ist bedingt. Er ist keine Falsch-Positiv-Quote für jedes Fach, Unternehmen oder jede Kennzahl.

Daraus entstehen vier getrennte Prüfungsfragen:

- Ist die Schätzung groß genug, um für die Entscheidung zu zählen?
- Ist das Intervall eng genug, um die relevanten Alternativen zu unterscheiden?
- War der Test- oder Spezifikationsprozess erklärt und kontrolliert?
- Bleibt die Schlussfolgerung in Population, Ergebnis, Modell und Zeitfenster?

Ein signifikanter Befund kann an jeder der letzten drei Fragen scheitern. Ein nicht signifikanter Befund kann scheitern, weil das Intervall breit oder die Information gering ist. Keines der beiden Labels ersetzt die Evidenz.

Wie sieht ein Inferenz-Review aus?

Die sechs Zeilen sind synthetisch. Sie veranschaulichen zulässige Schlussfolgerungen, keine Ergebnisse eines Unternehmens oder einer Studie.

Abbildung 1 Der synthetische statistische Inferenz-Review

ID	Gemeldetes Ergebnis	Unsicherheit und Information	Test- oder Modellkontext	Zulässige Schlussfolgerung	Noch nicht festgestellt
T-01	Schätzung +2,0 Punkte, p = 0,01	95-%-Intervall +0,5 bis +3,5; geplante Teststärke ausreichend	Ein erklärtes Ergebnis und Modell	Evidenz für positiven Zusammenhang in der Stichprobe	Praktische Bedeutung oder Kausalität
T-02	Schätzung +0,2 Punkte, p < 0,001	Enges Intervall +0,15 bis +0,25	Große Stichprobe; Skala ist klein	Präzise geschätzter kleiner Unterschied	Wesentlicher Geschäftswert
T-03	Schätzung +8 Punkte, p = 0,08	Breites Intervall -1 bis +17	Ein Ergebnis; begrenzte Information	Daten reichen für oder gegen Effekt nicht aus	Aussage „kein Effekt“
T-04	Schätzung +5 Punkte, p = 0,03	Intervall +0,4 bis +9,6	Zwölf Ergebnisse getestet; keine Testfamilienregel	Allenfalls markierter Befund bis zur Mehrfachtestprüfung	Bestätigte Entdeckung
T-05	Schätzung +10 Punkte, p = 0,02	Intervall +2 bis +18	Modell nach Sichtung des Ergebnisses geändert	Bedingter nachträglicher Befund	Vorabinterpretation
T-06	Schätzung +6 Punkte, p = 0,01	Intervall +2 bis +10	Eine Segment- und 14-Tage-Analyse	Positiver Befund in Segment und Zeitraum	Transfer auf alle Kunden oder 90-Tage-Verlängerung

Synthetischer Inferenz-Review des Autors auf Grundlage von Cohen (1994) und Ioannidis (2005). Schätzungen, Intervalle, Testverläufe und zulässige Schlussfolgerungen sind illustrativ.

T-02 zeigt, warum ein enges Intervall trotzdem einen trivialen Entscheidungseffekt umschließen kann. T-03 zeigt, warum Nichtsignifikanz keine Evidenz für keinen Effekt ist. T-04 und T-05 zeigen, warum eine korrekte Berechnung eine unerklärte Testfamilie oder nachträgliche Spezifikation nicht beseitigt. T-06 zeigt, dass statistische Schlussfolgerungsvalidität durch Population, Ergebnis und Horizont begrenzt bleibt.

Was sind Fehler erster und zweiter Art?

Unter einem erklärten Testverfahren ist ein Fehler erster Art die Ablehnung einer wahren Nullhypothese. Ein Fehler zweiter Art ist die Nichtablehnung einer Nullhypothese, wenn die erklärte Alternative gilt. Beide sind nicht austauschbar, und ein p-Wert berichtet keines dieser Risiken allein. Die Auslegung hängt von Null, Alternative, Design, Testfamilie, Schwelle und verfügbarer Information ab.

Multiples Testen verändert die Frage, weil mehrere Ergebnisse, Untergruppen, Modelle oder Zeitfenster geprüft worden sein können. Die Lösung ist nicht immer eine Formel. Sie besteht darin, die Familie sichtbar zu halten, das Kontrollverfahren zu benennen und zu berichten, was vor dem Ergebnis ausgewählt wurde.

Wie formuliert ein Team die Schlussfolgerung?

1. Einheit, Population, Ergebnis, Vergleich und Beobachtungsfenster angeben.
2. Schätzung und Unsicherheit in den Ergebnis-Einheiten berichten.
3. Test- oder Schätzverfahren und Mehrfachtestkontrolle nennen.
4. Teststärke oder Information relativ zum entscheidungsrelevanten Effekt festhalten.
5. Schreiben, ob der Befund beschreibend, assoziativ, kausal oder prognostisch ist.
6. Praktische Bedeutung und externen Transfer als eigene Fragen behandeln.

Verwenden Sie unter diesem Design nicht ausreichende Evidenz für den erklärten Effekt, wenn Intervall oder Design keine stärkere Aussage tragen. Verwenden Sie Evidenz für einen positiven Zusammenhang in dieser Population und diesem Zeitraum, wenn das die Analyse erlaubt. Vermeiden Sie „bewiesen“, „kein Effekt“ und „funktioniert überall“, wenn das Design diese Wörter nicht trägt.

Ein signifikanter Befund ist ein Feld im Inferenzdatensatz. Statistische Schlussfolgerungsvalidität heißt, jedes begrenzende Feld für den Satz sichtbar zu halten.

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

ENDE DES BEITRAGS

ZITIERWEISE

Isoglu, S. (2026, 6. September). Was ist statistische Schlussfolgerungsvalidität? Ein signifikanter Befund kann falsch sein. Sinan Isoglu. https://isoglu.com/de/insights/journal/was-ist-statistische-schlussfolgerungsvaliditaet/

WEITERLESEN

Aus der Forschung
Was ist Reproduzierbarkeit? Kann eine andere Person das Ergebnis rekonstruieren?
https://isoglu.com/de/insights/journal/was-ist-reproduzierbarkeit/

Aus der Forschung
Gemischte Reaktionen sind keine Markenpolarisierung
https://isoglu.com/de/insights/journal/gemischte-reaktionen-sind-keine-markenpolarisierung/

Aus der Forschung
Was ist Claim-Entailment? Stützt die Quelle genau diesen Satz?
https://isoglu.com/de/insights/journal/was-ist-claim-entailment/

Sinan Isoglu, MBA (Quantic)
Führungskraft für Umsatzwachstum, Dozent und Doktorand