

Aus der Forschung

Was ist kausale Inferenz?

Kausale Inferenz schätzt, was eine Änderung bewirkt hat, aus Ergebnissen, die nie alle beobachtet werden. Die Antwort tragen Design und Annahmen.

Autor	Veröffentlicht	Lesezeit	Wörter
Sinan Isoglu, MBA (Quantic)	18. Oktober 2026	22 Min.	4.628
Führungskraft für Umsatzwachstum, Dozent und Doktorand			

Management Summary Kausale Inferenz ist die Schätzung dessen, was eine Behandlung verändert hat: der Unterschied zwischen dem Ergebnis, das eine Einheit hatte, und dem Ergebnis, das sie unter der Alternative gehabt hätte. Dieser Konzeptleitfaden stellt das Potential-Outcomes-Modell vor (Rubin; Rosenbaum und Rubin), erklärt, warum ein roher Vergleich von behandelten und unbehandelten Gruppen verzerrt ist, und vergleicht fünf Wege zu einer Schätzung: randomisierte Experimente, Adjustierung für beobachtete Kovariaten, natürliche Experimente, Fixed Effects und Difference-in-Differences sowie Instrumentvariablen, jeweils mit ihren Annahmen. Er behandelt Estimands, Kausaldiagramme (Pearl) und ein einseitiges Identifikationsblatt für eine Kausalaussage; die Grenzen stammen aus Gangls Übersicht und aus denen, die Pearl für seine eigene Methode nennt.

Inhalt	§ 1	Was ist kausale Inferenz?	3
	§ 2	Warum ist Korrelation keine Kausalität?	4
	§ 3	Warum steht das randomisierte Experiment an erster Stelle?	6
	§ 4	Wie können Beobachtungsdaten eine Kausalaussage stützen?	7
	§ 5	Was leisten natürliche Experimente und wiederholte Beobachtungen?	9
	Abbildung 1	Fünf Wege zu einer kausalen Schätzung und was jeder annehmen muss	11
	§ 6	Welche Annahmen begleiten eine Kausalaussage?	14
	§ 7	Welcher Effekt wird geschätzt?	15
	§ 8	Wie unterscheiden sich potenzielle Ergebnisse und Kausaldiagramme?	16
	§ 9	Wie hält man eine Kausalaussage fest?	17
	Abbildung 2	Ein Identifikationsblatt	18

§ 10	Wo endet die Evidenz zur kausalen Inferenz?	20
	Referenzen	21

Was ist kausale Inferenz?

§ 1 von 10

Kausale Inferenz ist die Schätzung dessen, was eine Behandlung verändert hat. Gangl (2010)¹ beschreibt sie im Anschluss an Heckman als den Versuch, kausale Parameter sozialer Prozesse aus empirisch beobachteten Daten zurückzugewinnen („the attempt to recover causal parameters that describe social processes of interest from empirically observed data“). Die Behandlung kann ein Programm, ein Preis, eine Kampagne oder ein Ereignis sein; der kausale Effekt ist der Unterschied zwischen dem Ergebnis, das eine Einheit hatte, und dem Ergebnis, das sie unter der Alternative gehabt hätte.

¹ Gangl, M. (2010). Causal inference in sociological research. *Annual Review of Sociology*, 36, 21-47.

Rubin (1974)⁴ gibt die Alltagsfassung. Der kausale Effekt einer Behandlung gegenüber einer anderen, für eine bestimmte Einheit und ein Zeitintervall, ist der Unterschied zwischen dem, was unter der einen und unter der anderen geschehen wäre: Hätte ich vor einer Stunde zwei Aspirin statt eines Glases Wasser genommen, wären meine Kopfschmerzen jetzt weg („If an hour ago I had taken two aspirins instead of just a glass of water, my headache would now be gone“). Die Schwierigkeit steckt in der Definition selbst: Nie lassen sich beide Ergebnisse zugleich beobachten („we can never observe both $y(E)$ and $y(C)$ “), denn niemand kann in der Zeit zurückgehen und derselben Einheit die andere Behandlung geben. Rosenbaum und Rubin (1983)³ ziehen die Folgerung. Weil jede Einheit nur eine Behandlung erhält („Since each unit receives only one treatment“), fehlt immer eines der beiden Ergebnisse, und die Schätzung kausaler Effekte ist in diesem Sinn ein Problem fehlender Daten („In a sense, estimating the causal effects of treatments is a missing data problem“).

³ Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1), 41-55.

⁴ Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66(5), 688-701.

Nach Gangl (2010)¹ ist das kontrafaktische Modell, auch Rubin- oder Potential-Outcomes-Modell genannt, in den letzten drei Jahrzehnten zum konzeptionellen Standardwerkzeug für die Definition kausaler Effekte und das Identifikationsproblem geworden („the counterfactual, Rubin, or potential outcomes model of causality has, over the past three decades, become the standard conceptual tool“). Die Effekte je Einheit, die es definiert, sind grundsätzlich nicht beobachtbar und

müssen unter zusätzlichen Identifikationsannahmen aus beobachtbaren Daten geschätzt werden („are therefore unobservable in principle but need to be estimated from empirically observable data under auxiliary identifying assumptions“).

In diesem Leitfaden heißt kausale Inferenz: für eine benannte Population den Unterschied schätzen zwischen den Ergebnissen, die sie hatte, und denen, die sie unter einer benannten Alternative gehabt hätte, und die Annahmen nennen, unter denen ein beobachteter Vergleich für die nie beobachteten Ergebnisse eintreten kann. Die Daten liefern den Vergleich; kausal machen ihn erst das Design und seine Annahmen.

Mehrere Leitfäden auf dieser Website behandeln einen Teil dieses Problems. Der Leitfaden zur Inkrementalität wendet es auf den Umsatz an, den eine kommerzielle Maßnahme zusätzlich gebracht hat, und der Leitfaden zum A/B-Test behandelt das randomisierte Experiment. Die Leitfäden zu Selektionsbias, Interferenz und parallelen Trends greifen je eine Annahme auf, die scheitern kann, und der Leitfaden zur externen Validität fragt, ob eine Schätzung auf ein anderes Ziel übertragbar ist. Diese Seite legt den gemeinsamen Rahmen dar.

Warum ist Korrelation keine Kausalität?

§ 2 von 10

Weil sich die behandelten Einheiten meist schon vor der Behandlung von den anderen unterscheiden. Gangl (2010)¹ zeigt, dass der direkte Vergleich der Durchschnittsergebnisse beider Gruppen, mitunter naiver Schätzer genannt, den durchschnittlichen Behandlungseffekt verzerrt schätzt („the straightforward comparison of average outcomes in the two groups (sometimes called the naive estimator) will be a biased estimator of the ATE parameter“). Die Verzerrung hat zwei Teile: Die Gruppen könnten auch ohne die Behandlung unterschiedliche Ausgangsergebnisse gehabt haben („might have seen different baseline outcomes even absent the intervention“), und der Effekt kann sich zwischen denen, die die Behandlung erhielten, und den anderen unterscheiden. Er nennt Selektionsbias oder Heterogenität sowie Endogenität oder Selbstselektionsbias als die zwei Hauptquellen der Verzerrung („selection bias, or heterogeneity, and endogeneity, or self-selection bias, as the two principal sources of bias when recovering any causal parameter from empirical data“). Sein Selektionsbias ist ein Unterschied der Ausgangsergebnisse zwischen den Gruppen und kommt dem nahe, was Pearl weiter unten Confounding nennt; der Leitfaden zum Selektionsbias auf dieser Website verwendet den Begriff für Verzerrungen durch die Regel, nach der Einheiten in die Daten gelangen.

Pearl (1995)² beginnt mit einem Beispiel von Cochran: einer Studie zu Bodenbegasungsmitteln und Haferertrag, in der die Landwirte selbst entscheiden, welche Parzellen behandelt werden („farmers insist on deciding for themselves which plots are to be fumigated“). Ihre Wahl hängt vom Älchenbestand des Vorjahres ab, der eng mit dem diesjährigen zusammenhängt. Damit liegt ein klassischer Fall von Confounding vor, also einer Verzerrung durch eine gemeinsame Ursache von Behandlung und Ergebnis, und sie stört die Beurteilung des Effekts unabhängig von der Stichprobengröße („Thus we have a classical case of confounding bias, which interferes with the assessment of treatment effects, regardless of sample size.“). Mehr Daten aus demselben Vergleich beseitigen die Verzerrung nicht.

² Pearl, J. (1995). Causal diagrams for empirical research. *Biometrika*, 82(4), 669-688.

Dieselbe Struktur findet sich in kommerziellen Daten, so liest es dieser Leitfaden. Accounts, die eine Funktion übernehmen, an einem Schulungsprogramm teilnehmen oder einen Rabatt erhalten, werden oft aus Gründen ausgewählt oder wählen sich selbst aus, die mit dem Ergebnis zusammenhängen: Sie wuchsen schon, waren schon aktiv oder schon gefährdet. Ein Dashboard, das ihren Umsatz mit dem aller anderen vergleicht, misst diese Auswahl zusammen mit einem etwaigen Effekt.

Gangl (2010)¹ verwirft die Regression deshalb nicht: Korrelation sei zwar keine Kausalität, doch Regressionsmodelle identifizierten den kausalen Effekt, wenn Behandlung und Ergebnisse nicht konfundiert sind („whereas correlation clearly is not causation, regression models do identify the causal effect of interest if treatment status and outcomes are unconfounded“). Kausal wird eine Schätzung durch das Design und die Annahmen dahinter, nicht durch die Methode, die sie berechnet, auch wenn die Methode bestimmen kann, welchen Durchschnitt sie ausweist.

Warum steht das randomisierte Experiment an erster Stelle?

§ 3 von 10

Weil eine Zuteilung per Zufall die Verbindung kappt zwischen der Frage, wer behandelt wird, und der Frage, wie es ihm sonst ergangen wäre. Randomisierte Experimente beseitigen die Verzerrung („Randomized experiments in particular eliminate bias“), schreibt Gangl (2010)¹, weil sie eine aktive experimentelle Manipulation der Behandlung mit zufälliger Zuteilung verbinden, die die erwarteten Ergebnisse zwischen den Bedingungen ausgleicht. Rubin (1974)⁴ folgert, dass randomisiert werden sollte, wo immer es möglich ist, dass aber die Nutzung sorgfältig kontrollierter, nicht randomisierter Daten zur Schätzung kausaler Effekte in vielen Fällen ein vernünftiges und notwendiges Verfahren ist („that randomization should be employed whenever possible but that the use of carefully controlled nonrandomized data to estimate causal effects is a reasonable and necessary procedure in many cases“).

Rubins Vergleich ist vorsichtig. Für nicht ausdrücklich kontrollierte Variablen steht der Forschende mit einer randomisierten Studie besser da als mit einer nicht randomisierten („a randomized study leaves the investigator in a more comfortable position than does a nonrandomized study“). Drei Punkte gelten für beide Arten von Studien. Jede Adjustierung hängt zum Teil davon ab, ob das Modell dahinter stimmt. Man kann nie wissen, ob alle kausal relevanten Vorabvariablen kontrolliert sind, die sich zwischen den Gruppen systematisch unterscheiden („We can never know that all causally relevant prior variables that systematically differ in the E and C trials have been controlled“). Und irrelevante Vorabvariablen müssen auch dann unberücksichtigt bleiben, wenn sie sich zwischen den Gruppen unterscheiden; sonst lässt sich durch langes Suchen nach den passenden jede gewünschte Schätzung erzeugen („we can obtain any estimate desired“).

Ein Experiment ist kein Freibrief. Auch bei vollständiger Randomisierung, so Gangl (2010)¹, braucht die Identifikation die Stable Unit Treatment Value Assumption (SUTVA), nach der die Ergebnisse einer Einheit von der Zuteilung nur über ihre eigene Behandlung abhängen. Sie schließt Hawthorne- und John-Henry-Effekte aus („thus ruling out Hawthorne or John Henry effects“), ebenso soziale Interaktionen, Informationsdiffusion und Normbildung („social interactions, information diffusion, norm formation“) und andere übergreifende Effekte. Er warnt außerdem, dass sich experimentelle Schätzungen schwer verallgemeinern lassen können, wenn

Studienpopulation oder Behandlung nicht für einen größeren Kontext stehen („may be difficult to generalize if either the study population or the administered treatments cannot be considered representative of some larger context“). Rubin (1974)⁴ benennt den Zielkonflikt offen: Nicht randomisierte Studien haben meist repräsentativere Fälle, weil Experimente oft unter eingeschränkten Bedingungen laufen („Typically, nonrandomized studies have more representative trials than experiments since these are often conducted in constrained environments.“).

Wie können Beobachtungsdaten eine Kausalaussage stützen?

§ 4 von 10

Nur unter einer Annahme, die die Daten nicht beweisen können. Gangl (2010)¹ formuliert sie als bedingte Unabhängigkeit: Die beobachteten Kovariaten bilden eine ausreichende Menge gemeinsamer Ursachen von Behandlung und Ergebnis, sodass die Behandlung gegeben die Kovariaten so gut wie zufällig zugeteilt ist („so that, conditional on covariates, variation in D is as good as randomly assigned“). Rosenbaum und Rubin (1983) geben die stärkere Form, die starke Ignorierbarkeit der Zuteilung: Gegeben die Kovariaten sind die Ergebnisse unabhängig von der Zuteilung, und jede Einheit hätte jede Behandlung erhalten können. In einem randomisierten Versuch mit sauber erhobenen Daten ist der erste Teil bekanntermaßen erfüllt. In einer nicht randomisierten Studie ist das in der Regel nicht bekannt („This condition is usually not known to hold in a nonrandomized experiment.“).

Der Propensity Score ist ihr Werkzeug, um diese Annahme zu nutzen. Er ist die bedingte Wahrscheinlichkeit, einer bestimmten Behandlung zugeteilt zu werden, gegeben die beobachteten Kovariaten („The propensity score is the conditional probability of assignment to a particular treatment given a vector of observed covariates“), und die Adjustierung für ihn genügt, um die Verzerrung durch alle beobachteten Kovariaten zu beseitigen („adjustment for the scalar propensity score is sufficient to remove bias due to all observed covariates“). Im Wort beobachtet liegt die Grenze: Der Score gleicht aus, was gemessen wurde, und nichts sonst. Außerhalb eines randomisierten Versuchs ist die Funktion fast immer unbekannt („the propensity score function is almost always unknown“) und kann geschätzt werden, etwa mit einem Logit-Modell.

Welche Kovariaten in die Adjustierung gehören, ist eine Frage der Theorie, nicht der Modellgüte. Gangl (2010) schreibt, die Annahme bedingter Unabhängigkeit sei keine primär statistische, sondern eine

grundlegend theoretische Aussage über die gemeinsamen Ursachen von Behandlung und Ergebnis („is not a primarily statistical, but rather a fundamentally theoretical statement about the joint causes X of D and Y“). Gebraucht wird nicht jede gemeinsame Ursache von Behandlung und Ergebnis, sondern eine minimal ausreichende Menge von Kovariaten („a minimally sufficient set of covariates“), die die Abhängigkeit zwischen beiden aufhebt. Pearl (1995)² liefert dafür einen grafischen Test, das Back-Door-Kriterium, und hält fest, dass es der Ignorierbarkeitsbedingung von Rosenbaum und Rubin³ entspricht („is equivalent to the ignorability condition of Rosenbaum & Rubin (1983)“).

Die falsche Variable kann schaden. Gangl (2010)¹ warnt, dass die Kontrolle einer Variablen die kausale Deutung gefährden kann („put the causal interpretation in jeopardy“), wenn sie ein Mechanismus ist, über den die Behandlung auf das Ergebnis wirkt („is an intervening mechanism that transmits the effect of D on Y“), denn nur der Gesamteffekt der Behandlung auf das Ergebnis hat einen klaren kausalen Gehalt („it is only the total effect of D on Y that has straightforward causal content“). Er warnt im Anschluss an Pearl auch davor, auf Collider-Variablen zu konditionieren („conditioning on collider variables“), die meist einen unbeobachteten Faktor vertreten sollen: Das kann eine Abhängigkeit zwischen Behandlung und Ergebnis erzeugen, statt sie zu verringern („may inadvertently induce rather than alleviate dependency between treatment status and outcomes“). Kommerziell gelesen, und das ist die Lesart dieses Leitfadens: Wer prüft, ob ein Onboarding-Programm die Verlängerung erhöht hat, sollte nicht für die spätere Produktnutzung kontrollieren, wenn das Programm gerade über die Nutzung wirkt.

Was leisten natürliche Experimente und wiederholte Beobachtungen?

§ 5 von 10

Natürliche Experimente liefern Variation, die die Einheiten nicht selbst gewählt haben; wiederholte Beobachtungen machen jede Einheit zu ihrem eigenen Vergleich. Natürliche Experimente nutzen nach Gangl (2010)¹ alle ein äußeres Ereignis oder eine institutionelle Bedingung, die exogene Variation im untersuchten Prozess erzeugt („all share the idea of exploiting some external event or institutional condition that creates exogenous variation in the social process of interest“): einen Schock oder die Änderung eines Programms oder, in einem Regressions-Diskontinuitäts-Design, eine bekannte institutionelle Regel, die eine Behandlung gewährt oder verweigert („institutional rule to assign or withhold treatment that creates the relevant variation“). Die Kovariaten müssen trotzdem andere gleichzeitige Ursachen von Veränderung und Unterschiede zwischen den verglichenen Populationen abdecken.

Bei wiederholten Beobachtungen derselben Einheiten lassen sich Fixed-Effects- und Difference-in-Differences-Schätzer einsetzen, die zeitkonstante unbeobachtete Unterschiede herausrechnen („difference-in-differences estimators may again be employed to condition on time-invariant unobservables“) und so die Annahme bedingter Unabhängigkeit abschwächen. Der Preis sind zwei Annahmen. Die verbleibende Annahme bedingter Unabhängigkeit ist nach Gangl immer noch stark („is still a strong one“): Sie schließt Einheiten aus, die die Behandlung wählen, weil sie sich zu Recht einen Nutzen davon versprechen. Und der Fixed-Effects-Schätzer beruht auf gemeinsamen (parallelen) Zeittrends der Gruppen („rests on the assumption of common (parallel) time trends across groups“), dem Thema des Leitfadens zu parallelen Trends, und bei heterogenen Effekten haben seine Schätzungen womöglich wenig externe Validität („FE estimates may have little external validity“), weil nur Einheiten beitragen, deren Behandlung sich im Zeitfenster geändert hat.

Eine Instrumentvariable ist eine Variable, die die Behandlung vorhersagt, das Ergebnis aber nicht („that predicts treatment status D, but not outcomes“), außer über die Behandlung. Gangl (2010)¹ sagt deutlich, was sie leisten kann: Sie identifiziert den durchschnittlichen Behandlungseffekt nicht verlässlich („IV does not reliably identify the average treatment effect of D unless the“), es sei denn, der Effekt ist für alle Einheiten gleich. Unter einer Monotoniebedingung identifiziert sie den lokalen durchschnittlichen Behandlungseffekt für die Einheiten, die die Behandlung wegen des Instruments annahmen („those units i

that complied with treatment D because of the instrument Z“). Weil dieser lokale Effekt vom jeweiligen Instrument abhängt („LATE is defined by the specific instrument Z available to the researcher“), können zwei gültige Instrumente zwei verschiedene Antworten geben, und schwache Instrumente untergraben die Konsistenz und Effizienz des Schätzers („reliance on weak instruments undermines the statistical consistency and efficiency of the estimator“).

Abbildung 1 Fünf Wege zu einer kausalen Schätzung und was jeder annehmen muss

Jeder Weg baut den fehlenden Vergleich anders. Jeder muss etwas annehmen, was die Daten nicht zeigen, und sie schätzen nicht unbedingt denselben Effekt.

Weg	Woher der Vergleich kommt	Was er annehmen muss	Was er schätzt	Woran er scheitern kann
Randomisiertes Experiment	Ein bekannter Zufallsmechanismus teilt die Behandlung zu	Stabile Einheitswerte: Kein Ergebnis hängt von der Behandlung einer anderen Einheit oder vom Beobachtet werden ab	Den durchschnittlichen Behandlungseffekt in der Studienpopulation; bei unvollständiger Compliance nur den Effekt der Zuteilung (Intent-to-Treat-Effekt)	Spillover zwischen Einheiten, Hawthorne- oder John-Henry-Effekte, eine nicht repräsentative Studienpopulation
Adjustierung für beobachtete Kovariaten	Einheiten mit gleichen gemessenen Kovariaten oder gleichem Propensity Score werden verglichen	Bedingte Unabhängigkeit: Die Kovariaten bilden eine ausreichende Menge gemeinsamer Ursachen von Behandlung und Ergebnis (nicht notwendig alle), und jede Einheit hätte jede Behandlung erhalten können	Den durchschnittlichen Effekt oder den Effekt bei den Behandelten, wo sich die Gruppen überlappen	Eine ungemessene gemeinsame Ursache, die die Kovariaten nicht abdecken; Adjustierung für einen Mediator oder Collider

Weg	Woher der Vergleich kommt	Was er annehmen muss	Was er schätzt	Woran er scheitern kann
Natürliches Experiment	Ein äußeres Ereignis oder eine institutionelle Regel ändert die Bedingungen für manche Einheiten und für andere nicht	Gegeben die Kovariaten ist Ereignis oder Regel so gut wie zufällig: Die Kovariaten decken andere gleichzeitige Ursachen von Veränderung und Unterschiede zwischen den getrennten Gruppen ab	Den Effekt des Ereignisses oder der Regel selbst; den Effekt einer Behandlung, auf die Ereignis oder Regel wirkt, liefert erst der Weg über die Instrumentvariable	Gleichzeitige andere Veränderungen; Gruppen, die sich schon vorher unterschieden; wenige Ereignisse, die die Frage klären
Fixed Effects und Difference-in-Differences	Dieselben Einheiten oder Gruppen vorher und nachher beobachtet	Ungemessene Unterschiede sind zeitlich konstant, und die Gruppen teilen einen gemeinsamen Trend	Den Effekt für Einheiten, deren Behandlung sich im Zeitfenster geändert hat	Auseinanderlaufende Trends; Einheiten, die den Zeitpunkt der Behandlung nach dem erwarteten Nutzen wählen; eine selektive Gruppe von Wechslern

Weg	Woher der Vergleich kommt	Was er annehmen muss	Was er schätzt	Woran er scheitern kann
Instrumentvariable	Eine Variable, die die Behandlung bewegt, das Ergebnis aber nicht direkt	Das Instrument ist so gut wie zufällig zugeteilt und bewegt die Behandlung; es wirkt nur über die Behandlung auf das Ergebnis; die Einheiten reagieren nur in eine Richtung	Den lokalen Effekt bei den Einheiten, die das Instrument bewegt hat; ein anderes Instrument kann andere Einheiten bewegen und einen anderen lokalen Effekt liefern	Ein schwaches Instrument; ein zweiter Weg zum Ergebnis; ein Instrument, das eine Ursache mit dem Ergebnis teilt

Zuerst die dritte Spalte lesen. Ein Weg ist nur so stark wie die Annahme, die er nicht vollständig prüfen kann.

Quelle Synthese des Autors aus Rosenbaum und Rubin (1983), S. 41, 43; Gangl (2010), S. 24-29, 33-38; Rubin (1974), S. 699; und Pearl (1995), S. 671. Die Zeilen fassen die Bedingungen der Quellen in den Worten des Autors zusammen; es wird keine Schätzung berichtet.

Wenn die Kovariaten die bedingte Unabhängigkeit nicht voll rechtfertigen, verweist Gangl (2010)¹ auf die Sensitivitätsanalyse. Sie soll das Ausmaß der verbleibenden Verzerrung unter plausiblen Annahmen über das ungemessene Confounding abschätzen („to understand the extent of remaining bias in the empirical estimates under plausible assumptions about the degree of unmeasured confounding in the data“) und ist am aussagekräftigsten, wenn die durchgespielte Spanne empirischen Gehalt hat („are most meaningful if the range of simulated values has empirical content“), etwa wenn sich die wahrscheinliche Größe des Confoundings mit dem Effekt einer bekannten Kovariate vergleichen lässt.

Welche Annahmen begleiten eine Kausalaussage?

§ 6 von 10

Drei. Die erste und die dritte betreffen jeden Weg, die zweite jeden, auf dem Kovariaten die Adjustierung tragen. Die erste: Jede Einheit hat ein Ergebnis je Behandlung. Rosenbaum und Rubin (1983) halten fest, dass die Struktur der potenziellen Ergebnisse versagt, wenn die Reaktion einer Einheit von der Behandlung einer anderen abhängt, wie es geschehen kann, wenn sie um Ressourcen konkurrieren („as could happen if they compete for resources“); die Annahme, dass das nicht geschieht, heißt Stable Unit Treatment Value Assumption („has been called the stable unit-treatment value assumption“). Wo soziale Interaktionen für die Ergebnisse wichtig sind, sind Effekte je Einheit nach Gangl (2010)¹ grundsätzlich undefiniert („unit effects are inherently undefined“), und er nennt zwei Antworten: die Analyse auf eine höhere Ebene verlagern, etwa eine Organisation oder einen lokalen Arbeitsmarkt („An obvious response to potential violations of SUTVA is to move the analysis to a more aggregate level“), oder Standardschätzungen als lokal verstehen („standard treatment effect estimates are then to be considered as local“), also an ihre Zeit und Lage gebunden. In der Lesart dieses Leitfadens sind Vertriebsmitarbeitende, die sich ein Gebiet teilen, oder Accounts, die Angebote untereinander vergleichen, kommerzielle Fälle des Problems, das der Leitfaden zur Interferenz behandelt.

Die zweite ist die Überlappung, die Rosenbaum und Rubin (1983)³ in die starke Ignorierbarkeit aufnehmen. In einem randomisierten Experiment hat jede Einheit der Population eine Chance auf jede Behandlung („every unit in the population has a chance of receiving each treatment“); ein adjustierter Vergleich braucht dasselbe für jeden Wert der Kovariaten. Wo manche Arten von Einheiten immer oder nie behandelt werden, enthalten die Daten keinen Vergleich für sie: Gangl (2010)¹ merkt an, dass Matching keinen Effekt schätzen kann, wo sich die Verteilungen der Kovariaten nicht überlappen („matching cannot estimate a treatment effect in areas where covariate distributions do not overlap for the treatment and control group“), und dass die Schätzung nur für den gemeinsamen Wertebereich der Kovariaten gilt („only applies to the population defined over the common covariate support“).

Die dritte: Die Annahmen werden genannt und begründet, nicht stillschweigend vorausgesetzt. Rubin (1974) rät, gründlich über Variablen neben der Behandlung nachzudenken, die das Ergebnis kausal beeinflussen können, und ihre Kontrolle vorab zu planen („the

investigator should think hard about variables besides the treatment that may causally affect Y and plan in advance how to control for the important ones“), und anzugeben, wie weit die Annahmen der subjektiven Randomisierung und der subjektiven Zufallsstichprobe glaubhaft sind („it is important to indicate the extent to which the assumptions of subjective randomization and subjective random sampling can be believed“); mit diesen Begriffen meint er, Zuteilung und Stichprobe so zu behandeln, als wären sie zufällig.

Welcher Effekt wird geschätzt?

§ 7 von 10

Es gibt mehr als einen, und die Wahl kommt zuerst. Gangl (2010)¹ schreibt, bei heterogenen Effekten gebe es keine einzelne Zahl, die den kausalen Effekt angibt („there is no single number that would provide the causal effect of D on Y“), und die Wahl des Parameters sei die erste theoretische Frage, die der Analysierende klären muss („Determining which causal parameter(s) to focus on in any given study thus becomes a first theoretical issue to be settled by the analyst“). Der durchschnittliche Behandlungseffekt mittelt über alle in der Population. Der durchschnittliche Behandlungseffekt der Behandelten mittelt über die, die die Behandlung erhielten, und kann leichter zu identifizieren und theoretisch aufschlussreicher sein („may be both easier to identify and theoretically more informative“). Der lokale durchschnittliche Behandlungseffekt gilt nur für die, die ein Instrument bewegt hat.

Ein Durchschnitt kann in die Irre führen, auch wenn er stimmt. Rubin (1974)⁴ merkt an, dass der durchschnittliche kausale Effekt, wenn ein Einzeleffekt sehr groß ist und die übrigen klein sind, deutlich größer sein kann als alle Einzeleffekte bis auf einen („the average causal effect may be substantially larger than all but one of the individual causal effects“). Kommerziell gelesen: Der Effekt eines Rabatts auf die Accounts, die ihn erhielten, ist eine andere Zahl als sein Effekt, wenn man ihn allen Accounts anböte, und keine von beiden beschreibt den typischen Account, wenn wenige große das Ergebnis treiben.

Wie unterscheiden sich potenzielle Ergebnisse und Kausaldiagramme?

§ 8 von 10

Nach Pearls Darstellung fassen sie dieselben Annahmen in zwei Sprachen; er liest die Gleichungen hinter einem Diagramm als kontrafaktische Beziehungen und sieht darin eine klare Entsprechung zu Rubins Modell potenzieller Ergebnisse („forming a clear correspondence between causal diagrams and Rubin’s⁴ model of potential outcome“). Das Potential-Outcomes-Modell schreibt die Annahmen als Bedingungen an nicht beobachtete Ergebnisse, etwa als bedingte Unabhängigkeit. Pearl (1995)² schreibt sie als Diagramm, das daraufhin befragt wird, ob die verfügbaren Annahmen genügen, um kausale Effekte aus nicht experimentellen Daten zu identifizieren („diagrams are queried to determine if the assumptions available are sufficient for identifying causal effects from nonexperimental data“). Das Diagramm bildet ab, wie der Forschende die wichtigsten kausalen Einflüsse zwischen messbaren Größen versteht („the investigator’s understanding of the major causal influences among measurable quantities in the domain“), und sein Back-Door-Kriterium lässt sich direkt auf das Diagramm anwenden („can be applied directly to the causal diagram“). Gangl (2010)¹ stützt sich bei der Wahl der Kovariaten auf Pearls spätere Analyse, auch bei der Warnung vor Collidern.

Pearl benennt die Grenze offen: Die Ergebnisse beruhen auf den kausalen Annahmen im Graphen, und diese lassen sich in Beobachtungsstudien meist nicht prüfen („the results must rest on the causal assumptions shown in the graph, and that these cannot usually be tested in observational studies“). Manche lassen sich zumindest widerlegen: Einige, allen voran die zu Instrumentvariablen („most notably those associated with instrumental variables“), sind nach verwandten Arbeiten Pearls Falsifikationstests zugänglich. Gangl (2010)¹ ergänzt, dass die Ausschlussrestriktion eines Instruments grundsätzlich nicht prüfbar ist und theoretisch begründet werden muss („is untestable in principle and needs to be justified theoretically“). Pearls Argument für das Diagramm lautet, dass jede kausale Schlussfolgerung aus Beobachtungsstudien letztlich auf kausalen Annahmen beruht („any causal inferences from observational studies must ultimately rely on some kind of causal assumptions“), und das Diagramm biete eine wirksame Sprache, sie genau und ausdrücklich zu machen („an effective language for making those assumptions precise and explicit“).

Wie hält man eine Kausalaussage fest?

§ 9 von 10

In zwei Hälften. Die erste sagt, worum es in der Aussage geht. Der Essay zur Uplift-Spezifikation gibt dafür ein Blatt mit neun Feldern, geschrieben für Marketing-Uplift-Aussagen: Einheit, Behandlung, Gegenfaktum, Schätzgröße (Estimand), Interferenz, Zeitraum, Ergebnis, Messung und Entscheidung. Seine Prüfschritte vor der Skalierung fragen bereits nach der Zuweisungsregel; die zweite Hälfte fragt, ob diese Zuteilung den Vergleich trägt. Sie sagt, warum der Vergleich kausal ist: wie die Einheiten zu ihrer Behandlung kamen, welcher Weg den Vergleich baut, was dieser Weg annimmt, wie jede Annahme geprüft wurde und welche Aussage übrig bleibt. Das folgt Rubins Rat, die Variablen zu nennen, die das Ergebnis beeinflussen können, ihre Kontrolle vorab zu planen und zu sagen, wie weit die Annahmen glaubhaft sind, und Ganglis Hinweis, dass der gesuchte Effekt vor der Schätzung gewählt wird.

Abbildung 2 Ein Identifikationsblatt

Neben der Aussage festhalten, wie die Einheiten zugeteilt wurden, welcher Weg den Vergleich baut, was er annimmt, wie jede Annahme geprüft wurde und welche Aussage bleibt.

Aussage	Zuteilung	Weg	Annahmen	Prüfungen	Gestützte Aussage
Wie spezifiziert: Einheit, Estimand, Ergebnis, Zeitraum.	Per Zufall, durch eigene Wahl oder nach einer Regel?	Welcher der fünf Wege aus Abbildung 1 baut den Vergleich?	Was muss gelten, damit der Vergleich für das fehlende Ergebnis steht?	Welche Prüfung lief: Balance, Falsifikation, Sensitivität?	Durchschnitt, bei den Behandelten, lokal, ITT oder nur Zusammenhang?

Haben die Einheiten die Behandlung selbst gewählt und ist das Annahmenfeld leer, ist die gestützte Aussage ein Zusammenhang.

Eine Zeile je Aussage, neben ihrer Spezifikation geführt. Im Prüfungsfeld zeigt sich eine ungeprüfte Annahme.

Quelle Entscheidungsblatt des Autors. Die Felder folgen Rubin (1974), S. 700, zur Benennung und Kontrolle anderer Ursachen und zur Glaubhaftigkeit der Annahmen; Gangl (2010), S. 23-29, 32 und 36-37, zu Estimands, Identifikation, Sensitivitätsanalyse und lokalen Effekten; Rosenbaum und Rubin (1983), S. 43, zur Ignorierbarkeit; und Pearl (1995), S. 684, zu Falsifikationstests. Die Felder sind Eingaben der Lesenden; Unternehmensdaten werden nicht geliefert.

Ein durchgespieltes Beispiel ohne Unternehmensdaten: Ein Team sieht, dass Accounts, die ein Onboarding-Programm abgeschlossen haben, häufiger verlängert haben, und will das dem Programm zuschreiben. Die Aussage ist der Effekt des Programmabschlusses auf die Verlängerung nach zwölf Monaten, für die Accounts, die es abgeschlossen haben. Das Zuteilungsfeld sagt: durch eigene Wahl. Der rohe Unterschied trägt also die erste Verzerrungsquelle, die Gangl nennt, einen Unterschied der Ausgangsergebnisse: Accounts, die das Programm aus eigener Wahl abschlossen, hätten vermutlich auch ohne es häufiger verlängert. Seine zweite Quelle, unterschiedliche Effekte, betrifft den Durchschnitt über alle Accounts; Effekte bei den Behandelten lassen sich nach Gangl in der Regel unter etwas schwächeren Bedingungen identifizieren („treatment effects on the treated will generally be identified under somewhat weaker

conditions“), wobei der praktische Nutzen dieses Ergebnisses wohl begrenzt sei („although the practical utility of this result is probably limited“) und die Grenze zwischen beiden Quellen in der Praxis verschwimmen könne („may also blur in practice“). Zwei Designs würden helfen. Eine Adjustierung für Nutzung, Größe und Dauer der Kundenbeziehung, gemessen vor dem Onboarding, beruht auf bedingter Unabhängigkeit und braucht eine Sensitivitätsanalyse. Eine Randomisierung der Einladung liefert dagegen den Effekt der Einladung und, als Instrument genutzt, einen lokalen Effekt für die Accounts, die erst die Einladung zum Abschluss bewegt hat. Sofern auch Accounts ohne Einladung abschließen können, sind das nicht alle Accounts mit Abschluss, und die Complier lassen sich in den Daten nicht bestimmen („the population of compliers cannot actually be identified in the sample data“). Gangl merkt an, dass der lokale Effekt den durchschnittlichen kausalen Effekt der Complier identifiziert, während der Nettoeffekt der Maßnahme dem Intention-to-Treat-Effekt entspricht („identifies the complier average causal effect (CACE), whereas the net effect of the manipulation Z corresponds to the intention-to-treat (ITT) effect“). In beiden Fällen gehört ins Aussagefeld der Effekt, den das Design liefert. Bis eines von beiden läuft, ist die gestützte Aussage ein Zusammenhang.

Wo endet die Evidenz zur kausalen Inferenz?

§ 10 von 10

Bei der Methode. Die vier Quellen sind methodisch: Rubin (1974)⁴ schreibt für die pädagogisch-psychologische Forschung, Rosenbaum und Rubin (1983)³ und Pearl (1995)² für die Statistik, mit Beispielen aus Medizin und Landwirtschaft, und Gangl (2010)¹ fasst die Literatur für die Soziologie zusammen. Keine schätzt einen kommerziellen Effekt. Gangls eigenes Fazit setzt auf das Design: Für viele, wenn nicht die meisten Versuche, in der Soziologie kausale Parameter zu bestimmen, erscheine der Vorrang des Forschungsdesigns als der natürlichere Ausgangspunkt („an emphasis on the primacy of research design seems the more natural starting point for many, if not most, attempts to recover causal parameters in sociology“). Er fügt hinzu, dass (quasi-)experimentelle Methoden zwangsläufig nur ausgewählte kausale Parameter für bestimmte Populationen und Maßnahmen identifizieren („inevitably identify only selected causal parameters for specific populations and specific interventions“), sodass sie die Analyse von Beobachtungsdaten ergänzen statt ersetzen, und dass kausale Inferenz zwar eine gewaltige Herausforderung bleibe („causal inference will inevitably remain a formidable challenge“), die Aussichten kausaler Analyse in der Soziologie aber besser sein dürften als ihr Ruf („may thus be better than its current reputation would suggest“). Pearls Identifikationsanalyse nimmt zudem an, dass die Stichprobe so groß ist, dass sich die Stichprobenschwankung vernachlässigen lässt; ein identifizierter Effekt braucht also noch sein Konfidenzintervall.

Dieser Leitfaden behandelt weder die Schätzung im Einzelnen noch die statistische Inferenz für eine Schätzung noch Methoden, die nach 2010 erschienen sind. Die kommerziellen Beispiele sind Überlegungen des Autors und nicht geprüft.

 Ende des Beitrags

Referenzen

Zitate stehen in der Sprache des Originals; die deutschen Wiedergaben im Fließtext sind sinngemäß und stammen vom Autor.

- 1** Gangl, M. (2010). Causal inference in sociological research. *Annual Review of Sociology*, 36, 21-47.
<https://doi.org/10.1146/annurev.soc.012809.102702>
- 2** Pearl, J. (1995). Causal diagrams for empirical research. *Biometrika*, 82(4), 669-688. <https://doi.org/10.1093/biomet/82.4.669>
- 3** Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1), 41-55.
<https://doi.org/10.1093/biomet/70.1.41>
- 4** Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66(5), 688-701. <https://doi.org/10.1037/h0037350>

Zitierweise

Isoglu, S. (2026, 18. Oktober). Was ist kausale Inferenz? Sinan Isoglu.
<https://isoglu.com/de/insights/journal/was-ist-kausale-inferenz/>
Online-Fassung: <https://isoglu.com/de/insights/journal/was-ist-kausale-inferenz/>

Sinan Isoglu, MBA (Quantic)

Führungskraft für Umsatzwachstum, Dozent und Doktorand
© 2026 Sinan Isoglu · isoglu.com · Alle Rechte vorbehalten