

Aus der Forschung

Was ist das Simpson-Paradoxon?

Simpson-Paradoxon: Ein Zusammenhang, der in jeder Teilgruppe gilt, kehrt sich beim Zusammenfassen um. Welche Tabelle stimmt, ist eine kausale Frage.

Autor	Veröffentlicht	Lesezeit	Wörter
Sinan Isoglu, MBA (Quantic) Führungskraft für Umsatzwachstum, Dozent und Doktorand	18. Oktober 2026	18 Min.	3.735

Management Summary Das Simpson-Paradoxon ist die Umkehrung eines Zusammenhangs zwischen zwei Variablen, wenn nach einer Drittvariable aufgeteilte Daten zusammengefasst werden: Jede Teilgruppe zeigt in die eine Richtung, die Gesamtdaten in die andere. Dieser Konzeptleitfaden definiert es (Pearl; Simpson), erklärt die Arithmetik dahinter und warum sie überrascht, geht Berkeleys Zulassungsdaten zum Graduiertenstudium von 1973 durch, in denen 277 Frauen weniger zugelassen wurden, als die campusweite Erwartung vorsah, aber 60 mehr, als die Zulassungsquoten der einzelnen Fachbereiche erwarten ließen, und stellt Pearls kausales Kriterium dafür vor, welche Tabelle zu lesen ist: die zusammengefasste, die aufgeschlüsselten oder keine, mit einem Blatt für die Entscheidung. Die Quellen sind eine Notiz von 1951 mit konstruierten Beispielen, eine Studie zu den Zulassungsdaten einer Universität und ein Kommentar.

Inhalt	§ 1	Was ist das Simpson-Paradoxon?	3
	§ 2	Wie kann jeder Teil in die eine Richtung zeigen und das Ganze in die andere?	4
	§ 3	Warum wirkt eine Umkehrung unmöglich?	5
	§ 4	Was zeigten die Zulassungsdaten aus Berkeley?	6
	Abbildung 1	Zugelassene Frauen in Berkeley im Herbst 1973, verglichen mit zwei erwarteten Zahlen	8
	§ 5	War die Frage der Verzerrung mit der Rechnung je Fachbereich erledigt?	9
	§ 6	Welche Tabelle beantwortet die Frage?	10
	§ 7	Was sollte feststehen, bevor zwei Segmente zusammengefasst werden?	12
	Abbildung 2	Ein Blatt für die Entscheidung über das Zusammenfassen	13

§ 8	Welche Fragen müssen Analysten selbst beantworten?	14
	Referenzen	15

Was ist das Simpson-Paradoxon?

§ 1 von 8

Es ist ein Vergleich, der sich umkehrt, sobald man die Daten zusammenzählt. Pearl (2014)² gibt die strenge Definition: Das Simpson-Paradoxon bezeichnet ein Phänomen, bei dem der Zusammenhang zwischen zwei Variablen X und Y sein Vorzeichen umkehrt, wenn man nach einer dritten Variable Z aufteilt, und zwar gleich, welchen Wert Z annimmt („regardless of the value taken“, S. 8). Teilt man die Daten nach der Drittvariable auf, zeigt jeder Teil den Zusammenhang in die eine Richtung; fasst man die Teile zusammen, zeigt das Ganze ihn in die andere.

² Pearl, J. (2014). Comment: Understanding Simpson's paradox. *The American Statistician*, 68(1), 8–13.

Die Beobachtung ist älter als der Name, und ihre ersten Beispiele waren schwächer. In Pearls Darstellung der Geschichte berichteten Karl Pearson und Kollegen, Udny Yule und Simpson selbst von Zusammenhängen, die beim Zusammenfassen verschwinden, statt sich umzukehren („reported associations that disappear, rather than reversing“); die erste Umkehrung schreibt Pearl Cohen und Nagel im Jahr 1934 zu, den Namen Paradoxon Blyth im Jahr 1972 (S. 8). Er trennt außerdem zwei Dinge, die der Name vermengt: die Simpson-Umkehrung, ein arithmetisches Phänomen im Rechnen mit Anteilen („an arithmetic phenomenon in the calculus of proportions“), und das Simpson-Paradoxon, ein psychologisches Phänomen, das Überraschung und Unglauben auslöst („a psychological phenomenon that evokes surprise“, S. 9).

In diesem Leitfaden bedeutet Simpson-Paradoxon: Die Richtung eines Zusammenhangs zwischen zwei Variablen innerhalb jeder Teilgruppe, die eine Drittvariable bildet, also in den aufgeschlüsselten Daten, ist der Richtung in den zusammengefassten, aggregierten Daten entgegengesetzt. Pearls strenger Fall verlangt, dass alle Teilgruppen in dieselbe Richtung zeigen. Der Name wird auch für schwächere Fälle verwendet: für einen Zusammenhang, der beim Zusammenfassen verschwindet, wie in Simpsons eigenem Beispiel, oder für eine campusweite Zahl, die ihr Vorzeichen wechselt, sobald für jede Teilgruppe eine eigene Erwartung berechnet wird, wie in Berkeley. Keine der beiden Tabellen enthält einen Rechenfehler. Das Paradoxon ist ein Widerspruch zwischen zwei korrekten Zusammenfassungen, und die Arbeit besteht darin zu entscheiden, welche von ihnen die gestellte Frage beantwortet.

Wie kann jeder Teil in die eine Richtung zeigen und das Ganze in die andere?

§ 2 von 8

Weil eine zusammengefasste Quote ein gewichteter Durchschnitt ist und die Gewichte zwischen den verglichenen Gruppen verschieden sein können. Pearl fasst die Arithmetik in einem Satz zusammen: Studierende sind überrascht, wenn Quotienten von Summen, $(a + b)/(c + d)$, anders geordnet sind als die einzelnen Quotienten a/c und b/d (S. 9). Die zusammengefasste Quote einer Gruppe ist der Durchschnitt ihrer Quoten in den Teilgruppen, gewichtet danach, wie sich die Einheiten dieser Gruppe auf die Teilgruppen verteilen. Verteilen sich die beiden Gruppen unterschiedlich auf Teilgruppen mit unterschiedlichen Quoten, verschiebt sich der zusammengefasste Vergleich zugunsten der Gruppe, die dort konzentriert ist, wo die Quoten hoch sind, und diese Verschiebung kann schwerer wiegen als alles, was innerhalb der Teilgruppen geschieht. Dieselbe Arithmetik steckt in einem Durchschnittspreis oder einem durchschnittlichen Rabatt, sobald sich der Mix des Verkauften ändert; der Leitfaden zur Preis-Volumen-Mix-Analyse trennt einen Mixeffekt von Preis und Volumen und hält fest, dass eine solche Zerlegung die Arithmetik erklärt, nicht aber, warum Käufer ihr Verhalten geändert haben.

Bickel et al. (1975)¹ geben ein hypothetisches Beispiel, das die Gefahr unvorsichtigen Zusammenfassens zeigt („incautious pooling of data“). In ihrer Tabelle 2 bewerben sich bei einem Fachbereich, den sie Machismatik nennen, 400 Männer und 200 Frauen, und er nimmt von beiden die Hälfte auf, 200 Männer und 100 Frauen. Ein zweiter Fachbereich, Sozialkriegsführung, erhält Bewerbungen von 150 Männern und 450 Frauen und nimmt von beiden ein Drittel auf, 50 Männer und 150 Frauen. Zusammengefasst werden 250 von 550 Männern und 250 von 650 Frauen zugelassen, und eine aus den Gesamtsummen berechnete Erwartung ergibt ein Defizit von rund 21 Frauen, obwohl sich gezeigt hatte, dass beide Fachbereiche ihre Bewerberinnen und Bewerber völlig fair behandelt hatten („seen to have been absolutely fair“, S. 400). Das Defizit entsteht daraus, wohin sich die Bewerbungen richteten: Rund 73 Prozent der Männer bewarben sich bei dem Fachbereich, der die Hälfte aufnahm, rund 69 Prozent der Frauen bei dem, der ein Drittel aufnahm.

¹ Bickel, P. J., Hammel, E. A., & O'Connell, J. W. (1975). Sex bias in graduate admissions: Data from Berkeley. *Science*, 187(4175), 398–404.

Simpson (1951)³ bewies, wann Zusammenfassen das Bild nicht verändern kann. Für sein Maß des Zusammenhangs in einer $2 \times 2 \times$

2-Kontingenztafel ohne Wechselwirkung zweiter Ordnung zeigt die zusammengefasste Tafel genau dann denselben Zusammenhang wie die einzelnen Tafeln, wenn das Merkmal, das beim Zusammenfassen verschwindet, von einem der beiden anderen Merkmale unabhängig ist, und zwar auf jeder Stufe des jeweils verbleibenden („if and only if the classification C which is being submerged is independent“, S. 241). Bickel et al.¹ sagen dasselbe mit ihren eigenen Worten: Das Zusammenfassen der Fachbereiche hätte keine Rolle gespielt, wenn sie gleich schwer zugänglich gewesen wären („were equally difficult to enter“, S. 399). Nach der Lesart dieses Leitfadens braucht eine Umkehrung daher eine Drittvariable, die sowohl mit der Gruppierung als auch mit dem Ergebnis zusammenhängt.

³ Simpson, E. H. (1951). The interpretation of interaction in contingency tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 13(2), 238–241.

Warum wirkt eine Umkehrung unmöglich?

§ 3 von 8

Nach Pearls Darstellung liegt es daran, dass man die Vergleiche in den Teilgruppen als Wirkungen liest. Ist die Drittvariable etwas, das eine Behandlung nicht verändern kann, etwa Alter oder Geschlecht, dann kann, so die Überlegung, kein Medikament Männern wie Frauen schaden und zugleich der Gesamtbevölkerung nützen. Er formuliert das Ergebnis als Sure-Thing-Theorem aus seinem Buch: Eine Handlung A, die die Wahrscheinlichkeit eines Ereignisses B in jeder Teilpopulation erhöht, muss sie auch in der Gesamtpopulation erhöhen, sofern die Handlung die Verteilung der Teilpopulationen nicht verändert („does not change the distribution of the subpopulations“, S. 10).

Auf diese Bedingung kommt es an. Ein Zusammenhang kann sich beim Zusammenfassen umkehren; nach Pearls Theorem kann die Wirkung einer Handlung, die die Größe der Teilgruppen unverändert lässt, das nicht. Er folgert, die Überraschung sei kaum zu erklären, ohne anzunehmen, dass die menschliche Intuition einem kausalen Kalkül folgt und statistische Zusammenhänge beharrlich kausal deutet (S. 10). Das ist ein Argument aus seiner Theorie, kein Befund darüber, wie Menschen denken.

Was zeigten die Zulassungsdaten aus Berkeley? § 4 von 8

Bickel et al. (1975)¹ untersuchten die Bewerbungen für ein Graduiertenstudium an der University of California, Berkeley, für das Herbstquartal 1973: 12.763 vollständige Bewerbungen bei 101 Fachbereichen, von 8.442 Männern und 4.321 Frauen. Zugelassen wurden rund 44 Prozent der Männer und rund 35 Prozent der Frauen. Die Autoren fragten, ob die Entscheidung über Zulassung oder Ablehnung vom Geschlecht der Bewerbenden beeinflusst war, und gingen absichtlich naiv vor, weil sie annahmen, andere, die sich für Fragen der Verzerrung interessieren, würden genauso vorgehen. Im ersten Schritt verglichen sie die Zulassungen mit erwarteten Häufigkeiten aus den campusweiten Summen, unter zwei Annahmen: dass sich Männer und Frauen, die sich für dasselbe Fach bewerben, in ihrer Qualifikation nicht unterscheiden (Annahme 1), und dass die Geschlechterverteilung der Bewerbungen auf die Fächer mit keinen anderen Faktoren der Zulassung nennenswert zusammenhängt (Annahme 2). Danach wurden 277 Frauen weniger und 277 Männer mehr zugelassen als erwartet, bei einem Chi-Quadrat von 110,8 (S. 398 bis 399, Tabelle 1).

Über Zulassungen entscheiden vor allem die Fachbereiche, also suchten die Autoren als Nächstes die verantwortlichen Fachbereiche. Nach Ausschluss von 16 Fachbereichen, die keine Bewerberinnen hatten oder alle Bewerbungen annahmen, fanden sie unter den übrigen 85 vier, die auf dem 5-Prozent-Niveau zulasten der Frauen verzerrt waren, mit einem gemeinsamen Defizit von 26 Frauen, und sechs, die zulasten der Männer verzerrt waren, mit einem Defizit von 64 Männern (S. 399). Diese Ergebnisse seien verwirrend, schrieben sie („These results are confusing“): Ein Defizit von 277 dürfe nicht einfach verschwinden („ought not simply to disappear“).

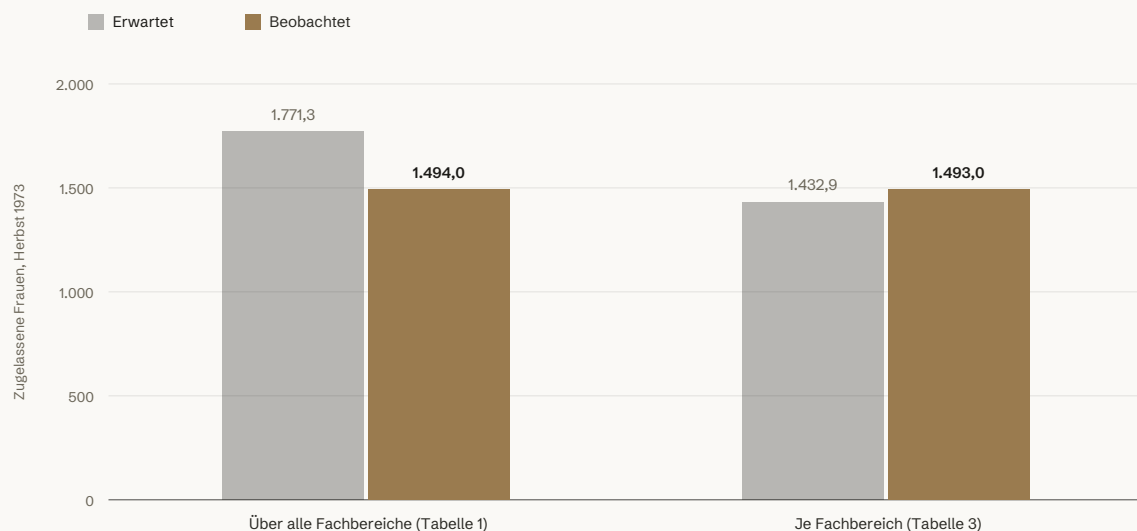
Ihre Diagnose: Sie seien auf ein Paradoxon gestoßen, das in diesem Zusammenhang mitunter nach Simpson benannt werde („We have stumbled onto a paradox“; „sometimes referred to as Simpson’s³“). Annahme 2 war falsch. Männer und Frauen bewarben sich bei verschiedenen Fachbereichen: Fast zwei Drittel der Bewerbungen für Englisch, aber nur 2 Prozent der Bewerbungen für Maschinenbau kamen von Frauen, und die Fachbereiche unterschieden sich stark darin, wie schwer sie zugänglich waren. Der Frauenanteil unter den Bewerbungen war tendenziell hoch in Fachbereichen, die schwer zugänglich waren, und niedrig in leicht zugänglichen (S. 399). Gewichtet nach dem Anteil jedes Fachbereichs an allen Bewerbungen, korrelierten die Zulassungsquote eines Fachbereichs und der

Männeranteil unter seinen Bewerbungen über die 85 Fachbereiche mit 0,56 (S. 399 und Anmerkung 8).

Ihre Abhilfe bestand darin, die Erwartung Fachbereich für Fachbereich aufzubauen. Für jeden Fachbereich ist die erwartete Zahl zugelassener Frauen seine Zulassungsquote über alle Bewerbungen mal die Zahl der Bewerberinnen; die campusweite Erwartung ist die Summe. So gerechnet wurden 1.432,9 Frauen erwartet und 1.493,0 zugelassen, 60,1 mehr als erwartet, bei $P = 0,003$, zweiseitig (S. 402, Tabelle 3). Die Rechnung umfasste 96 Fachbereiche, ohne fünf mit nur je einer Bewerbung, und verwendete einen Test, den Cochran sowie Mantel und Haenszel in einem anderen Zusammenhang vorgeschlagen hatten (S. 401 bis 402). Abbildung 1 stellt beide Erwartungen nebeneinander.

Abbildung 1 Zugelassene Frauen in Berkeley im Herbst 1973, verglichen mit zwei erwarteten Zahlen

Gegenüber der campusweiten Erwartung wurden 277 Frauen weniger zugelassen als erwartet. Gegenüber der Zulassungsquote jedes Fachbereichs waren es 60 mehr.



Die Erwartung in Tabelle 1 unterstellt, dass der Frauenanteil unter den Bewerbungen nicht damit zusammenhängt, wie schwer ein Fachbereich zugänglich ist; die Erwartung in Tabelle 3 verzichtet auf diese Annahme, hält aber an gleicher Qualifikation fest. Tabelle 3 lässt fünf Fachbereiche mit je einer Bewerbung aus, deshalb zählt sie 1.493,0 zugelassene Frauen gegenüber 1.494 in Tabelle 1. Ihre Differenz war auf dem Niveau $P = 0,003$ signifikant, zweiseitig.

Quelle Bickel et al. (1975), S. 399, Tabelle 1 (12.763 Bewerbungen bei 101 Fachbereichen; erwartete Zahlen aus den campusweiten Summen unter den Annahmen 1 und 2) und S. 402, Tabelle 3 (96 Fachbereiche; erwartete Zahlen aus der Zulassungsquote jedes Fachbereichs über alle Bewerbungen, nur unter Annahme 1). Verwaltungsdaten zu Zulassungen einer Universität.

Dieselben Zulassungen von Frauen sind gegenüber der einen Erwartung ein Defizit und gegenüber der anderen ein Überschuss. In der Zusammenfassung der Autoren zeigen die aggregierten Daten ein klares, aber irreführendes Muster der Verzerrung zulasten der Bewerberinnen („a clear but misleading pattern“); wenige Fachbereiche weichen signifikant von der erwarteten Zahl zugelassener Frauen ab, ungefähr ebenso viele scheinen Frauen wie Männer zu bevorzugen, und richtig zusammengefasst zeigen die Daten eine kleine, aber statistisch signifikante Verzerrung zugunsten der Frauen (S. 403). Streng genommen ist das nicht Pearls Fall, denn die Fachbereiche zeigten nicht alle in dieselbe Richtung. Was sich

umkehrte, war die campusweite Zahl, sobald für jeden Fachbereich eine eigene Erwartung berechnet wurde.

War die Frage der Verzerrung mit der Rechnung je Fachbereich erledigt? § 5 von 8

Für sich genommen nicht, und die Autoren sagen, warum. Auch die bereinigte Rechnung beruht auf Annahme 1, gleicher Qualifikation, die sie in ihrem weitesten Sinn für eindeutig falsch halten („its most extensive sense“): Die mathematische Vorbildung unterschied sich, und gerade die Fachbereiche, die weniger Mathematik verlangten, waren schwerer zugänglich (eine Pearson-Korrelation von 0,38 zwischen dem Anteil zugelassener Bewerbungen und den im Grundstudium empfohlenen oder vorgeschriebenen Einheiten in Mathematik oder Statistik, über 83 Programme; S. 402). Ein niedriger Gesamtwert kann zudem Fachbereiche verbergen, die in entgegengesetzte Richtungen verzerrt sind, denn niedrige Werte könnten ebenso gut durch gegenseitige Aufhebung entstehen („arise as a consequence of cancellation“, S. 402): Ein Fachbereich, der zu wenige Frauen aufnimmt, und einer, der zu viele aufnimmt, können sich zu null summieren. Der Beitrag zur Forecastgenauigkeit zeigt dasselbe für vorzeichenbehaftete Fehler, die sich in einer Summe aufheben.

Auch der Blick auf einzelne Fachbereiche hat seine Probleme. Manche Fachbereiche waren klein, und bei manchen durchschnittlich großen bewarben sich nur sehr wenige Frauen; wo die Zahlen sehr klein sind, ist die Evidenz für die Entscheidung zwischen keiner und grober Verzerrung nach den Worten der Autoren wertlos („really worthless“, S. 401). Wer 85 Fachbereiche gleichzeitig prüft, findet außerdem schon durch Zufall Extremfälle: Beim Fachbereich mit der stärksten Verzerrung zulasten der Frauen entstände eine so große Verzerrung durch Zufall nur 69-mal in 100.000 Fällen, wenn man ihn allein betrachtete; dass unter 85 Fachbereichen einer so verzerrt ist, ergibt sich durch Zufall aber etwa 57-mal in 1.000 Fällen (S. 401).

Ihr Schluss ist begrenzt. Der Campus als Ganzes habe Bewerberinnen nicht diskriminiert, schreiben sie, und die Daten für 1969 bis 1972 deuteten auf wenig Verzerrung irgendeiner Art (S. 402); das Urteil schließe Einzelfälle von Benachteiligung nicht aus („the possibility of individual cases“, S. 403). Ihre Erklärung des Musters reicht über die Zulassungsdaten hinaus: Die Verzerrung in den aggregierten Daten rühre anscheinend von einer Vorauswahl auf früheren Stufen des Bildungssystems her („apparently from prior screening“, S. 403). Diese Erklärung begründet der Aufsatz mit Argumenten; seine Daten prüfen sie nicht.

Welche Tabelle beantwortet die Frage?

§ 6 von 8

Die Daten können es nicht sagen. Simpson (1951)³ zeigte das mit einem Satz Zahlen und zwei Geschichten. In einem konstruierten Beispiel prüft jemand ein Kartenspiel, von dem einige Karten schmutzig sind, weil ein Kleinkind damit gespielt hat, und findet sowohl unter den schmutzigen als auch unter den sauberen Karten einen positiven Zusammenhang zwischen roter Farbe und Zahlenkarten („a positive association between redness and plainness both among the dirty cards and among the clean“); die zusammengefasste Tafel zeigt keinen, und gerade sie liefere, was man die vernünftige Antwort nennen würde, nämlich dass es keinen solchen Zusammenhang gibt („which provides what we would call the sensible answer, namely, that there is no such association“, S. 240 bis 241). Dann benennt Simpson die Klassen um: männlich und weiblich, unbehandelt und behandelt, lebend und tot. Die Wahrscheinlichkeiten sind genau dieselben wie in seiner Tabelle 2 („The probabilities are exactly the same as in Table 2“), und nun gibt es unter Männern wie unter Frauen einen positiven Zusammenhang zwischen Behandlung und Überleben, in der Gesamtpopulation aber keinen. Diesmal liefern die aufgeschlüsselten Tafeln die vernünftige Antwort: Die Behandlung könne kaum als wertlos verworfen werden, wenn sie bei Männern und bei Frauen nützt („The treatment can hardly be rejected as valueless to the race when it is beneficial when applied to males and to females.“, S. 241).

Pearl zieht daraus den Schluss. Lindley und Novick hätten, so berichtet er, gezeigt, dass kein statistisches Kriterium vor falschen Schlüssen warnt oder anzeigt, welche Daten die richtige Antwort enthalten („that there was no statistical criterion“); bei denselben Daten müsse die Entscheidung daher von zusätzlicher Information aus dem Kontext getragen werden, nicht von statistischen Überlegungen („our decision must be driven not by statistical considerations“, S. 8 bis 9). In seiner Lesart, die Hernán, Clayton und Keiding folgt, unterscheiden sich Simpsons zwei Beispiele in ihrer kausalen Struktur: In der medizinischen Fassung war das Geschlecht eine gemeinsame Ursache der beiden anderen Variablen, bei den Karten war die Verschmutzung eine gemeinsame Wirkung von Farbe und Kartenart (S. 9, Anmerkung 4).

Sein Kriterium für eine einzelne Drittvariable Z , das Back-Door-Kriterium, wird an einem Kausaldiagramm abgelesen und hat zwei Bedingungen: Z ist kein Nachkomme von X , die Behandlung X kann Z also nicht verändern („ Z is not a descendant of X “), und Z blockiert jeden Pfad, der mit einem Pfeil in X endet („ Z blocks every

path that ends with an arrow into X“). Gelten beide, liefern die aufgeschlüsselten Tafeln die Wirkung von X auf das Ergebnis Y; muss kein solcher Pfad blockiert werden, liefert sie die zusammengefasste Tafel (S. 11). Eine gemeinsame Wirkung verhält sich umgekehrt: Konditioniert man auf sie, öffnet man einen Pfad, statt ihn zu schließen. Das ist das allgemeine Phänomen, das als Berksons Paradoxon bekannt ist; in Pearls Beispiel werden die Ergebnisse zweier unabhängiger Münzwürfe abhängig, sobald jemand bezeugt, dass mindestens eine der Münzen Zahl zeigt („the outcomes of two independent coins are rendered dependent by the testimony that at least one of them is a tail“, S. 11). Der Leitfaden zum Selektionsbias behandelt die Konditionierung auf eine solche gemeinsame Folge als einen Selektionspfad.

Ob die Behandlung die Drittvariable verändern kann, ist wichtig, entscheidet die Wahl aber nicht allein. In Pearls Blutdruckbeispiel beeinflusst das Medikament den Blutdruck, und auf die Frage, welches Medikament wirksamer ist, lautet seine Antwort unter der Annahme, dass es keine ungemessenen Confounder gibt, also keine nicht gemessenen gemeinsamen Ursachen von Behandlung und Ergebnis: den Blutdruck ignorieren und die zusammengefassten Daten lesen (S. 12). In einem anderen seiner Diagramme aber ist die Drittvariable von der Behandlung unbeeinflusst und soll trotzdem nicht kontrolliert werden („is not affected by the treatment, and still, it should not be controlled for“, S. 12). Und in manchen Modellen liegt die richtige Antwort weder in den aufgeschlüsselten noch in den aggregierten Daten („may not lie in either the disaggregated or the aggregated“), weil die Drittvariable nicht ausreicht, um einen Pfad in die Behandlung zu blockieren, und weitere Kovariaten nötig sein können (S. 11).

Nach der Lesart dieses Leitfadens zeigt Berkeley, dass die Frage ebenso viel entscheidet wie das Diagramm. Männer und Frauen bewarben sich bei verschiedenen Fachbereichen; ist das Geschlecht die Behandlung, ist der Fachbereich ihm also nachgelagert und verfehlt Pearls erste Bedingung für den Gesamteffekt des Geschlechts auf die Zulassung. Bickel et al.¹ stellten eine engere Frage, nämlich ob die Zulassungsentscheidung vom Geschlecht der Bewerbenden beeinflusst war, und für diese Frage ist der Fachbereich der Ort der Entscheidung und die Variable, die konstant zu halten ist, sofern sich Männer und Frauen, die sich bei demselben Fachbereich bewerben, in ihrer Qualifikation nicht unterscheiden. Diese Bedingung ist die Annahme 1 der Autoren, ohne die sich, wie sie schreiben, jeder Unterschied in der Zulassung nach Geschlecht auf die Qualifikation zurückführen ließe (S. 398), und die sie in ihrem weitesten Sinn für

eindeutig falsch halten (S. 402). Pearl hält fest, dass Zwischenvariablen für andere Fragen als die nach Gesamteffekten, etwa wie viel einer Wirkung über eine Drittvariable läuft, durchaus nützliche Information liefern können („intermediate variables can of course provide useful“, S. 12).

Pearl nennt zwei Wege zur richtigen Tabelle: einen mathematischen und, für Lesende, die eine Simulation eher überzeugt, eine Prüfung durch Simulation (S. 10). In seinem Ratespiel prüft jemand, der das Modell hinter den Daten kennt, die Tabellenwahl der Analysierenden gegen den Goldstandard eines auf dem Modell simulierten randomisierten Experiments („the gold standard of a randomized trial, simulated on the model“, S. 11), und Pearl argumentiert, dass moderne kausale Werkzeuge mathematisch zu bestimmen erlauben, ob die aggregierten oder die aufgeschlüsselten Befunde gelten („permit us to determine mathematically whether the aggregated“, S. 12). Der Kommentar begründet das aus der kausalen Theorie, mit Diagrammen und einem Simulationsspiel, das er beschreibt, aber nicht durchführt, nicht mit Daten. Der Aufsatz von Armistead, auf den er antwortet, hielt fest, dass Drittvariablen, ob kausal oder nicht, wichtige Information tragen können; Pearl stimmt dem allgemein zu, entgegnet aber, dass sich für den Gesamteffekt der Behandlung die richtige Tabelle ohne weitere Prüfung übernehmen lässt (S. 12).

Was sollte feststehen, bevor zwei Segmente zusammengefasst werden?

§ 7 von 8

Fünf Dinge, schriftlich festgehalten, bevor die Tabellen gelesen werden: die Frage (welche Handlung, welches Ergebnis), die Drittvariable, nach der die Daten aufgeteilt werden, ob diese Variable sowohl mit den verglichenen Gruppen als auch mit dem Ergebnis zusammenhängt, ob die Handlung sie verändern kann und ob sie jeden Pfad in die Handlung schließt. Die letzten beiden sind Pearls Bedingungen; die dritte folgt Simpsons Bedingung dafür, wann Zusammenfassen keine Rolle spielen kann. Die Antwort kommt in eine letzte Spalte: die zusammengefasste Tabelle, die aufgeschlüsselten Tabellen oder keine von beiden, bis mehr gemessen ist.

Abbildung 2 Ein Blatt für die Entscheidung über das Zusammenfassen

Frage, Drittvariable und die Antworten auf drei Prüfungen festhalten, bevor die zusammengefasste oder die aufgeschlüsselte Tabelle gewählt wird.

Frage	Drittvariable	Mit beidem verknüpft	Durch Handlung verändert	Pfade geschlossen	Tabelle
Welche Handlung, welches Ergebnis?	Wonach wird aufgeteilt?	Mit Gruppen und Ergebnis?	Kann die Handlung sie ändern?	Schließt sie jeden Pfad hinein?	Zusammengefasst, aufgeteilt, keine?

Aufteilen, wenn die Handlung sie nicht ändern kann und sie jeden Pfad hinein schließt; zusammenfassen, wenn kein Pfad hinein offen ist; sonst mehr messen.

Eine Zeile je Vergleich. Die dritte Spalte prüft anhand von Simpsons Bedingung: Für sein Maß des Zusammenhangs in einer $2 \times 2 \times 2$ -Kontingenztafel ohne Wechselwirkung zweiter Ordnung kann Zusammenfassen den Zusammenhang nur verändern, wenn die Drittvariable mit beidem verknüpft ist. Die vierte und fünfte sind Pearls zwei Bedingungen für eine einzelne Drittvariable. Ein Pfad hinein ist jeder Pfad, der mit einem Pfeil in die Handlung endet; in die fünfte Spalte gehört auch, ob ohne Aufteilung einer offen ist, nicht nur, ob die Drittvariable jeden schließt.

Quelle Entscheidungsblatt des Autors. Das Feld zur Verknüpfung mit beidem folgt Simpson (1951), S. 241; die Felder zur Veränderung durch die Handlung und zu geschlossenen Pfaden folgen Pearl (2014), S. 11; die Entscheidung folgt Pearl (2014), S. 11 bis 12. Die Felder sind Eingaben der Lesenden; es werden keine Unternehmensdaten geliefert.

Ein Beispiel mit erfundenen Zahlen und ohne Unternehmensdaten: Zwei Vertriebsteams gewähren Rabatte auf kleine und große Abschlüsse. Team A gibt im Schnitt 10 Prozent auf seine 30 kleinen und 20 Prozent auf seine 70 großen Abschlüsse; Team B gibt 12 Prozent auf 80 kleine und 22 Prozent auf 20 große. Innerhalb jeder Größenklasse gibt Team A weniger Rabatt, über alle Abschlüsse kommt es aber auf 17 Prozent gegenüber 14 Prozent bei Team B, weil sieben von zehn seiner Abschlüsse groß sind. Angenommen, Abschlüsse werden bei ihrer Anlage nach ihrer Größe einem Team zugewiesen, und nichts anderes, was den Rabatt beeinflusst, entscheidet über die Zuweisung. Dann erfüllt die Abschlussgröße beide Bedingungen Pearls, und der aufgeschlüsselte Vergleich beantwortet, ob Team A bei vergleichbaren Abschlüssen mehr Rabatt gibt; in diesen Zahlen tut es das nicht.

Vergrößert Team A Abschlüsse außerdem, indem es vor dem Angebot weitere Nutzerlizenzen hinzufügt, sind einige seiner großen Abschlüsse wegen des Teams groß. Die Abschlussgröße ist dann teilweise dem Team nachgelagert, und die aufgeschlüsselte Tabelle lässt den Teil der Wirkung des Teams aus, der über die Abschlussgröße läuft. Die Größe bei der Zuweisung festzuhalten, bevor eines der Teams den Abschluss bearbeitet hat, löst dieses Problem nur, wenn die Größe das Einzige war, was die Zuweisung nutzte und was zugleich die Rabatte beeinflusst. Das sind Beispiele des Autors, und sie sind ungeprüft.

Welche Fragen müssen Analysten selbst beantworten?

§ 8 von 8

Wie oft Umkehrungen in kommerziellen Daten vorkommen und wie man eine Tabelle wählt, wenn die kausale Geschichte strittig ist. Simpson (1951)³ ist eine vierseitige Notiz über Kontingenztafeln mit konstruierten Beispielen. Bickel et al. (1975)¹ analysieren die Zulassungsdaten einer Universität für einen Zyklus im Detail, mit campusweiten Prüfungen für 1969 bis 1973, unter einer Annahme gleicher Qualifikation, die sie selbst in ihrem weitesten Sinn für falsch halten. Pearl (2014)² ist ein Kommentar zu einem anderen Aufsatz: Seine Kriterien stammen aus der kausalen Theorie, auf die er verweist, und werden an Diagrammen und einem Simulationsspiel gezeigt, das er beschreibt, aber nicht durchführt, nicht an neuen Daten; reicht eine Drittvariable nicht aus, führt die Wahl weiterer Kovariaten über den Rahmen der Notiz hinaus (S. 11).

Nach der Lesart dieses Leitfadens lassen sich zwei Dinge mit einiger Zuversicht auf Preis- und Vertriebsdaten übertragen. Jede Quote und jeder Durchschnitt, die über Segmente mit unterschiedlichem Mix zusammengefasst werden, können sich umkehren; ein zusammengefasster Vergleich ist also eine Aussage über den Mix ebenso wie über die Gruppen. Und die Wahl zwischen zusammengefasstem und aufgeschlüsseltem Vergleich ist eine kausale Entscheidung, die begründet werden muss, bevor man die Zahlen liest, und die sich nicht in ihnen findet. Ob sich ein bestimmter Vergleich von Rabatten, Abschlussquoten oder Verlängerungen umkehrt und welche Tabelle für ihn richtig ist, sagen die Quellen nicht.

 Ende des Beitrags

Referenzen

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

- 1** Bickel, P. J., Hammel, E. A., & O'Connell, J. W. (1975). Sex bias in graduate admissions: Data from Berkeley. *Science*, 187(4175), 398–404.
<https://doi.org/10.1126/science.187.4175.398>
- 2** Pearl, J. (2014). Comment: Understanding Simpson's paradox. *The American Statistician*, 68(1), 8–13.
<https://doi.org/10.1080/00031305.2014.876829>
- 3** Simpson, E. H. (1951). The interpretation of interaction in contingency tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 13(2), 238–241. <https://doi.org/10.1111/j.2517-6161.1951.tb00088.x>

Zitierweise

Isoglu, S. (2026, 18. Oktober). Was ist das Simpson-Paradoxon? Sinan Isoglu.
<https://isoglu.com/de/insights/journal/was-ist-das-simpson-paradoxon/>
Online-Fassung:
<https://isoglu.com/de/insights/journal/was-ist-das-simpson-paradoxon/>

Sinan Isoglu, MBA (Quantic)

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