

From the research bench

What is Simpson's paradox?

Simpson's paradox: an association that holds in every subgroup reverses when the groups are pooled. Which table to trust is a causal question.

Author	Published	Reading time	Words
Sinan Isoglu, MBA (Quantic) Commercial growth leader, lecturer and doctoral researcher	18 October 2026	16 min	3,506

Management summary

Simpson's paradox is the reversal of an association between two variables when data split by a third variable are pooled: every subgroup points one way and the combined data point the other. This concept guide defines it (Pearl; Simpson), explains the arithmetic behind it and why it surprises, works through the 1973 Berkeley graduate admissions data, in which a shortfall of 277 women against the campus-wide expectation became a surplus of 60 against each department's own admission rate, and sets out Pearl's causal criterion for reading the pooled table, the split tables or neither, with a record for a pooling decision. The sources are one 1951 note with constructed examples, one study of one university's admissions records and one commentary.

Contents	§ 1	What is Simpson's paradox?	2
	§ 2	How can every part point one way and the whole the other?	3
	§ 3	Why does a reversal feel impossible?	4
	§ 4	What happened in the Berkeley admissions data?	4
	Figure 1	Women admitted at Berkeley in fall 1973, against two expected counts	6
	§ 5	Did the department-by-department count settle the question of bias?	7
	§ 6	Which table answers the question?	8
	§ 7	What should be settled before two segments are pooled?	10
	Figure 2	A pooling decision record	10
	§ 8	Which questions do these sources leave to the analyst?	11
	References	12	

What is Simpson's paradox?

§ 1 of 8

It is a comparison that turns round when the data are added up. Pearl (2014)² gives the strict definition: the association between two variables, X and Y, reverses its sign when the data are conditioned on a third variable, Z, “regardless of the value taken” by Z (p. 8). Split the data by the third variable and every part shows the association in one direction; pool the parts and the whole shows it in the other.

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

The observation is older than the name, and its first examples were weaker. In Pearl's² history, Karl Pearson and colleagues, Udny Yule and Simpson himself all “reported associations that disappear, rather than reversing” signs on aggregation; Pearl credits the first reversal to Cohen and Nagel, in 1934, and the name paradox to Blyth, in 1972 (p. 8). He also separates two things the name runs together: Simpson's³ reversal, “an arithmetic phenomenon in the calculus of proportions”, and Simpson's paradox, “a psychological phenomenon that evokes surprise” and disbelief (p. 9).

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.

In this guide, Simpson's³ paradox means that the direction of an association between two variables inside each subgroup defined by a third variable, the disaggregated data, is the opposite of its direction in the pooled, or aggregated, data. Pearl's² strict case has every subgroup pointing the same way. The name is also used for weaker cases: an association that vanishes on pooling, as in Simpson's own example, or a campus-wide count that changes sign once each subgroup is given its own expectation, as at Berkeley. Neither table contains an arithmetic error. The paradox is a disagreement between two correct summaries, and the work is to decide which one answers the question being asked.

How can every part point one way and the whole the other?

§ 2 of 8

Because a pooled rate is a weighted average, and the weights can differ between the groups being compared. Pearl puts the arithmetic in one line: students are surprised when ratios of sums, $(a + b)/(c + d)$, are ordered differently from the individual ratios a/c and b/d (p. 9). Each group's pooled rate is the average of its subgroup rates, weighted by how that group's own units are spread across the subgroups. When the two groups are spread differently across subgroups whose rates differ, the pooled comparison tilts in favour of the group concentrated where rates are high, and the tilt can outweigh what happens inside every subgroup. The same arithmetic sits inside an average price or discount whenever the mix of what was sold changes; the [price-volume-mix guide](#) separates a mix effect from price and volume and says that such a decomposition explains the arithmetic, not why buyers changed their behaviour.

Bickel et al. (1975)¹ give a hypothetical example of the danger of “incautious pooling of data”. In their Table 2, a department they call machismatics receives 400 men and 200 women and admits half of each, 200 men and 100 women. A second department, social warfare, receives 150 men and 450 women and admits a third of each, 50 men and 150 women. Pooled, 250 of 550 men and 250 of 650 women are admitted, and an expectation computed from the pooled totals shows a deficit of about 21 women, although both departments were “seen to have been absolutely fair” in dealing with their applicants (p. 400). The deficit comes from where the applicants went: about 73 per cent of the men applied to the department that admitted half, and about 69 per cent of the women to the one that admitted a third.

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 proved when pooling cannot change the picture. For his measure of association in a $2 \times 2 \times 2$ table with no second order interaction, the pooled table shows the same association as the separate tables “if and only if the classification C which is being submerged is independent” of one of the other two attributes at each level of the remaining one (p. 241). Bickel et al.¹ make the same point in their own terms: pooling the departments would not have mattered if they “were equally difficult to enter” (p. 399). On this guide's reading, a reversal therefore needs a third variable that is linked both to the grouping and to the outcome.

Why does a reversal feel impossible?

§ 3 of 8

Because, on Pearl's² account, people read the subgroup comparisons as effects. If the third variable is something a treatment cannot change, such as age or sex, then, as the reasoning goes, no drug can harm both males and females and yet benefit the population as a whole. He states the result as the sure-thing theorem from his book: an action that increases the probability of an event in each subpopulation must also increase it in the population as a whole, provided that the action “does not change the distribution of the subpopulations” (p. 10).

The proviso is the point. An association can reverse on pooling; on Pearl's² theorem, the effect of an action that leaves the sizes of the subgroups unchanged cannot. He concludes that the surprise is hard to explain unless human intuition follows causal calculus and people persistently read statistical associations as causal (p. 10). That is an argument from his theory, not a finding about how people reason.

What happened in the Berkeley admissions data?

§ 4 of 8

Bickel et al. (1975)¹ examined the applications for graduate study at the University of California, Berkeley, for the fall 1973 quarter: 12,763 complete applications to 101 departments, from 8,442 men and 4,321 women. About 44 per cent of the men and about 35 per cent of the women were admitted. The authors asked whether the decision to admit or deny was influenced by the applicant's sex, and began naively on purpose, because they thought it likely that others interested in bias would proceed in just the same way. Their first step compared the admissions with expected frequencies computed from the campus totals under two assumptions: that men and women applying to any one discipline do not differ in their qualifications (assumption 1), and that the sex ratios of applicants to the various fields are not substantially associated with any other factor in admission (assumption 2). On that basis 277 fewer women and 277 more men were admitted than expected, with a chi-square of 110.8 (pp. 398 to 399, Table 1).

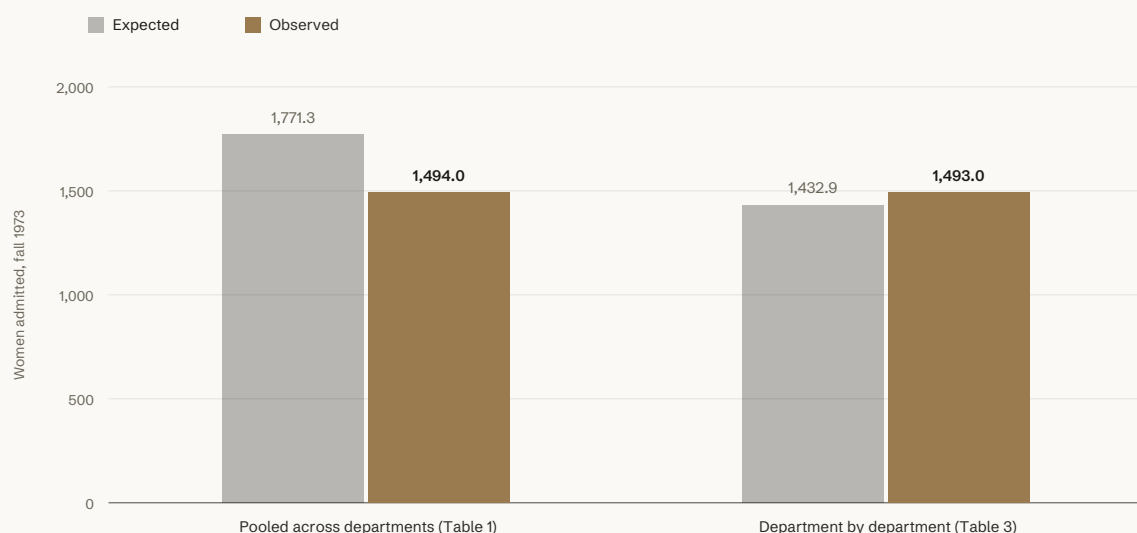
Departments mainly decide admissions, so the authors then looked for the departments responsible. After excluding 16 departments that had no women applicants or admitted every applicant, they found among the remaining 85 four departments biased against women at the 5 per cent level, with a combined deficit of 26 women, and six biased against men, with a deficit of 64 men (p. 399). “These results are confusing,” they wrote: a deficit of 277 “ought not simply to disappear”.

Their diagnosis: “We have stumbled onto a paradox”, one “sometimes referred to as Simpson’s³” in this context. Assumption 2 was false. Men and women applied to different departments, almost two-thirds of the applicants to English but only 2 per cent of the applicants to mechanical engineering being women, and the departments differed widely in how hard they were to enter. The share of women among applicants tended to be high in departments that were hard to get into and low in those that were easy (p. 399). Weighting each department by its share of applicants, the correlation between a department’s admission rate and its share of men among applicants was .56 across the 85 departments (p. 399 and note 8).

Their remedy was to build the expectation department by department. For each department, the expected number of women admitted is its admission rate for all applicants times the number of women who applied, and the campus expectation is the sum. On that basis 1,432.9 women were expected and 1,493.0 admitted, 60.1 more than expected, with $P = .003$, two-tailed (p. 402, Table 3). The computation covered 96 departments, leaving out five that had a single applicant each, and used a test that Cochran and Mantel and Haenszel had proposed in another context (pp. 401 to 402). Figure 1 sets the two expectations side by side.

Figure 1 Women admitted at Berkeley in fall 1973, against two expected counts

Against the campus-wide expectation, 277 fewer women were admitted than expected. Against each department's own admission rate, 60 more were.



The Table 1 expectation assumes that the share of women among applicants is unrelated to how hard a department is to enter; the Table 3 expectation drops that assumption but keeps the one of equal qualifications. Table 3 leaves out five departments with one applicant each, which is why it counts 1,493.0 women admitted against 1,494 in Table 1. Its difference was significant at $P = .003$, two-tailed.

Source Bickel et al. (1975), p. 399, Table 1 (12,763 applications to 101 departments; expected counts from the campus totals under assumptions 1 and 2) and p. 402, Table 3 (96 departments; expected counts from each department's admission rate for all applicants, under assumption 1 only). Administrative admissions records of one university.

The same admissions of women are a shortfall against one expectation and a surplus against the other. In the authors' summary, the aggregate data show "a clear but misleading pattern" of bias against female applicants; few departments depart significantly from the expected admissions of women, about as many appear to favour women as to favour men, and, properly pooled, the data show a small but statistically significant bias in favour of women (p. 403). Strictly, this is not Pearl's² case, because the departments did not all point the same way. What turned round was the campus-wide count, once each department was given its own expectation.

Did the department-by-department count settle the question of bias? § 5 of 8

Not on its own, and the authors say why. The adjusted count still rests on assumption 1, equal qualifications, which they call clearly false in “its most extensive sense”: preparation in mathematics differed, and the departments requiring less mathematics were the harder ones to enter (a Pearson r of .38 between the share of applicants admitted and the undergraduate units in mathematics or statistics recommended or required, across 83 programmes; p. 402). A low aggregate measure can also hide departments biased in opposite directions, since low values could equally well “arise as a consequence of cancellation” (p. 402): a department that admits too few women and another that admits too many can sum to nothing. The forecast accuracy essay makes the same point about signed errors that cancel in a total.

Looking department by department has its own problems. Some departments were small, and in some of ordinary size very few women applied; where the numbers are very small, the evidence for deciding between no bias and gross bias is, in the authors’ words, “really worthless” (p. 401). Examining 85 departments at once also throws up extreme cases by chance: for the department most biased against women, a bias that large would arise by chance only 69 times in 100,000 in one department examined on its own, but a department that biased would turn up by chance about 57 times in 1,000 among 85 (p. 401).

Their conclusion is bounded. The campus as a whole, they write, did not engage in discrimination against women applicants, and the data for 1969 to 1972 suggested little evidence of bias of any kind (p. 402); the judgement does not eliminate “the possibility of individual cases” of prejudicial treatment (p. 403). Their explanation of the pattern goes beyond the admissions records: the bias in the aggregate stems “apparently from prior screening” at earlier levels of the educational system (p. 403). The paper argues that explanation; its data do not test it.

Which table answers the question?

§ 6 of 8

The data cannot say. Simpson (1951)³ showed this with one set of numbers and two stories. In an artificial example, an investigator checks a pack of cards, some of them dirty because a baby had played with them, and finds “a positive association between redness and plainness both among the dirty cards and among the clean”; the combined table shows none, and it is the combined table “which provides what we would call the sensible answer, namely, that there is no such association” (pp. 240 to 241). Simpson then renames the classes: male and female, untreated and treated, alive and dead. “The probabilities are exactly the same as in Table 2”, and now there is a positive association between treatment and survival among males and among females but none in the combined population. This time the split tables give the sensible answer: “The treatment can hardly be rejected as valueless to the race when it is beneficial when applied to males and to females.” (p. 241)

Pearl draws the conclusion. Lindley and Novick, he reports, showed “that there was no statistical criterion” to warn the investigator against the wrong conclusion or to indicate which data held the correct answer; with the very same data, “our decision must be driven not by statistical considerations” but by additional information taken from the context (pp. 8 to 9). In his reading, which follows Hernán, Clayton and Keiding, Simpson’s³ two examples differ in causal structure: in the medical version sex was a common cause of the other two variables, while in the cards a card’s being dirty was a common effect of its colour and its face type (p. 9, note 4).

His criterion for a single third variable Z , the back-door criterion, is read off a causal diagram and has two conditions: “ Z is not a descendant of X ”, which means the treatment X cannot change it, and “ Z blocks every path that ends with an arrow into X ”. When both hold, the split tables give the effect of X on the outcome Y ; when no such path needs blocking, the pooled table does (p. 11). A common effect behaves the other way round: conditioning on it opens a path instead of closing one, the general phenomenon known as Berkson’s paradox, by which, in Pearl’s² example, “the outcomes of two independent coins are rendered dependent by the testimony that at least one of them is a tail” (p. 11). The [selection bias guide](#) treats conditioning on such a shared consequence as one pathway to [selection bias](#).

Whether the treatment can change the third variable matters, but it does not settle the choice by itself. In Pearl’s² blood pressure example the drug affects blood pressure, and for the question of which drug is

more effective, assuming no unmeasured confounders, his answer is to ignore blood pressure and read the pooled data (p. 12). Yet in another of his diagrams the third variable “is not affected by the treatment, and still, it should not be controlled for” (p. 12). And in some models the correct answer “may not lie in either the disaggregated or the aggregated” data, because the third variable is not enough to block a path into the treatment and further covariates may be needed (p. 11).

On this guide’s reading, Berkeley shows that the question decides as much as the diagram. Men and women sought entry to different departments, so if sex is the treatment, the department is downstream of it and fails Pearl’s² first condition for the total effect of sex on admission. Bickel et al.¹ asked a narrower question, whether the decision to admit was influenced by the applicant’s sex, and for that question the department is where the decision was made and is the thing to hold fixed, provided that men and women applying to the same department did not differ in their qualifications. That proviso is the authors’ assumption 1, without which, they write, any difference in acceptance by sex could be put down to qualifications (p. 398), and which they call clearly false in its widest sense (p. 402). Pearl notes that for queries other than total effects, such as how much of an effect runs through a third variable, “intermediate variables can of course provide useful” information (p. 12).

Pearl offers two routes to the right table: a mathematical one and, for readers who find simulation more convincing, a check by simulation (p. 10). In his guessing game, a challenger who knows the model behind the data tests the analyst’s choice of table against “the gold standard of a randomized trial, simulated on the model” (p. 11), and Pearl argues that modern causal tools “permit us to determine mathematically whether the aggregated” or the disaggregated findings should be accepted (p. 12). The commentary argues this from causal theory, with diagrams and a simulation game that it describes but does not run, not from data. The article it answers, by Armistead, held that third variables, causal or not, can convey critical information; Pearl agrees in general but replies that, for the total effect of the treatment, the right table can be accepted without further examination (p. 12).

What should be settled before two segments are pooled?

§ 7 of 8

Five things, written down before the tables are read: the question (which action, which outcome), the third variable that splits the data, whether that variable is linked both to the groups compared and to the outcome, whether the action can change it, and whether it closes every path into the action. The last two are Pearl's² conditions; the third follows Simpson's³ condition for when pooling cannot matter. The answer goes in a last column: the pooled table, the split tables, or neither until more is measured.

Figure 2 A pooling decision record

Write down the question, the third variable and the answers to three tests before choosing the pooled or the split table.

Question	Third variable	Linked to both	Moved by action	Paths closed	Table
Which action, which outcome?	What splits the data?	To the groups and the outcome?	Can the action change it?	Does it close every path in?	Pooled, split or neither?

Split if the action cannot change it and it closes every path in; pool if no path in is open; otherwise measure more.

One row per comparison. The third column screens with Simpson's condition: for his measure of association in a $2 \times 2 \times 2$ table without second order interaction, pooling cannot change the association unless the third variable is linked to both. The fourth and fifth are Pearl's two conditions for a single third variable. A path in is any path that ends with an arrow into the action; note in the fifth column whether one is open without the split, as well as whether the third variable closes each one.

Source Author's decision worksheet. The linked-to-both field follows Simpson (1951), p. 241; the moved-by-action and paths-closed fields follow Pearl (2014), p. 11; the decision follows Pearl (2014), pp. 11 to 12. The fields are reader inputs; no company data is supplied.

An illustration, with invented numbers and no company data: two sales teams grant discounts on small and large deals. Team A averages 10 per cent on its 30 small deals and 20 per cent on its 70 large ones;

team B averages 12 per cent on 80 small deals and 22 per cent on 20 large ones. Within each deal size team A discounts less, yet across all its deals it averages 17 per cent against team B's 14 per cent, because seven in ten of its deals are large. Suppose deals are routed to a team by their size when they are created, and nothing else that moves the discount decides the routing. Then deal size meets both of Pearl's² conditions, and the split comparison answers whether team A discounts more on comparable deals; in these numbers it does not. If team A also enlarges deals, by adding seats before it quotes, some of its large deals are large because of the team. Deal size is then partly downstream of the team, and the split table leaves out the part of the team's effect that runs through deal size. Recording the size at routing, before either team has touched the deal, removes that problem only if size was the one thing routing used that also moves discounts. These are the author's examples and have not been tested.

Which questions do these sources leave to the analyst?

§ 8 of 8

How often reversals occur in commercial data, and how to choose a table when the causal story is disputed. Simpson (1951)³ is a four-page note on contingency tables whose examples are artificial. Bickel et al. (1975)¹ analyse one university's admissions records for one cycle in depth, with campus-wide checks for 1969 to 1973, under an assumption of equal qualifications that they themselves call false in its widest sense. Pearl (2014)² is a commentary on another article: its criteria come from the causal theory it cites and are shown with diagrams and a simulation game it describes but does not run, not with new data, and when one third variable is not enough, the choice of further covariates lies beyond the scope of the note, Pearl says (p. 11).

On this guide's reading, two things carry over to pricing and sales data with some confidence. Any rate or average pooled across segments with different mixes can reverse, so a pooled comparison is a statement about mix as well as about the groups. And the choice between the pooled and the split comparison is a causal decision that has to be argued before the numbers are read, not found in them. Whether a given discount, win rate or renewal comparison reverses, and which table is right for it, the sources do not say.

 End of essay

References

- 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>

How to cite this essay

Isoglu, S. (2026, October 18). What is Simpson's paradox? Sinan Isoglu. <https://isoglu.com/insights/journal/what-is-simpsons-paradox/>
Online version: <https://isoglu.com/insights/journal/what-is-simpsons-paradox/>

Sinan Isoglu, MBA (Quantic)

Commercial growth leader, lecturer and doctoral researcher
© 2026 Sinan Isoglu · isoglu.com · All rights reserved