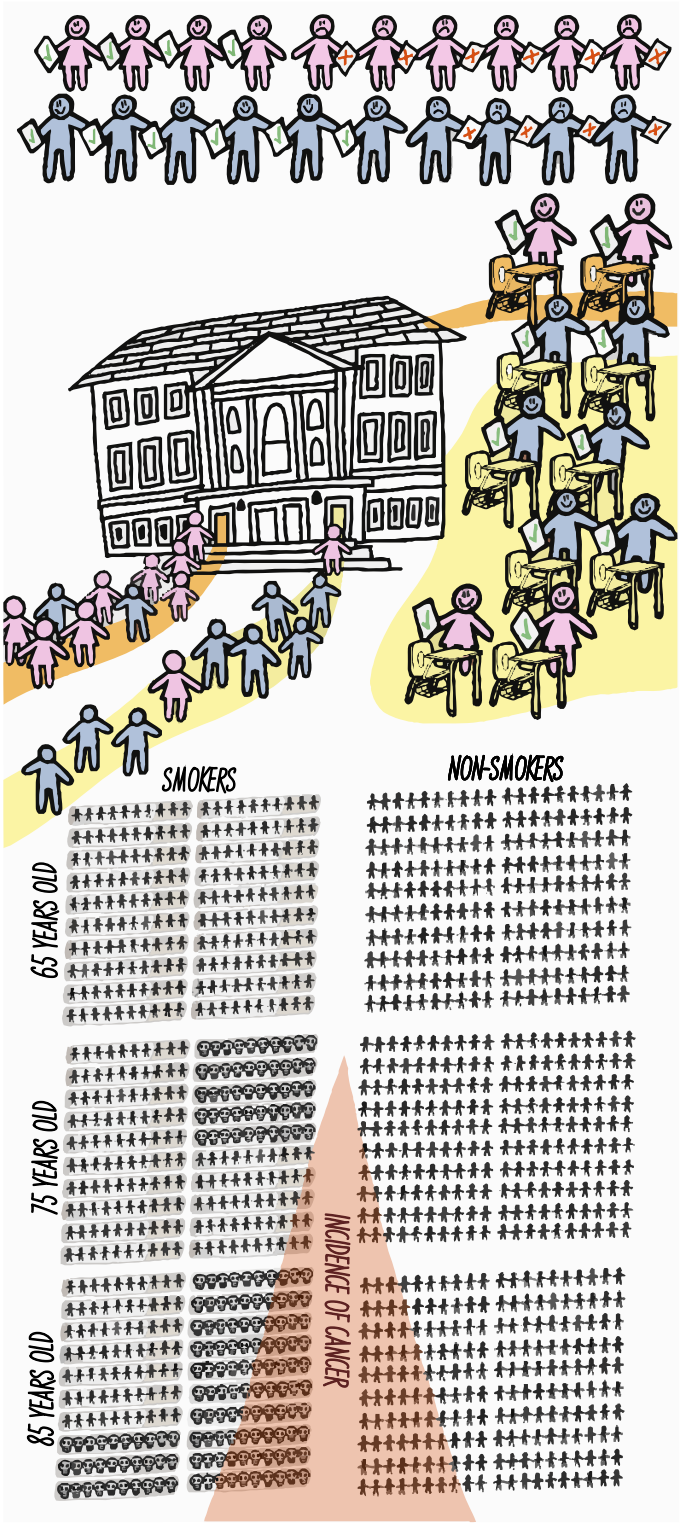
Divide and Conquer

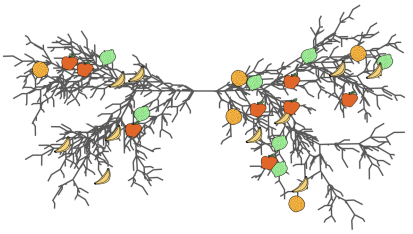
Top (Simpson's Paradox – Admissions): At the aggregate level, a higher percentage of men are admitted compared to women, suggesting gender bias. But disaggregating by department reveals that women applied more often to highly competitive departments with low acceptance rates, while men applied to departments with higher admission rates. Within departments, women were often admitted at equal or higher rates. This reversal — where an apparent bias in subgroups contradicts the overall trend — is known as *Simpson's Paradox*.

Bottom (Survivorship Bias – Cancer in Smokers): Cancer incidence increases sharply with age, but smokers have higher early mortality from other causes. As a result, fewer smokers survive into the high-risk age brackets where cancer becomes common. This skews the population-level data, making it seem as though smokers have lower cancer rates than non-smokers. The truth is hidden by *survivorship bias*: smokers often die before cancer can occur.



Divide and Conquer

Simpson's Paradox occurs when a statistical trend present in separate groups reverses when the groups are combined. This effect is a result of unequal group sizes or hidden confounding variables that distribute non-uniformly across the data. For example, a treatment might show positive effects in both male and female subgroups yet appear harmful in the aggregate population if the treatment is disproportionately given to patients with more severe conditions and males and females differ in average severity. The apparent paradox demonstrates that causal inference requires careful consideration of the causal relationship rather than relying solely on raw correlations.



Simpson's Paradox ∘ Correlation Reversal ∘ Weighted Averages ∘ Kidney Stone Example ∘ Gerrymandering Math ∘ UC Berkeley Bias ∘ Subgroup vs Aggregate ∘ Confounding Variables ∘ Statistical Grouping ∘ North Carolina 2012 ∘ Data Partitioning

"If you torture the data long enough,
it will confess to anything."
— Ronald H. Coase

"Correlation is not causation,
but they keep close company."
— Huw Price, 2026

Divide and Conquer

Early statisticians, including G. Udny Yule (1903) and Karl Pearson, documented reversals that occur when aggregated data obscure subgroup relationships. In several early case studies, trends within groups differed from the pooled trend, foreshadowing the formal statement later articulated by Simpson.

Edward H. Simpson published a four-page paper in 1951 in the Journal of the Royal Statistical Society. He proved that for any proportions $a/b < c/d$ and $e/f < g/h$, it remains algebraically possible that $(a+e)/(b+f) > (c+g)/(d+h)$. Simpson's formulation established the general conditions under which combining data groups reverses their individual trends.

The paradox gained widespread attention through the 1973 Berkeley graduate admissions analysis by P. J. Bickel, E. A. Hammel, and J. W. O'Connell. Raw data showed women had a 35% acceptance rate compared to 44% for men. Department-specific analysis revealed no discrimination — women had equal or higher acceptance rates in four of six departments. Women applied more frequently to highly competitive departments with much lower acceptance rates, while men applied more to less selective departments. The differing application patterns created the aggregate disparity.

Simpson's result appears in medical trials when treatments are assigned based on patient severity, in election analysis when votes are aggregated by district, and in machine learning when training sets are partitioned. Colin R. Blyth coined the term “Simpson's paradox” in 1972, though Udny Yule had described similar reversals in 1903. The phenomenon is sometimes called the Yule-Simpson effect.

The same mathematics that creates accidental reversals enables deliberate manipulation through gerrymandering. Elbridge Gerry signed a redistricting bill in Massachusetts in 1812 that created a salamander-shaped district to concentrate opposition voters, giving the practice its name. The underlying strategy of manipulating representation, however, has deeper roots in British “rotten boroughs”, which were used to control elections by packing (concentrating them in a few districts) or cracking (splitting them across many) voters.

With the rise of computer-assisted mapping after the 1990 census, gerrymandering became a science, allowing partisan mapmakers to secure durable advantages in evenly divided electorates. After a mid-decade redistricting in Texas in 2003 reshaped the congressional balance of power, courts increasingly employed mathematical diagnostics — such as the efficiency gap, mean-median difference, and ensemble simulation tests — to determine when district boundaries purposefully waste opposition votes through packing or cracking.

In North Carolina’s 2012 congressional elections, Democratic candidates received about 51% of votes statewide yet won only 4 of 13 seats. Republicans secured 9 seats with roughly 49% of votes. The reversal occurred through district boundaries — lines drawn to group voters in ways that inverted the relationship between votes and representation.

©2026 David H. Silver, CC BY-NC 4.0

Statistical association hinges on how data are grouped. The same population can yield opposite conclusions depending on the partition chosen. In extreme cases, a relationship positive in every subgroup becomes negative when groups combine — or a democratic majority becomes a legislative minority through strategic line-drawing.

Simpson's paradox demonstrates correlation reversal (Simpson, 1951). A kidney stone treatment shows 93% success for small stones and 73% for large stones. A competing treatment achieves only 87% for small stones and 69% for large stones. The first treatment beats the second in both categories. Yet overall success rates reverse: 79% versus 85.5%. The reversal arises because doctors used the superior treatment primarily on difficult cases — 70% of its patients had large stones, while 91% of the inferior treatment's patients had small stones.

Gerrymandering engineers deliberate reversal. Both major parties employ this tactic when they control redistricting. Wisconsin's 2012 state assembly elections saw Democrats win 53% of votes but only 39% of seats. The mechanism: district lines packed Democratic voters into urban districts where they won by 70–80% margins, while Republican victories spread efficiently across suburban and rural districts with 55–60% margins. Both phenomena — Simpson's paradox and gerrymandering — exploit the mathematics of aggregation, but with different intent. The mathematics underlying both phenomena reduces to weighted averages. When calculating any aggregate statistic — whether treatment success rates or electoral outcomes — the result depends on two factors: the values within each group and the relative sizes of groups. Change either factor and the aggregate changes.

In formal terms, if groups have success rates $p_1, p_2, ..., p_k$ and sizes $n_1, n_2, ..., n_k$, the overall rate is:

$$\bar{p} = \frac{\sum_{i=1}^{k} n_i p_i}{\sum_{i=1}^{k} n_i}.$$

Simpson's paradox occurs when natural imbalances in group sizes (n_i) cause $\bar{p}$ to misrepresent the relationship seen in individual p_i values. Gerrymandering manipulates the same formula by choosing group boundaries to engineer specific (n_i) values.

The kidney stone example illustrates the reversal:

	Success Rate	Patients	Successes
Treatment A, Small stones	93%	30	28
Treatment B, Small stones	87%	200	174
Treatment A, Large stones	73%	70	51
Treatment B, Large stones	69%	20	14
Treatment A, Overall	79%	100	79
Treatment B, Overall	85.5%	220	188

Treatment A wins in both stone categories yet loses overall. Treatment A handled 70% difficult cases (large stones), Treatment B only 9%. When groups combine, Treatment B's easy-case bias overwhelms its inferior performance.

Gerrymandering employs similar mathematics with manipulative intent. Consider a simplified state with 10 districts, 5 million voters split evenly between parties:

District	Voters (A)	Voters (B)	Winner	Strategy
1	20%	80%	B	Packed (B stronghold)
2	22%	78%	B	Packed (B stronghold)
3	18%	82%	B	Packed (B stronghold)
4	56%	44%	A	Cracked (A edge win)
5	55%	45%	A	Cracked (A edge win)
6	54%	46%	A	Cracked (A edge win)
7	57%	43%	A	Cracked (A edge win)
8	53%	47%	A	Cracked (A edge win)
9	56%	44%	A	Cracked (A edge win)
10	55%	45%	A	Cracked (A edge win)
Totals	**50%**	**50%**	**A wins 7, B wins 3**	Gerrymandered for A

In the UC Berkeley admissions case (1973), similar patterns appeared. Women showed lower overall acceptance rates (35%) than men (44%), suggesting discrimination. At the department level, women had equal or higher acceptance rates in 4 of 6 departments. The reversal occurred because women disproportionately applied to competitive departments — English admitted 3.4% of applicants while Engineering admitted 65%.

Simpson's paradox and gerrymandering exploit the disagreement between local and global measures. In Simpson's paradox, local measures (department-specific admission rates) tell the truth while global measures (overall rates) mislead. In gerrymandering, local measures (district-level victories) are manipulated to distort global truth (statewide voter preference).

For any partition of data into groups, the overall average equals:

$$\bar{y} = \sum_i w_i \bar{y}_i$$

where $w_i = n_i/N$ represents the fraction of data in group i, and $\bar{y}_i$ is that group's average.

Simpson's paradox exposes existing groupings in the data; gerrymandering constructs groupings to exploit the same arithmetic.

The efficiency gap quantifies gerrymandering's success (Stephanopoulos & McGhee, 2015) by measuring "wasted" votes — those beyond the 50%+1 needed to win a district. A party that wins districts by slim margins while losing others by wide margins achieves maximum efficiency. The formula:

$$\text{Efficiency Gap} = \frac{|\text{Wasted}_A - \text{Wasted}_B|}{\text{Total Votes}}.$$

Values around 7–8% have been proposed in the political science literature as a heuristic threshold for durable advantage; courts have not adopted a single standard, and experts treat it as one indicator among others. Thus, gerrymandering leaves fingerprints. Districts snake through neighborhoods, splitting cities and joining disparate communities. Pennsylvania's

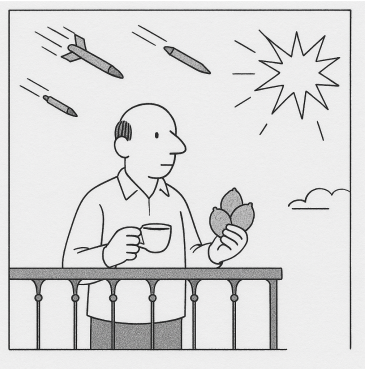
7th district (pre-2018) stretched like tentacles across five counties to link Republican areas while avoiding Democratic ones. Maryland's 3rd district exhibits similar contortions, engineered by Democrats to dilute Republican votes across Baltimore suburbs.

Simpson's paradox is revealed through careful analysis. Statisticians discover reversals by examining subgroups. Early COVID-19 comparisons illustrated how age structure can confound: countries with older populations showed higher overall death rates even when age-specific rates were comparable. Proper age standardization is necessary before drawing conclusions.

Simpson's paradox warns that natural parameters (patient severity, department selectivity) can mislead when ignored. Gerrymandering demonstrates that artificial boundaries can be weaponized to subvert democratic representation.

Solutions to Simpson's paradox require disaggregating data and examining subgroups. Medical trials now routinely report results by patient characteristics. Universities analyze admissions by department.

The solution to gerrymandering requires judicial reform: independent redistricting commissions, mathematical constraints on district compactness, or algorithmic districting that minimizes partisan advantage. Several states now use efficiency gap calculations in legal challenges to districting plans.



What are the odds a bomb hits the only person holding three guavas?

More Statistical Paradoxes and Interpretation Failures

1. Berkson's Paradox *Conditioning on a common effect induces spurious negative (Berkson, 1946) correlation.* If two independent variables both affect a selection criterion, then restricting attention to cases that satisfy that criterion creates an artificial negative correlation. This occurs in hospital datasets, where independent risk factors may appear inversely related when conditioned on admission. The association is real in the conditional data but does not reflect a relationship in the population.

2. Ecological Fallacy *Group-level associations are wrongly projected onto individuals (Robinson, 1950).* When a statistical association holds across aggregated units — such as regions or schools — it does not necessarily hold within them. For example, a country with higher average education may have higher average income, but this does not imply that more educated individuals earn more within each region. Unlike Simpson's paradox, ecological fallacy involves misapplying group-level trends to individual inference without requiring any reversal. The error lies in cross-level extrapolation, not confounding.

3. Will Rogers Phenomenon *Reclassification improves group averages without improving any member.* If individuals from the low end of one group are reclassified into another group with an even lower average, both groups may show improved mean outcomes. This occurs in cancer staging and school performance tracking, and reflects the fact that averages are sensitive to how groups are defined.

4. Modifiable Areal Unit Problem (MAUP) *Statistical results depend on the choice of spatial or administrative (Openshaw, 1984)ative boundaries.* In spatial analysis, correlations and rates can shift significantly depending on how geographic regions are aggregated. A pattern observed at the county level may not hold at the district level or when boundaries are redrawn.

5. Low Birth-Weight Paradox *Conditioning on an intermediate variable reverses risk comparisons.* Infants born to smoking mothers have higher rates of low birth-weight, and low birth-weight is associated with higher mortality. But among low birth-weight babies, those born to smokers may show lower mortality than those of non-smokers. The paradox appears because birth-weight is both an effect of smoking and a predictor of mortality. Conditioning on it introduces collider bias, obscuring causal direction.

6. Prosecutor's Fallacy *Confusing the likelihood of evidence with the probability of guilt.* In forensic contexts, the probability of observing the evidence assuming innocence is often mistaken for the probability of innocence given the evidence. For example, a DNA match with a false positive rate of $1/1000$ is incorrectly interpreted as implying a 0.1% chance of innocence, ignoring base rates. The fallacy reflects improper inversion of conditional probability.

Two Mathematical Realizations of Simpson's Paradox

Reversal in Pearson Correlation

Suppose two subgroups yield:

$$\begin{aligned}\text{Corr}(X, Y \mid Z = 1) &= +0.8,\\ \text{Corr}(X, Y \mid Z = 2) &= +0.7,\end{aligned}$$

yet the marginal correlation is:

$$\text{Corr}(X, Y) = -0.3.$$

This reversal can occur when the subgroup means oppose each other:

$$\begin{aligned}\mathbb{E}[X \mid Z = 1] &\ll \mathbb{E}[X \mid Z = 2],\\ \mathbb{E}[Y \mid Z = 1] &\gg \mathbb{E}[Y \mid Z = 2].\end{aligned}$$

The total covariance decomposes as:

$$\begin{aligned}\text{Cov}(X, Y) = \mathbb{E}[\text{Cov}(X, Y \mid Z)]\\ + \text{Cov}(\mathbb{E}[X \mid Z], \mathbb{E}[Y \mid Z]).\end{aligned}$$

The first term represents the true structural relationship. The second term arises from between-group mean shifts. When subgroup trends are consistent but means shift in opposite directions, this second term can dominate and flip the sign.

In such cases, the subgroup correlation reflects the actual relationship between the variables. The marginal correlation is an artifact of mixed distributions and should not be used to infer these relationships.

Why Subgroup Correlations Reflect Structure

The Pearson correlation coefficient assumes a homogeneous population. When data consist of subgroups (e.g., children vs adults), the overall correlation reflects two effects:

- the correlation within each group,
- the shift in means across groups.

This decomposes as:

$$\begin{aligned}\text{Corr}(X, Y) = \text{within-group structure} +\\ \text{between-group shift}.\end{aligned}$$

If the subgroups differ in both $\mathbb{E}[X \mid Z]$ and $\mathbb{E}[Y \mid Z]$, the between-group term may dominate and flip the marginal sign — even if each group has a positive internal trend.

Subgroup correlations hold Z fixed and reveal how X relates to Y when background is controlled. The marginal correlation, in contrast, entangles structure with population imbalance.

For variable relationships inference — e.g., how height relates to foot size, or how score relates to study time — $\text{Corr}(X, Y \mid Z)$ provides the interpretable relationship. The marginal $\text{Corr}(X, Y)$ may be distorted by mixing.

Visual example: Imagine both kids and adults show that larger plates come with higher calorie counts. But if kids mostly use small plates and pile them with calorie-dense snacks, while adults take large plates but fill them with vegetables, the overall data may suggest that smaller plates correspond to higher calories. This reflects sample composition, not individual-level relationships.

How Likely is Simpson's Paradox?

Pavlides and Perlman (2009) studied how often Simpson's paradox arises in $2 \times 2 \times 2$ contingency tables. Under a uniform distribution over all such tables, they showed:

$$\boxed{\text{1 in 60 tables exhibits a reversal.}}.$$

This corresponds to a prior probability of ≈ 0.0166 that conditional trends align while the aggregate trend opposes them.

The paradox becomes rarer with more subgroups; under similar uniform assumptions, the chance decreases further as the number of conditioning groups increases.

References:

Simpson, E. H. (1951). *J. R. Stat. Soc. B*, 13(2), 238–241.

Pavlides, M. G., and Perlman, M. D. (2009). *J. Stat. Plan. Inference*, 139(1), 198–213.