When Numbers Mislead Us

Arthur Charpentier

When Numbers Mislead Us

Arthur Charpentier

Université du Québec à Montréal
charpentier.arthur@uqam.ca

Believing that there is a single, objective way to describe phenomena using numbers is to forget that data does not "speak" for itself. Collecting data involves making choices: what to measure, how, when, on whom, etc. This implies implicit (or even ideological) assumptions about what counts as a measurable fact. And in any data analysis, what is not measured can be as important as what is observed. When an influential variable is overlooked—whether ignored, neglected, or simply unknown—the apparent relationships between other variables can become misleading. This is known as "omitted variable bias": a hidden effect distorts comparisons and can make a correlation appear where there is none, or mask a real one. Sometimes, introducing this "forgotten" variable can even completely reverse the conclusions that would have been drawn from a naive reading of the data.

1 Rankings Hospitals, Doctors, Schools, etc.

In an era of widespread rankings—from the best hospitals to the best doctors, including schools, as frequently found in the press—performance indicators are playing an increasingly important role in public decisions and social perceptions. But as Goodhart's law states, "when a measure becomes a target, it ceases to be a good measure." When a hospital seeks to climb the rankings, it may do so not by actually improving the quality of care, but by acting on the levers that influence the indicators: admitting less serious patients, avoiding complex cases, or optimizing its discharge statistics. The risk is then to produce a biased image of efficiency, disconnected from medical reality. Consider the following (entirely fictional) example, from Table 1, with two hospitals of the same size, very loosely based on a famous example of kidney stone treatment (with real data) that appeared in the British Medical Journal in the 1980s. Based, on a quick comparison of mortality rates in two hospitals, Hospital

A has better outcomes within both subgroups (healthy and non-healthy patients), its overall death rate is higher than Hospital B's.

	hospital A				hospital B		
	total	death	ratio		total	death	ratio
non-healthy	60	36	60%	$<$	20	14	70%
healthy	20	4	20%	$<$	60	18	30%
total	80	40	50%	$>$	80	32	40%

Table 1: Mortality rates by hospital and patient condition (source: author).

2 Not Really a Paradox...

In 1951, British statistician Edward Simpson published an article on contingency table analysis, in Simpson [1951]. In it, he described a phenomenon whereby the relationship between two variables changes when a third hidden variable is taken into account. He did not consider this a paradox, but rather a statistical effect worth noting. It was not until the 1970s that this behavior was rediscovered, named in his honor, and elevated to the status of a "paradox" because it defies our intuition. Despite its name, Simpson's paradox is not a paradox in the mathematical sense, as it is not based on any logical contradiction or error in formal reasoning. Rather, it is a statistical surprise linked to our intuition about averages. In reality, this phenomenon occurs when ratios calculated in subgroups (e.g., recovery rates in two categories of patients) are compared and then aggregated without taking into account the structure of the data (group size, distribution of cases). The misinterpretation stems from the fact that the overall ratio is not the average of the ratios: in general, percentages from groups of different sizes cannot be added together and used to derive a meaningful average without weighting them. In the previous example, for Hospital A, the total death rate is not the average of the two rates (otherwise we would have obtained 40%, the average of 60% and 20%) but 50%. And then we have a small counterintuitive property about fractions, namely that we can have

$$\frac{36}{60} < \frac{14}{20} \text{ and } \frac{4}{20} < \frac{18}{60}$$

but, at the same time,

$$\frac{36+4}{60+20} = \frac{40}{80} > \frac{14+18}{20+60} = \frac{32}{80}$$

	group A				group B		
	total	deaths	ratio		total	deaths	ratio
subgroup 0	a	b	$\frac{b}{a}$	$<$	A	B	$\frac{B}{A}$
subgroup 1	c	d	$\frac{d}{c}$	$<$	C	D	$\frac{D}{C}$
total	$a+c$	$b+d$	$\frac{b+d}{a+c}$	$>$	$A+C$	$B+D$	$\frac{B+D}{A+C}$

Table 2: Formal version of Table 1, with counts, by hospital and patient condition.

More formally, in Table 2, we have an abstract representation of the paradox (inspired from Table 1). Each group (A and B) is split into two subgroups. Even though group B shows worse ratios in both subgroups, its overall rate is better—due to differences in the distribution of totals across groups.

A graphical depiction of the paradox: line slopes represent subgroup ratios. In both cases, group B's ratios are higher, but once aggregated, group A's overall slope becomes steeper—demonstrating the paradox in action.

xxx

$$\frac{b}{a} < \frac{B}{A} \text{ and } \frac{d}{c} < \frac{D}{C}$$

but, at the same time,

$$\frac{b+d}{a+c} > \frac{B+D}{A+C}$$

These inequalities can be seen in Figure 1. The segment connecting (20,4) from the origin (0,0), in red, has a slope of $4/20$, which is below the segment connecting (60,18), because its slope of $18/60$ is greater. The same applies to the red segment connecting (60,36), which is below the blue segment connecting (20,14) because $14/20 > 36/60$. However, when we aggregate, we obtain the points on the right, (80,32) in blue and (80,40) in red, and the blue slope is lower than the red slope.

This discrepancy between the overall ratio and the partial ratios fuels many well-known "paradoxes" in mathematics concerning averages: harmonic mean vs. arithmetic mean, average speed, or even the "average of averages." Simpson's paradox is a sophisticated version of the latter case: it reminds us that a misinterpreted average can not only be misleading, but can completely reverse a trend. It is therefore not a mathematical oddity, but an invitation to caution in the interpretation of aggregates.

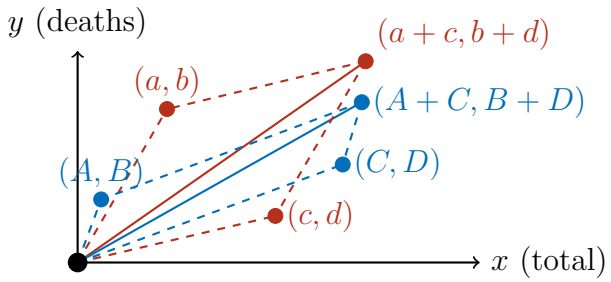

Figure 1: Geometric visualization of Simpson's paradox, based on Figure 2.

3 Berkeley Admissions and Gender Bias: A Real Case

A famous case highlighted this pitfall: in 1973, the University of California, Berkeley was accused of discriminating against women in graduate admissions. The overall figures were unequivocal: 44% of men were admitted, compared to only 35% of women—an alarming difference that was enough to trigger a lawsuit. This significant difference raised suspicions of gender discrimination in the selection process. However, upon closer examination of the data, something surprising emerged: in most departments, the admission rate for women was equal to or higher than that for men. The data provided by Peter Bickel, Eugene Hammel, Eugene, and William O'Connell shows the same kind of phenomenon observed in hospitals, in Bickel et al. [1975] or Bickel et al. [1977].

Table 3 reveals that in four of the six departments, women had a higher admission rate than men. Nevertheless, their overall admission rate was lower—a seemingly contradictory outcome. Why? Because women applied more to the most competitive departments, where admission rates are low for everyone. Men, on the other hand, applied in greater numbers to departments with high admission rates. As a result, the overall rates give an impression that is the opposite of what we get when analyzing each subgroup. Once the data are properly stratified, Bickel and his colleagues even conclude that there is a slight bias in favor of women: "If the data are properly pooled... there is a small but statistically significant bias in favor of women." What this example shows is that overall averages can mask structural effects. One might think that women are disadvantaged, when in fact they have simply chosen more competitive fields. This is the crux of Simpson's paradox: it teaches us to be wary of unadjusted overall comparisons. To fully understand a statistical situation, it is often necessary to analyze the data according to relevant subgroups, rather than

	Men				Women		
	Total	Admitted	Rate		Total	Admitted	Rate
A	825	512	62.1%	<	108	89	82.4%
B	560	353	60.0%	<	25	17	68.0%
C	325	120	36.9%	>	593	202	34.0%
D	417	138	33.1%	<	375	131	34.9%
E	191	53	27.7%	>	393	94	23.9%
F	373	22	5.9%	<	341	24	7.0%
Total	2691	1198	44.5%	>	1835	557	30.4%

Table 3: UC Berkeley graduate admissions, from Bickel et al. [1975] (real admission data).

simply looking at the aggregate figures. This paradox is therefore an excellent lesson in scientific rigor: understanding is not just about observing numbers, it is about knowing what they tell us—and what they hide.

4 Two More Examples of the Paradox

In 1986, demographer Joel Cohen studied mortality rates in Costa Rica and Sweden—two very different countries, with Sweden renowned for its high life expectancy, in Cohen [1986]. However, he noted a surprising fact: the overall mortality rate is higher in Sweden (9.29‰) than in Costa Rica (8.12‰). This would be surprising in itself, but another detail reinforces the paradox: in each age group, Sweden has a lower mortality rate than Costa Rica, as shown in Figure 2. How is this possible?

The answer lies in the age structure of the populations. Comparing the age pyramids, we find that Costa Rica had a very young population at the time—half of the inhabitants were under 20—while Sweden had a much higher proportion of older people. However, as the probability of dying increases with age, an older population will automatically have a higher overall mortality rate, even if it is healthier at each age. So, if you happen to meet a woman on the street in Costa Rica, she is statistically less likely to be elderly—and therefore less likely to die within the year—than if you meet a woman in Sweden. It is not that Swedish women die more often: it is that the Swedish women you meet are, on average, older. That is the subtlety of the

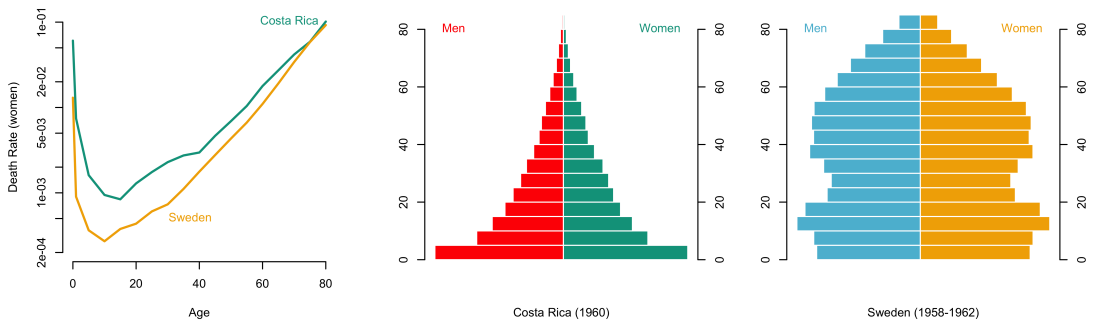

Figure 2: Age-specific mortality rates and population structures in Costa Rica and Sweden, from Charpentier [2024]. The left panel shows the annual probability of death (on a logarithmic scale) for women by age: at every age, mortality is lower in Sweden than in Costa Rica. The center and right panels display population pyramids from the same period, revealing that Costa Rica had a much younger population, while Sweden had a higher proportion of older individuals. This explains why Sweden's overall mortality rate was higher despite having lower age-specific death rates—a striking example of Simpson's paradox in demography.

paradox.

In 2004, Gary Davis, a researcher at the University of Minnesota, looked at the relationship between average vehicle speed and the number of accidents involving pedestrians in different neighborhoods of a city, in Davis [2004]. At first glance, the results of his model were puzzling: they seemed to show that reducing the speed limit from 30 to 25 miles per hour led to an increase in the number of accidents. Such a conclusion goes against intuition, the results of controlled experiments, and road safety logic. But upon closer inspection, Davis identified an interpretation error due to data aggregation—a typical effect of Simpson's paradox. The main problem was that residential areas, which naturally have less traffic and therefore fewer accidents, were also the most likely to adopt the new 25 mph speed limit. As a result, 30 mph streets remained associated with higher traffic volumes. When comparing all streets globally, without taking these contextual differences into account, it appears that the reduction in speed is linked to an increase in accidents — when in reality, it is simply that very different areas are being compared. The model did not take into account this confounding variable: the nature of the neighborhood (residential or non-residential), which determines both traffic volume and accident risk.

The Ecological Inference Paradox

Simpson's paradox can be formulated as follows: "*what is true in each group may be false in the whole*." A reverse interpretation is also possible: "*what is true at the group level is not necessarily true at the individual level*." This is often referred to as the "*ecological inference paradox*," which originated in an article published in 1950 by American sociologist William S. Robinson (who was the first to use the term "*ecological inference paradox*," in Robinson [1950]). Robinson was studying the correlation between immigration rates and literacy rates in the United States, comparing data by state. He observed that at the state level, the higher the rate of immigration, the higher the average literacy rate. But when looking at individual data, the opposite was true: immigrants were, on average, less literate than non-immigrants. In other words, the correlations observed at the group level (countries, regions, social classes, etc.) can be very different—even opposite—from those observed at the individual level. The paradox arises when we infer (or draw conclusions) about individuals from aggregated data, which is often a mistake. In this context, the term "ecological" does not refer to environmental ecology, but to analysis by "collective units," such as social groups, geographical areas, or institutions—as opposed to individual units.

To illustrate, we can look at the two examples of Figure 3, which show, for a large number of countries around the world, gross domestic product per hour worked as a function of per capita coffee consumption (left hand side); or life expectancy at birth as a function of per capita cigarette consumption per year (right hand side).

In both figures, there is a positive correlation at the aggregate level, by country. But it would be risky to claim that the link observed between the variables at the country level reflects a similar relationship at the individual level. In the first case, the more coffee a country consumes per capita, the higher its GDP per hour worked. Does this suggest that drinking coffee makes you productive? No, it does not mean that an employee who drinks more coffee will be more efficient individually. It is even possible that within a country, heavy coffee drinkers have a comparable or even lower level of productivity than those who do not drink coffee. The explanation is quite simple: coffee is consumed more in rich industrialized countries, where working conditions, automation, technology, and work organization explain productivity. Coffee, here, is just a cultural marker (present in Northern countries, for example), correlated with wealth, but not causal. In the second case, the more cigarettes a country consumes (per capita), the higher the life expectancy seems to be. This seems to contradict everything we know about the effects of tobacco on health! Individually, smoking reduces life expectancy. Again, in these data, the countries that consume

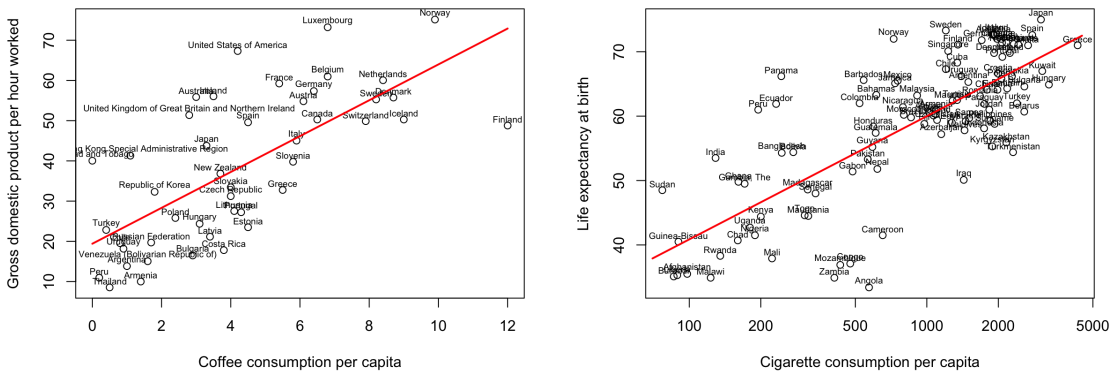

Figure 3: Cigarette consumption vs. life expectancy (left) and coffee consumption vs. labor productivity (right, per country,)

the most cigarettes were also, in general, the richest countries (Japan, France, Greece, etc.), which have better healthcare systems, better nutrition, better living conditions, etc. Conversely, countries with low cigarette consumption are often poorer, and life expectancy there is reduced for structural reasons (access to healthcare, infant mortality, wars, etc.). We therefore observe a positive correlation at the country level, but the individual effect of tobacco remains negative.

5 Consequences

These examples remind us that aggregating data can mask or reverse trends. When an important variable is ignored, the overall results become misleading. This is known as omitted variable bias, or confounding variable bias. The consequences are not trivial: they can lead to wrong decisions, or even miscarriages of justice, as Cathleen O'Grady points out in 2023, in O'Grady [2023], referring to the case of Lucia de Berk (convicted of murdering five children) and the report published by the Royal Statistical Society in 2022 (in Green et al. [2022]). Indeed, some nurses or interns, who mainly care for more seriously ill patients, are statistically associated with more deaths, not because they are at fault, but because their department treats more critical cases. This effect is a classic case of omitted variable bias: the severity of the patients' condition (the hidden variable) influences both the death and the assignment to the caregiver, completely distorting the perception of the mortality

rate.

References

Peter J Bickel, Eugene A Hammel, and J William O'Connell. Sex bias in graduate admissions: Data from berkeley: Measuring bias is harder than is usually assumed, and the evidence is sometimes contrary to expectation. *Science*, 187(4175):398–404, 1975.

Peter J Bickel, Eugene A Hammel, and J William O'Connell. Sex bias in graduate admissions: Data from berkeley. *Statistics and Public Policy*, pages 113–130, 1977.

Arthur Charpentier. *Insurance: biases, discrimination and fairness*. Springer Verlag, 2024.

Joel E Cohen. An uncertainty principle in demography and the unisex issue. *The American Statistician*, 40(1):32–39, 1986.

Gary A Davis. Possible aggregation biases in road safety research and a mechanism approach to accident modeling. *Accident Analysis & Prevention*, 36(6):1119–1127, 2004.

Peter Green, Richard D Gill, Neil Mackenzie, Julia Mortera, and William Thompson. *Healthcare Serial Killer or Coincidence? Statistical Issues in Investigation of Suspected Medical Misconduct*. Royal Statistical Society, 2022.

Cathleen O'Grady. Unlucky numbers. *Science*, 379(6629):228–233, 2023.

WS Robinson. Ecological correlations and the behavior of individuals. *American Sociological Review*, 15(3):351–357, 1950.

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