Simpson's Paradox

Ivan Koswara, Ananya Aaniya, Pi Han Goh, and 5 others contributed

Relevant For...
Logic > Paradoxes in Probability

Simpson's paradox occurs when groups of data show one particular trend, but this trend is reversed when the groups are combined together. Understanding and identifying this paradox is important for correctly interpreting data.

For example, you and a friend each do problems on Brilliant, and your friend answers a higher proportion correctly than you on each of two days. Does that mean your friend has answered a higher proportion correctly than you when the two days are combined? Not necessarily!

Day	You	Your friend
Saturday	$\frac{7}{8} = 87.5\%$	$\frac{2}{2} = \mathbf{100\%}$
Sunday	$\frac{1}{2} = 50\%$	$\frac{5}{8} = \mathbf{62.5\%}$
Total	$\frac{8}{10} = \mathbf{80\%}$	$\frac{7}{10} = 70\%$

This seemingly unintuitive possibility is referred to as *Simpson's paradox*.

Let's go back to our example on problem accuracy competition to see how this can occur.

- On Saturday, you solved 7 out of 8 attempted problems, but your friend solved 2 out of 2. You had solved more problems, but your friend pointed out that he was more accurate, since $\frac{7}{8} < \frac{2}{2}$. Fair enough.
- On Sunday, you only attempted 2 problems and got 1 correct. Your friend got 5 out of 8 problems correct. Your friend gloated once again, since $\frac{1}{2} < \frac{5}{8}$.

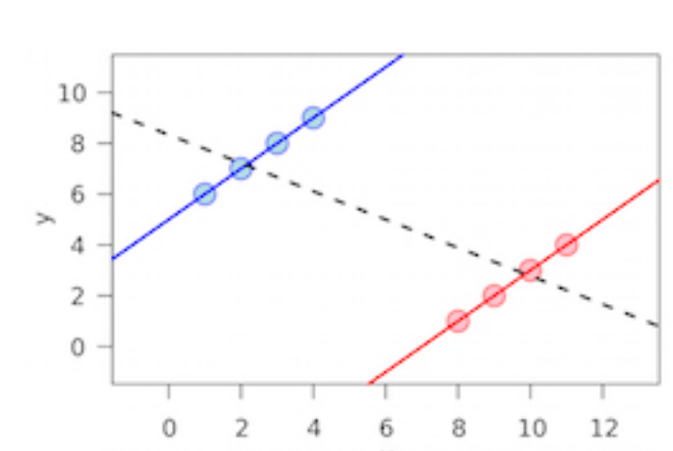
However, the competition is about the one who solved more accurately *over the weekend*, not on individual days. Overall, you have solved 8 out of 10 problems whereas your friend has solved 7 out of 10 problems. Thus, despite your friend solving a higher proportion of problems on each day, you actually won the challenge by solving the higher proportion for the entire weekend! While your friend got furious, you calmly pointed him to this page: you had just shown an instance of Simpson's paradox.

On this page, we'll give a formal definition of the paradox, show some interesting real-world examples, and provide an opportunity for you to add your own encounters with Simpson's paradox.

Contents

Definition

In layman's term, **Simpson's paradox** occurs when some groups of data show a certain relationship in each group, but when the data is combined, that relationship is reversed:



In the example above, we saw that when the problems were grouped into Saturday and Sunday, your friend solved a higher proportion correctly each day, but when the problems were combined into both days, you actually solved a higher proportion correctly.

This common form of Simpson's paradox can be defined as follows:

DEFINITION

Consider n groups of data such that group i has A_i trials and $0 \le a_i \le A_i$ "successes". Similarly, consider an analogous n groups of data such that group i has B_i trials and $0 \le b_i \le B_i$ "successes". Then, **Simpson's paradox** occurs if

$$\frac{a_i}{A_i} \le \frac{b_i}{B_i} \text{ for all } i = 1, 2, \ldots, n \text{ but } \frac{\sum_{i=1}^n a_i}{\sum_{i=1}^n A_i} \ge \frac{\sum_{i=1}^n b_i}{\sum_{i=1}^n B_i},$$

and at least one of the inequalities is strict (meaning that it is not in the equality case). Of course, we could also flip the inequalities and still have the paradox, since A and B are chosen arbitrarily. $\square$

To gain an intuition, let's see how this definition applies to our example above. There, the A_i/B_i are the number of problems you/your friend attempt on each day, and the a_i/b_i are the number you/your friend solve correctly on each day. As seen before,

$$\frac{7}{8} = \frac{a_1}{A_1} < \frac{b_1}{B_1} = \frac{2}{2} \text{ and } \frac{1}{2} = \frac{a_2}{A_2} < \frac{b_2}{B_2} = \frac{5}{8}, \text{ yet } \frac{7+1}{8+2} = \frac{a_1+a_2}{A_1+A_2} > \frac{b_1+b_2}{B_1+B_2} = \frac{2+5}{2+8}.$$

However, this is not the only way in which Simpson's paradox can occur. In general, Simpson's paradox is exhibited whenever there is a trend projected by individual categories of data, but the trend reverses when all categories are combined. While this template only considers binary "successes", where each individual datum contributes a single "yes" or a single "no", this can be easily generalized into numbers where the measure used for the trend is its average. We can even use some other measure (such as median). This is discussed in the section "Other Applications" below.

Now, let's try an example to test your ability to identify if Simpson's paradox is occurring!

TRY IT YOURSELF

You and your friend decide to spend the whole weekend doing a Brilliant marathon. The winner is whoever has answered the most number of questions right out of a massive set of 1500 problems at the end of the 2 day weekend.

On the first day, you answer 1200 questions, out of which about 62.2% were right. Your friend answers 700 questions, out of which about 63.6% were correct. Your friend teases you over the phone about how he has got a higher percentage of questions correct than you at the end of the first day.

On the second day, you answer 300 questions, out of which about 58.3% were correct. Your friend answers 800 questions, out of which he answers about 58.8% were correct. Once again, your friend teases you because the percentage of the number of questions he has got correct is greater than yours.

Who won this marathon?

- It's a draw.
- Your friend
- You

Why It Occurs

You've seen the results, but why does it occur?
Normally, one would expect that winning all groups means winning overall. However, this is only guaranteed to be the case if the group sizes are equal. When group sizes differ, the totals for each side might be dominated by particular groups, but these groups belong to different categories. In the introductory example above, the totals are dominated by the days where each player solved 8 problems, and in this case, you actually won $(7/8 > 5/8)$, which explains why you could win overall (when both days are combined). As an exaggerated example, consider a variant:

Day	You	Your friend
Saturday	$\frac{98}{99} = 98.99\%$	$\frac{1}{1} = \mathbf{100\%}$
Sunday	$\frac{0}{1} = 0\%$	$\frac{1}{99} = \mathbf{1.01\%}$
Total	$\frac{98}{100} = \mathbf{98\%}$	$\frac{2}{100} = 2\%$

The dominating groups are clearly those with 99 problems attempted. Those with 1 only affects the winner of the corresponding day; they barely do anything to the total.

When we line up the groups with equal sizes, we can see this paradox vanishing:

Size	You	Your friend
Big	$\frac{98}{99} = \mathbf{98.99\%}$	$\frac{1}{99} = 1.01\%$
Small	$\frac{0}{1} = 0\%$	$\frac{1}{1} = \mathbf{100\%}$
Total	$\frac{98}{100} = \mathbf{98\%}$	$\frac{2}{100} = 2\%$

Let's try another example to illustrate these ideas:

TRY IT YOURSELF

Two new drugs *AntiCynicismia* and *AntiMisantropia* are currently in a clinical trial phase, where the pharmacists determine whether the drugs are safe to use. The trial was divided into five distinct groups and the results of the experiment are as follows:

Drug name	AntiCynicismia	AntiMisantropia
Group A	436 out of 545 people were cured, or **80%** success rate	9 out of 10 people were cured, or **90%** success rate
Group B	245 out of 350 people were cured, or **70%** success rate	16 out of 20 people were cured, or **80%** success rate
Group C	48 out of 80 people were cured, or **60%** success rate	21 out of 30 people were cured, or **70%** success rate
Group D	10 out of 20 people were cured, or **50%** success rate	180 out of 300 people were cured, or **60%** success rate
Group E	2 out of 5 people were cured, or **40%** success rate	320 out of 640 people were cured, or **50%** success rate

It may seem like AntiMisantropia is a more effective drug based on the success rates on different groups. However, this is not true!

Given that $x\%$ is the difference in success rates between these two drugs, find the value of x.

Submit your answer

Other Applications

The above definition provides one common form of Simpson's paradox. However, it can occur in other ways.

The data, instead of "yes" and "no" (binary successes), may be arbitrary real numbers, and we can still have Simpson's paradox with the averages:

Category	Faction 1	Faction 2
Category 1	$6, 7, 8, 9 \to 7.5$	$10 \to \mathbf{10}$
Category 2	$0 \to 0$	$1, 2, 3, 4 \to \mathbf{2.5}$
Total	$0, 6, 7, 8, 9 \to \mathbf{6}$	$1, 2, 3, 4, 10 \to 4$

One real-world example using the average of real numbers occurs with income tax. Between 1974 and 1978, the U.S. tax rate decreased for every category of earning (under \$5000, \$5000-\$10000, etc). When aggregated across all of the people, however, the average tax rate increased!

The trend can also be the **median** instead of the average:

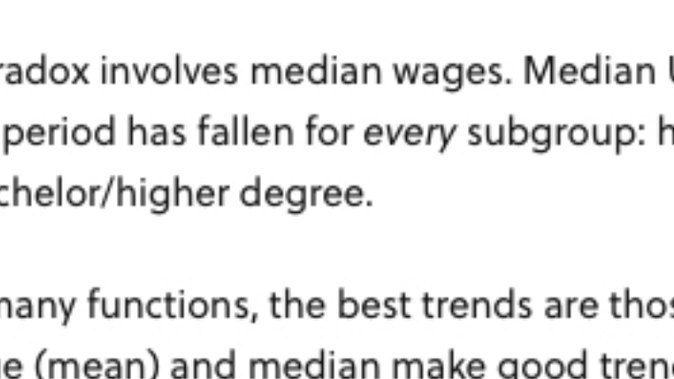
Category	Faction 1	Faction 2
Category 1	$6, 7, 8, 9 \to 7.5$	$10 \to \mathbf{10}$
Category 2	$0 \to 0$	$1, 2, 3, 4 \to \mathbf{2.5}$
Total	$0, 6, 7, 8, 9 \to \mathbf{7}$	$1, 2, 3, 4, 10 \to 3$

In fact, one real-world example of Simpson's paradox involves median wages. Median US wage between 2000 and 2012 has risen (about 1%). However, median US wage across the same period has fallen for *every* subgroup: high school dropouts, high school graduates but no college education, college education, and Bachelor/higher degree.

While technically the trend can be exhibited by many functions, the best trends are those that people wouldn't expect to be able to be reversed when the groups are combined. Average (mean) and median make good trends; these are likely counter-intuitive, which exactly explains the name paradox.

Real-World Examples

University of California Admission Rates



A study showed that, overall, men were accepted more than women (44% vs 35%). However, looking at each department, women were usually accepted at a rate equal to or higher than the rate at which men were accepted. What was happening? In fact, women tended to apply to departments which were harder to be admitted into.

Kidney Stone Treatment / Ambulances vs. Helicopters

Advanced surgical procedures should perform better than traditional treatment on kidney stones. When the data is grouped into treating small kidney stones and large kidney stones, the advanced surgical procedures outperform traditional treatment in each group. However, when all of the cases are combined, the traditional treatment outperforms!

How could this be? Well, the advanced surgical procedures were used more frequently when the kidney stones were large. Accordingly, these cases had high failure rates relative to smaller stones. Thus, because the advanced surgical procedure was used most in "tough" surgeries, it performed "worse" overall than traditional treatment.

In fact, there is an analogous result with medical evacuation helicopters and traditional ambulances. In the overall data, the helicopters actually do *worse* at saving lives than ambulances, but this is because they are sent to the higher-risk situations.

Low Birth Weight Paradox

Babies born to smokers have a higher mortality rate than babies born to non-smokers.

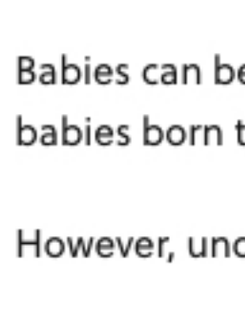
Babies can be born under-weight. It turns out that normal-weight babies born to smokers have an equal mortality rate to normal-weight babies born to non-smokers.

However, under-weight babies born to smokers have a *lower* mortality rate than under-weight babies born to non-smokers.

Can you guess why this might be the case?

Batting Averages

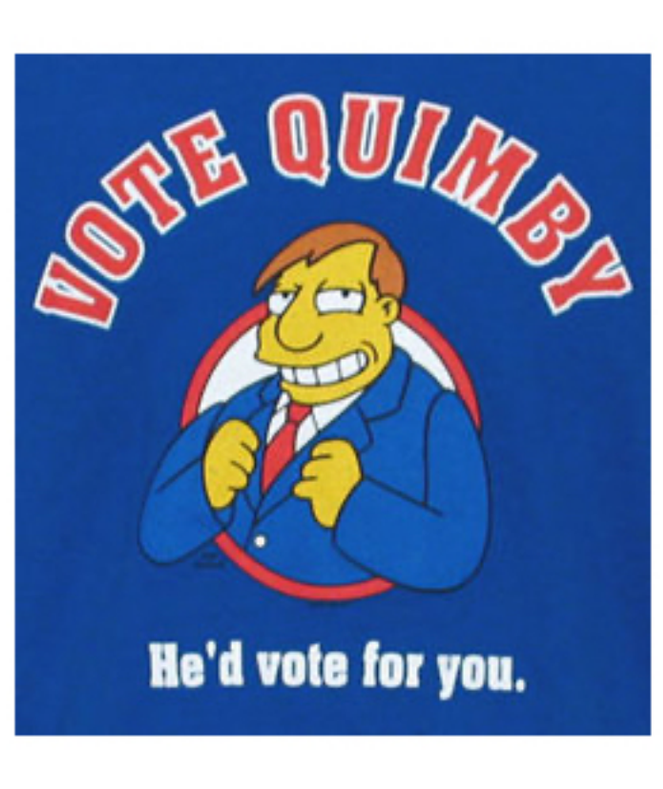
A baseball player can have higher batting average than another on each of two years, but lower than the other when the two are combined. In one case, David Justice had a higher batting average than Derek Jeter in 1995 and 1996, but across the two years, Jeter's average was higher.



Gerrymandering

It is possible to win a higher percentage of votes in multiple areas, yet lose the overall vote. This is a real-world phenomenon that can be partially seen in the U.S. electoral college model. Try your hand at this "Simpson's" example:

TRY IT YOURSELF



In a recent election, Sideshow Bob and Joe Quimby decided to run for election for Mayor of Springfield. The decision depends on the results in two districts: City and Countryside. Whichever candidate wins both districts wins the election.

The results shows that

In the City: 15000 out of 25000 voted for Sideshow Bob, while 4000 out of 5000 voted for Joe Quimby.

In the Countryside: 1000 out of 5000 voted for Sideshow Bob, while 7500 out of 25000 voted for Joe Quimby.

Tabulation of data:

Candidate	City	Countryside
Sideshow Bob	$\frac{15000}{25000} = 60\%$	$\frac{1000}{5000} = 20\%$
Joe Quimby	$\frac{4000}{5000} = \mathbf{80\%}$	$\frac{7500}{25000} = \mathbf{30\%}$

Because there's a higher percentage of people who have voted for Joe Quimby in both the City and Countryside, Joe Quimby was re-elected as the Mayor of Springfield.

If the decision does not depend on winning individual districts, but instead on the winning more votes in the entire population, then show that Sideshow Bob would have won the election. Also, if Sideshow Bob beat Joe Quimby by $x\%$ of the total voters, what is the value of x?

Image Credit: Simpsons Wikia.

Submit your answer

Post Your Own Example!

Feel free to add your own examples below! If you can polish it up, your example might even be featured above.

Andrew: I saw a real life example of Simpson's paradox when I did data analysis for my old company. The passing rates on the algebra end of course exam went up in each grade from 2012 to 2013, but the overall passing rate went down!

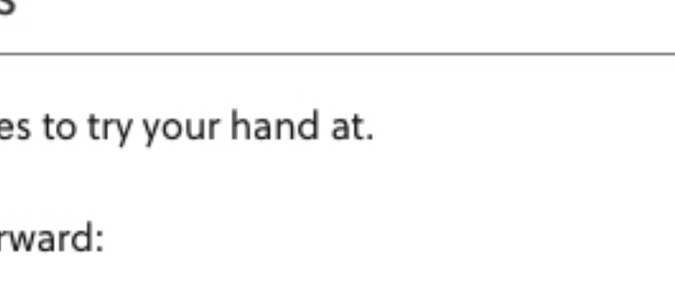
Write yours below!

Challenging Examples

Here's a few interesting examples to try your hand at.

This first one is pretty straightforward:

TRY IT YOURSELF



A survey was conducted on different age groups to determine whether more people prefer cars with automatic transmission or manual transmission.

	Age 16 to 21	Age 22 to 30	Age 31 to 50	Age 51 and higher
Automatic	90% of 100 people	60% of 200 people	50% of 300 people	50% of 400 people
Manual	80% of 400 people	40% of 300 people	40% of 200 people	40% of 100 people

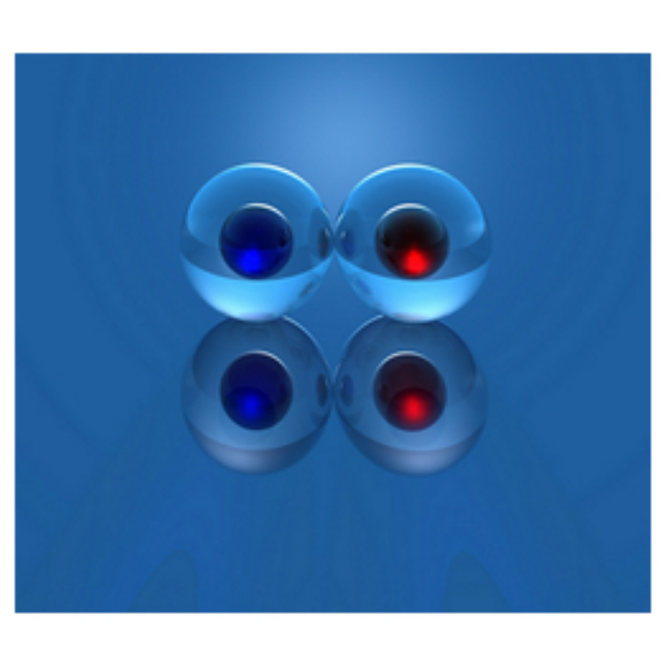
According to the survey result above, what is the conclusion?

Image Credit: Flickr Sophie.

- More people prefer cars with automatic transmission.
- More people prefer cars with manual transmission.
- There is no difference.

Here's more of a challenge:

TRY IT YOURSELF



I have four boxes, each containing a number of red marbles and blue marbles.

	Box A	Box B	Box C	Box D
Red marbles	70	y	2	7
Blue marbles	30	3	98	53

If the probability of randomly selecting a red marble from Box A is a, and the probability of randomly selecting a red marble from Box B is b, then $a < b$.

Suppose we group all the marbles in Box A and Box C into another Box AC; likewise we group all the the marbles in Box B and Box D into another Box BD. Now, there is a higher probability of randomly selecting a red marble from Box AC than from Box BD.

What is the sum of the smallest and the largest possible values of y for which the above criteria is satisfied?

Image Credit: Flickr Lyle.

Submit your answer

If you're looking for something really tough, check this one out:

TRY IT YOURSELF

Find the smallest example of (strict) Simpson's paradox; that is, construct such table where the number of cases is minimum. Formally, suppose that a, b, x, y are nonnegative integers and A, B, X, Y are positive integers such that $a \le A, b \le B, x \le X, y \le Y$, and also $\frac{a}{A} > \frac{x}{X}$, $\frac{b}{B} > \frac{y}{Y}$, but $\frac{a+b}{A+B} < \frac{x+y}{X+Y}$. Determine the minimum value of $A+B+X+Y$.

Example: There are two kinds of kidney stone problems, those with small stones and those with large stones. There are also two kinds of treatments, a simple treatment and a complex treatment. The number of success cases, divided by the number of cases for each stone/treatment combination, is displayed in the table below.

	Small stone	Large stone	Both
Complex treatment	81/87 **(93%)**	192/263 **(73%)**	273/350 (78%)
Simple treatment	234/270 (87%)	55/80 (69%)	289/350 **(83%)**

As one can see, the complex treatment performs better with small stone cases, and so as with large stone cases, but when the data is combined, the simple treatment performs better.

In the sample above, there are a total of 700 cases considered, with 350 complex treatments and 350 simple treatments (or alternatively 357 small stone cases and 343 large stone cases). This problem asks for the minimum possible total number of cases considered.

Clarification: In usual Simpson's paradox, it's allowed to have several weak inequalities (some of the inequalities above may actually be equalities). This problem thus has a stronger form of Simpson's paradox, where none of the inequalities may be an equality.

Submit your answer