

You are the Marketing Director for a large corporation, who is considering whether to rehire your current advertising agency, or move to a new one. The CEO of the current agency sends you the following e-mail message:

I'm happy to report the following results concerning the new national television advertising campaign. Over a two-week period ending ten days ago, we talked to consumers in both New York City and Los Angeles (2000 total consumers: 1000 consumers in each city, and 1000 consumers each week), to estimate the "recall rate" of the campaign. In New York City, the recall rates were 30% in the first week, and 40% in the second week; in Los Angeles, the recall rates were 80% in the first week and 90% in the second week. These increasing recall rates in both markets show that the campaign is catching on, just as we hoped.

Being pleased with these results, you sign the ad agency to a new three-year contract.

Two weeks later, the CEO of your company comes into your office. "I've got some bad news. I took a look at that data about the ad campaign, and I've found that, in fact, overall recall rates were **declining** in the data, not increasing." "Do you mean they lied to us? We'll sue!" you shout. "Well, no," she replies. "The numbers they gave you were correct, but what I'm saying is correct, too."

Should you:

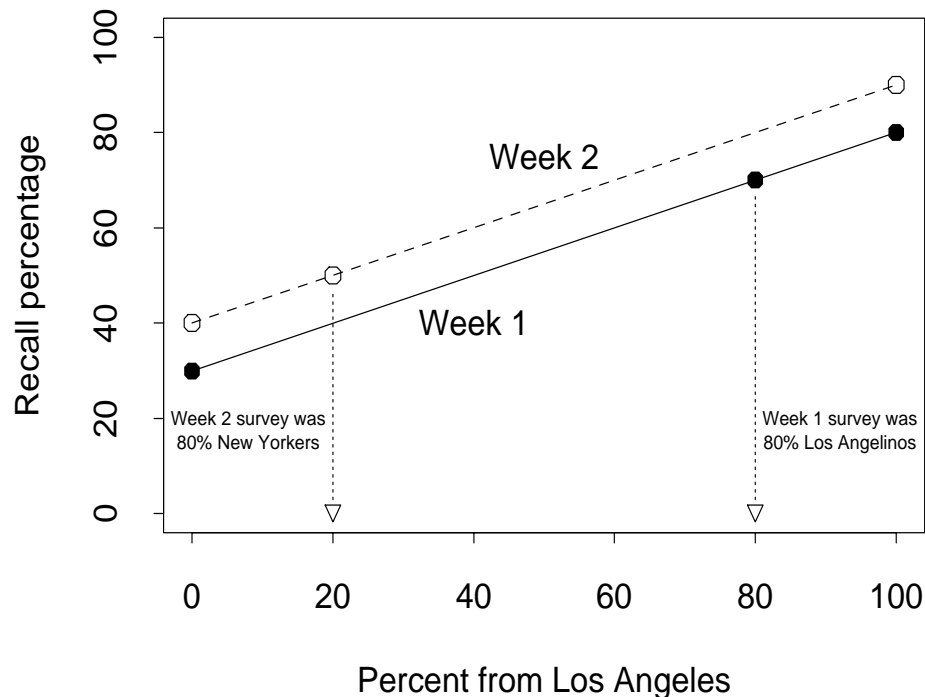
- (1) look for a new job, since it's now obvious to you that the CEO is such an idiot that she can't see that what she's saying is impossible?
- (2) look for a new job, since it's now obvious to the CEO that you are such an idiot that you couldn't see that this has happened?
- (3) go get drunk, and hope that this all blows over by the morning?

Here are the detailed results of the market research study:

	New York City	Los Angeles	
Week 1	60/200 (30%)	640/800 (80%)	700/1000 (70%)
Week 2	320/800 (40%)	180/200 (90%)	500/1000 (50%)

So, even though the recall rate was *increasing* within each city, it is *decreasing* overall. This is called **Simpson's paradox**, and it is an example of how associations within subgroups are not always consistent with overall associations ignoring subgroups. The reason this happened here is that New York City, which had generally lower recall rates, was more heavily sampled the second week, while Los Angeles, with generally higher recall rates, was more heavily sampled in the first week. This results in the overall recall rate appearing higher in the first week (being closer to the Los Angeles rate) and lower in the second week (being closer to the New York rate).

The figure on the next page represents Simpson's paradox in a graphical form (this figure comes from my 2003 book, *Analyzing Categorical Data*). The two lines represent what the overall recall percentage in week 1 (solid line) and week 2 (dashed line) would have been, as a function of the percentage of the total sample that was taken in Los Angeles. So, for example, in week 1, if none of the survey had been taken in Los Angeles, the overall recall rate would have been the week 1 New York rate, or 30%. Similarly, if in week 1 all of the survey had been taken in Los Angeles, the overall recall rate would have been the week 1 Los Angeles rate, or 80%. In fact, 80% of the respondents in week 1 were from Los Angeles, so the overall recall rate was 70%, represented by the marked point on the week 1 line. The corresponding values for week 2 are also given in the figure. The consistently higher line for week 2 compared to week 1 reflects the subgroup pattern that recall rates were generally higher in week 2, but the higher position of the marked point for week 1 compared to that for week 2 demonstrates the overall pattern that the overall rate was higher in week 1 than in week 2.



This diagram shows that Simpson's paradox actually reflects a sampling problem, in the sense that the problem only can arise when a control variable Z (in this case city) is related to the variables of interest X and Y (here week and recall, respectively). This suggests that when it is possible to control the level of the control variable for each observation, this should be done so as to minimize any chance of a relationship between it and the other variables. So, for example, if the marketing survey had been based on 500 respondents in each city for each week, city and time period would have been unrelated to each other, and Simpson's paradox could not have occurred (this corresponds to the week 2 line always being above the week 1 line in the figure if the percentage of the survey from Los Angeles is the same in the two weeks). In a study of the relationship between dosage of a drug and survival, for example, with gender a control variable, men and women should be assigned to dosage level randomly (and evenly), preventing the possibility of Simpson's paradox.

Here is a real example of this effect. The table refers to the death penalty verdict for the 674 defendants in indictments involving cases with multiple murders in Florida between 1976 and 1987 (source: A. Agresti, *An Introduction to Categorical Data Analysis*, Wiley, 1996). Overall, 53 of 483 white defendants were convicted with a death penalty

verdict (11%), while 15 of 191 black defendants (7.9%) were convicted. But, here is a table separated by race of the victim:

	White victim	Black victim	
White defendant	53/467 (11.3%)	0/16 (0%)	53/483 (11%)
Black defendant	11/48 (22.9%)	4/143 (2.8%)	15/191 (7.9%)

Note that the defendant race difference in conviction rate for white victims and for black victims are both in the same direction (higher conviction rates for black defendants). However, the overall rate for white defendants is higher than that for black defendants, because the weighted average is more heavily weighted towards the higher rate for white victims for white defendants and more heavily weighted towards the lower rate for black victims for black defendants.

Note, by the way, that the occurrence of Simpson’s paradox has important implications in business applications as well, and is not just limited to proportions. Consider, for example, a portfolio that consists of a large cap mutual fund, a bond fund, and a real estate fund. Say in year one the portfolio is weighted (.2, .2, .6), and has return performance (12.0%, 10.0%, 8.0%). The overall return of the portfolio is 9.2%. In year two the portfolio is weighted (.6, .3, .1), and has return performance (11.0%, 9.5%, 7.0%); the overall return of the portfolio is 10.15%. In which year did the “market” do better? Clearly year 1, since the returns of each part of the portfolio were higher than in year 2. But in which year did the portfolio do better? Year 2, of course. We might very well say that the portfolio manager deserves credit for reweighting the portfolio to emphasize the large cap and bond funds in year 2, but the fact is that the overall (marginal) return pattern is precisely the opposite of the specific product (conditional) returns, and ignoring this leads to misunderstanding.