

More Stratification: For years observational studies have shown that people who carry lighters are more likely to get lung cancer. However, this does not mean that carrying lighters causes you to get cancer. Smoking is an obvious confounder! If we weren't sure about this, how can we determine whether it's the lighters or the confounders or (maybe some combination of both) that is causing the lung cancer?

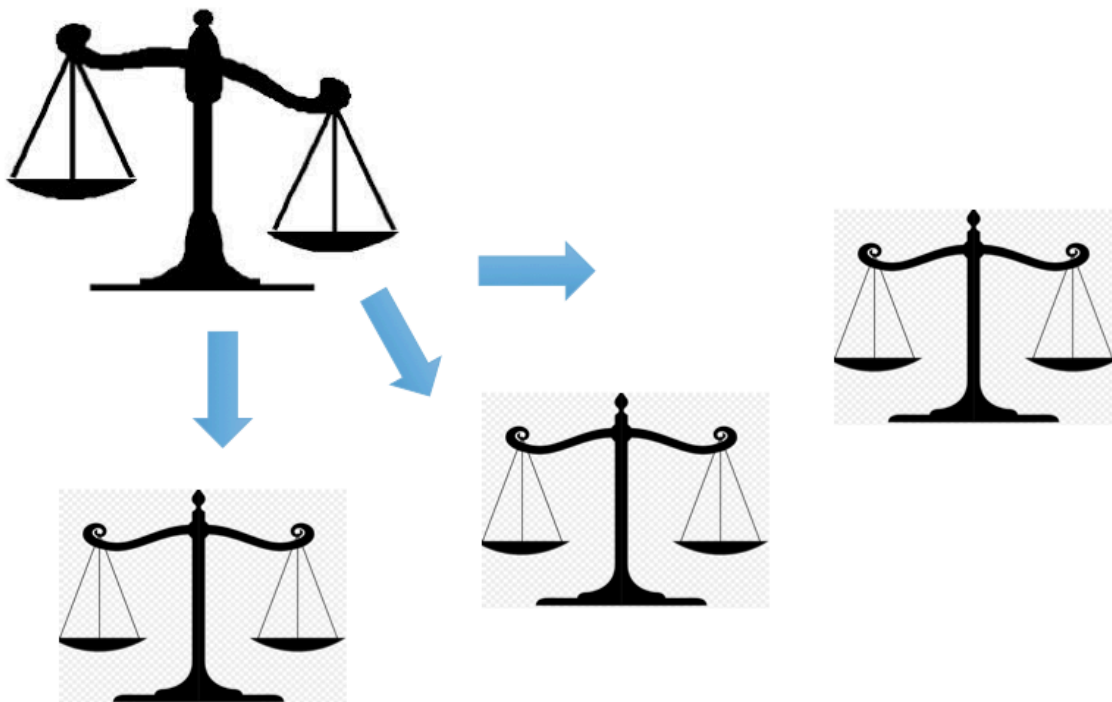
Stratification

Break (stratify) the population into subgroups where the confounding factor is the same.

Non-Smokers	Compare lung cancer rates of those who carry lighters to those who don't.
Moderate Smokers	Compare lung cancer rates of those who carry lighters to those who don't.
Heavy Smokers	Compare lung cancer rates of those who carry lighters to those who don't.

In this case, there's no difference in cancer rates between those who carry lighters and those who don't **within** each group. Of course, the heavy smoker group has the highest cancer rates but rates between those who carry lighters and those who don't are the same in that group.

Thus, the lighters are just a marker for people who smoke—smoking, not lighters, causes cancer.



Simpson's Paradox for _____: A clear-cut case of confounding that is easily adjusted for by stratification (dividing into subgroups).

Example 1: Sex Bias in Berkeley Graduate Admissions?

In 1973, thousands of men and thousands of women applied for admission to graduate school at UCB. 44% of the men and 35% of the women were admitted¹.

Is this a controlled experiment or observational study?

Assuming that men and women are equally qualified, is this evidence of sex bias?

Break it down by major:

		Men		Women	
		# Applied	% Admitted	# Applied	% Admitted
M a j o r	A	825	62	108	82
	B	560	63	25	68
	C	325	37	593	34
	D	417	33	375	35
	E	191	28	393	24
	F	373	6	341	7
Totals		2690		1835	

There is no sex bias against women if you adjust for major. Women's overall admission rate is worse simply because most of the women applied to the harder majors (C-F), while most of the men applied to the easier majors (A and B).

Example 2: Could smoking prolong life?

A 20-year (1974-94) study of 1314 British women compared death rates of smokers to non-smokers and found that 23.9% (139/582) of the smokers had died compared to 31.4% (230/732) of the non-smokers. Could this be evidence that smoking helps you live longer?

Control for age:

<i>Age in 1974</i>	<i># of Smokers</i>	<i>Death Rate of Smokers</i>	<i># of Non-smokers</i>	<i>Death Rate of Non-smokers</i>
18-34	179	$5/179 = 2.8\%$	219	$6/219 = 2.7\%$
35-64	354	$92/354 = 26.0\%$	320	$59/320 = 18.4\%$
Over 65	49	$42/49 = 85.7\%$	193	$165/193 = 85.5\%$
Total	582	$139/582 = 23.9\%$	732	$230/732 = 31.4\%$

Non-smokers had a high overall death rate because 26.4% (193/732) of them were old, while only 8.4% (49/582) of the smokers were old. If you control for the age you get the opposite: Smoking shortens your life.

Simpson's Paradox again: Overall percentage is misleading because of a confounder. Once the confounder is controlled for by looking at sub-groups separately, the overall effect is reversed.

Example 3: While they were in college, Karle Flanagan and her husband Steve decided to take some courses online. The courses are divided into 2 groups (A & B) that cover the same subjects but at a different level of difficulty. Group A courses are very difficult & include courses such as Japanese, Physics, & Calculus with Several Variables. Group B courses are easier & include courses such as Japanese Movies, Physical Education, & Calculators with Several Fingers. Here's a chart of how they did:

	Karle		Steve	
	# Courses	GPA	# Courses	GPA
Group A	1	1.0	9	3.0
Group B	9	3.6	1	3.9
Total	10	3.34	10	3.09

- Who had the **higher** GPA for Group A courses? **a)** Karle **b)** Steve **c)** Cannot be determined
- Who had the **higher** GPA for Group B courses? **a)** Karle **b)** Steve **c)** Cannot be determined
- Who had the **higher** overall GPA? **a)** Karle **b)** Steve **c)** Cannot be determined
- Which conclusion is best? **Choose one:**
 - Here it's best to compare total GPAs and clearly, Karle is the better student. She has a better overall GPA (3.34) than Steve who has an overall GPA of 3.09.
 - If you look at Group A and B courses separately, it is obvious that Steve is the better student. His overall GPA is lower because the results are confounded by the difficulty level of the courses.
 - Here we cannot determine any conclusion because we don't know if Karle and Steve actually did their homework for each class.
 - There must have been a mistake when calculating the total GPAs for Karle and Steve since Karle's total GPA is better than Steve's.

SUMMARY

Simpson's Paradox: Overall percentage is misleading because of the confounding effect of class difficulty level. Once the confounder is controlled for by looking at sub-groups separately, the overall effect is reversed. High schools adjust overall GPA's by weighting harder classes more than easier ones to deal with this same confounder.

References

- Stat 100 Incomplete Lecture Notes Workbook by Ellen Fireman, Karle Flanagan, & John Marden*
- Stat 200 Incomplete Lecture Notes Workbook by Ellen Fireman, Karle Flanagan, & John Marden*
- Data 8 Online Textbook: The Foundations of Data Science by Ani Adhikari & John DeNero*