Simpson's Paradox and Lurking Variables

Example 1

New York Times, Jan 12, 1990: *"The results of a government study on death rates in nearly 6,000 hospitals were challenged today by researchers who said the federal analyses failed to account for variations in the severity of patients' illness when they were hospitalized. As a result, they said, some hospitals were treated unfairly in the findings, which named hospitals with higher-than-expected death rates.*

The idea is that forgetting to account for severity of illness, a *lurking variable*, can affect our conclusions about the relationship between two variables, *hospital* and *death rate* that we care about.

Here is a simplified example of this phenomenon.

Looking only at hospital and death rate

	Hospital A	Hospital B
Died	63	16
Survived	2037	784
Total	2100	800

	Hospital A	Hospital B
Died	3%	2%
Survived	97%	98%
Total	100%	100%

Accounting for the lurking variable

Patients not so severe

	Hosp A	Hosp B
Died	6	8
Surv	594	592
Total	600	600

Patients severely ill

	Hosp A	Hosp B
Died	57	8
Surv	1443	192
Total	1500	200

Patients not so severe

	Hosp A	Hosp B
Died	1.00%	1.33%
Surv	99.00%	98.67%
Total	100.00%	100.00%

Patients severely ill

	Hosp A	Hosp B
Died	3.80%	4.00%
Surv	96.20%	96.00%
Total	100.00%	100.00%

Lurking Variable

A variable that you did not include in your analysis, that could substantially change your interpretation of the data if you did include it, is a **lurking variable**.

Including a lurking variable may

- Have no effect
- Make you re-think the cause of a phenomenon
- Make you re-think the direction (increasing vs de-creasing) of an association

Simpson's Paradox

When including a lurking variable causes you to re-think the direction of an association, this is called **Simpson's paradox**.

Failing to think about and include lurking variables is the primary reason that statisticians often say:

Correlation and association **do not** *imply causation.*

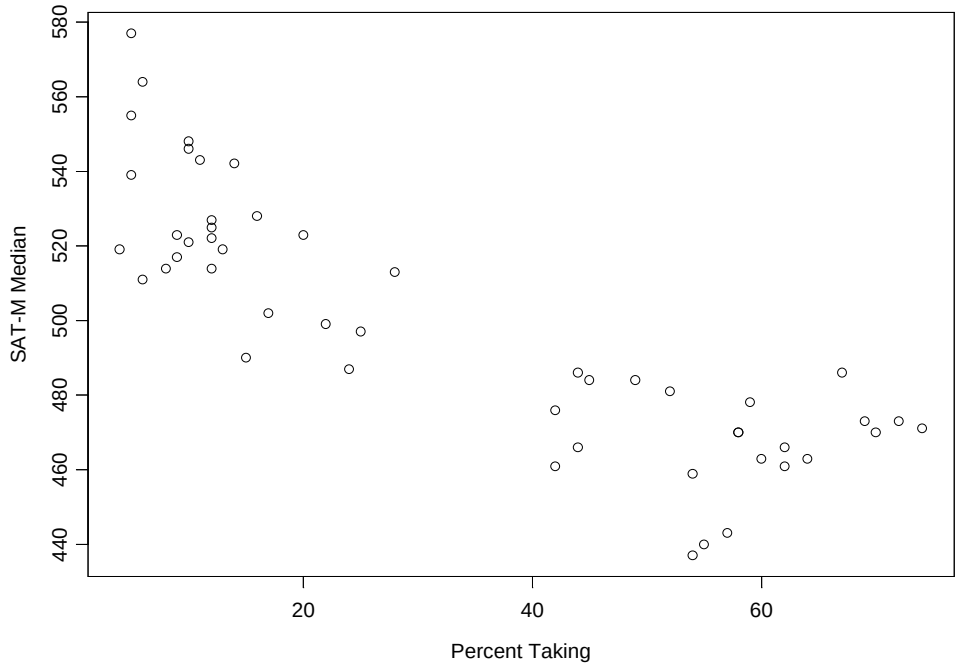
Example 2

There is great interest in comparing states in the US on educational achievement. One proposal that is often made is that states could be compared on a standard college entrance exam, like the SAT. Following are median SAT math scores, together with the percent of students who take the SAT in each state. On the next page is a scatter plot of this data.

ST	SAT	PCT	ST	SAT	PCT
AL	514	8	MT	523	20
AK	476	42	NE	546	10
AZ	497	25	NV	487	24
AR	511	6	NH	486	67
CA	484	45	NJ	473	69
CO	513	28	NM	527	12
CT	471	74	NY	470	70
DE	470	58	NC	440	55
FL	466	44	ND	564	6
GA	443	57	OH	499	22
HI	481	52	OK	523	9
ID	502	17	OR	484	49

[Continued. . .]

ST	SAT	PCT	ST	SAT	PCT
IL	528	16	PA	463	64
IN	459	54	RI	461	62
IA	577	5	SC	437	54
KS	548	10	SD	555	5
KY	521	10	TN	525	12
LA	517	9	TX	461	42
ME	463	60	UT	539	5
MD	478	59	VT	466	62
MA	473	72	VA	470	58
MI	514	12	WA	486	44
MN	542	14	WV	490	15
MS	519	4	WI	543	11
MO	522	12	WY	519	13



- Plotting, transforming if need be
- Relations between the variables
- Clustering
- Outliers (in X or Y)
- Unequal Variability

What is the lurking variable?

Lurking Variable

A variable that you did not include in your analysis, that could substantially change your interpretation of the data if you did include it, is a **lurking variable**.

Including a lurking variable may

- Have no effect
- Make you re-think the cause of a phenomenon
- Make you re-think the direction (increasing vs de-creasing) of an association

Failing to think about and include lurking variables is the primary reason that statisticians often say:

Correlation and association **do not** *imply causation.*