

PS2030

Political Research and Analysis

Spring 2025
Multiple Regression

Paradoxes of Statistical “Control”

A “Normal” Pattern First: Bivariate Relationship: High Media and High Political Knowledge

```
. tab highknow highmedia, col
```

Key
<i>frequency</i> <i>column percentage</i>

RECODE of know	RECODE of media2		Total
	1	2	
1	309 73.40	231 44.51	540 57.45
2	112 26.60	288 55.49	400 42.55
Total	421 100.00	519 100.00	940 100.00

Difference between Percentage of Low Media Users who are “High” on Knowledge (27%) and Percentage of High Media Users who are “High” on Knowledge (55%) =28 percentage point difference.

Pearson Correlation=.29

43% of Overall Sample is “High” on Knowledge (400/940)

55% of Overall Sample is “High” on Media Use (519/940)

Controlling for College Education: Multivariate Analysis

-> college = 1

Key
frequency column percentage

RECODE of know	RECODE of media2		Total
	1	2	
1	282 78.99	176 61.75	458 71.34
2	75 21.01	109 38.25	184 28.66
Total	357 100.00	285 100.00	642 100.00

-> college = 2

Key
frequency column percentage

RECODE of know	RECODE of media2		Total
	1	2	
1	27 42.19	55 23.50	82 27.52
2	37 57.81	179 76.50	216 72.48
Total	64 100.00	234 100.00	298 100.00

Among no college educated group (68% of overall sample):

Percentage Difference=16%

Pearson Correlation=.19

29% are “High” on Knowledge

44% are “High” Media Users

Taken together, Media and College explain Knowledge better than either alone: 21% of non-college low media users are “high” on knowledge; 77% of college, high media users are “high” on knowledge

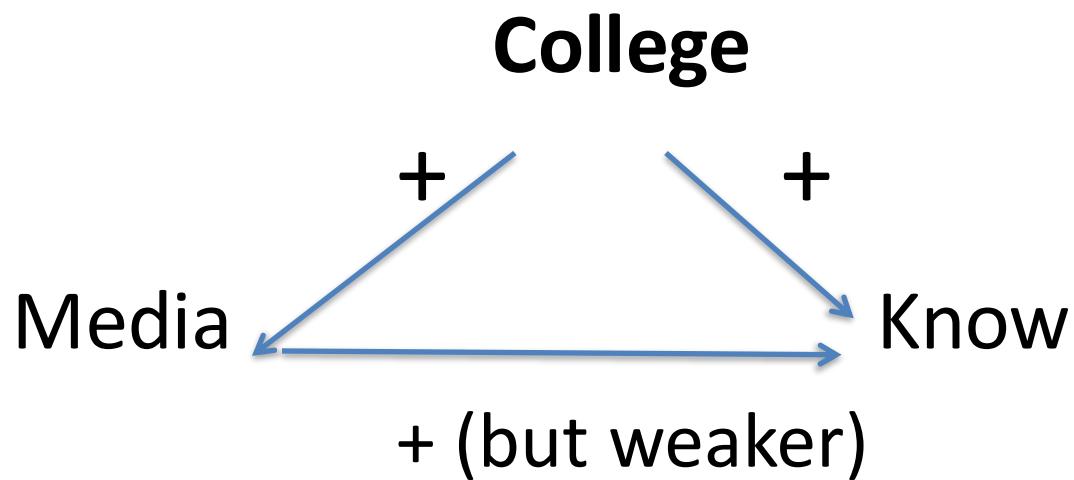
Among college educated group (32% of overall sample):

Percentage Difference=18%

Pearson Correlation=.17

72% are “High” on Knowledge

79% are “High” Media Users



Why?

College is positively associated with Media use and positively associated with Knowledge, so some of the bivariate relationship was due to the joint association of Media Use and Knowledge with College

SIMPSON'S PARADOX: A REVERSAL OF AN OVERALL BIVARIATE RELATIONSHIP WHEN EXAMINING SUB-GROUPS

Bivariate Relationship: KIDNEY TREATMENTS “A” AND “B” AND SUCCESSFUL OUTCOMES

(SUCCESS IS ROW OUTCOME 1; FAILURE IS ROW OUTCOME 2)

TREATMENT A is COLUMN 1; TREATMENT B IS COLUMN 2

row	col		Total
	1	2	
1	273 78.00	289 82.57	562 80.29
2	77 22.00	61 17.43	138 19.71
Total	350 100.00	350 100.00	700 100.00

Difference between TREATMENT A (COLUMN 1) SUCCESS and TREATMENT B(COLUMN 2) SUCCESS = $78 - 82.6\% = -4.6\%$

SO TREATMENT B LOOKS TO HAVE **HIGHER** SUCCESS RATE

Controlling for Size of Kidney Stone (Severity of Problem): Multivariate Analysis

row	col		Total
	1	2	
1	192 73.00	55 68.75	247 72.01
2	71 27.00	25 31.25	96 27.99
Total	263 100.00	80 100.00	343 100.00

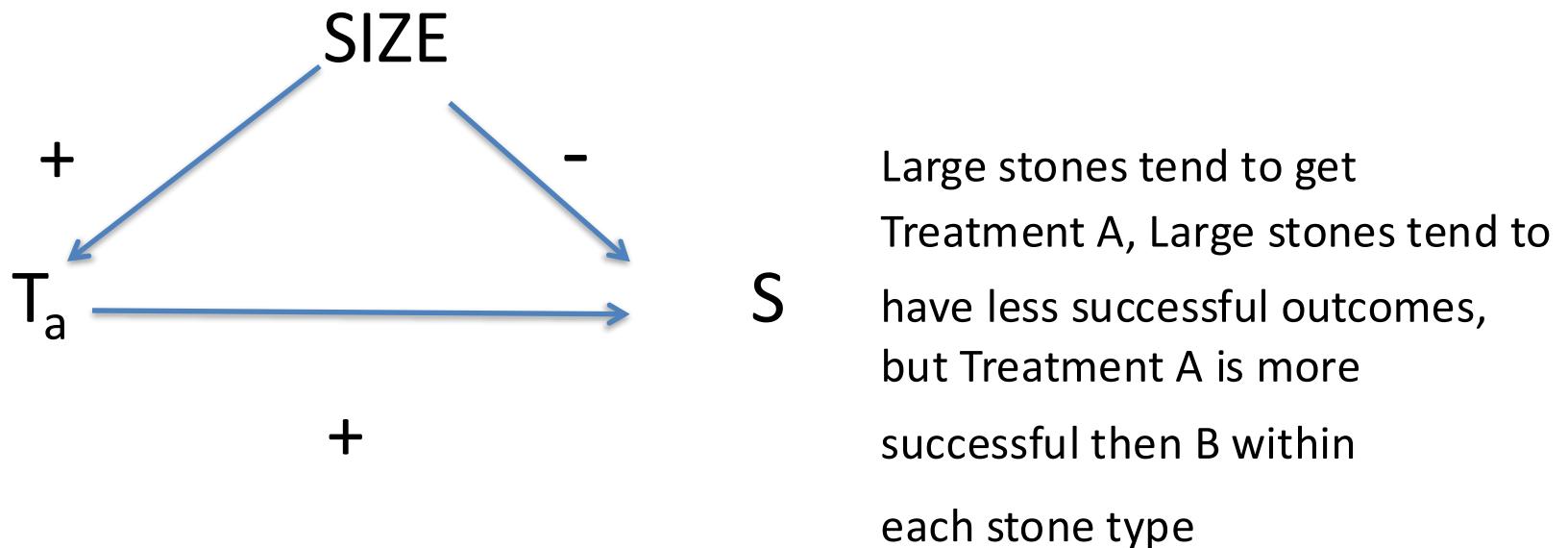
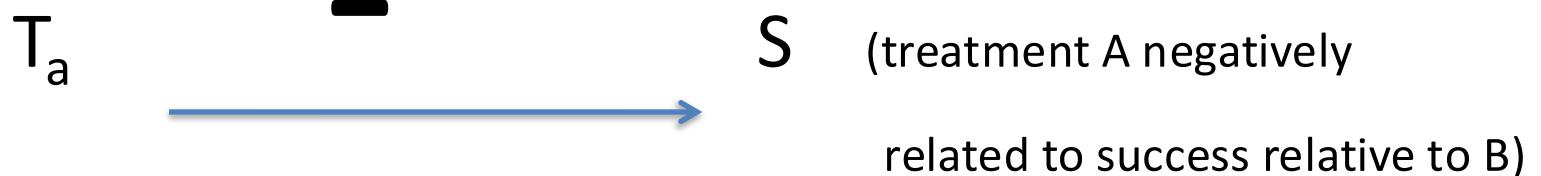
row	col		Total
	1	2	
1	81 93.10	234 86.67	315 88.24
2	6 6.90	36 13.33	42 11.76
Total	87 100.00	270 100.00	357 100.00

Among LARGE STONE GROUP:
TREATMENT A SUCCEEDS 74%
TREATMENT B SUCCEEDS 69%

Among SMALL STONE GROUP:
TREATMENT A SUCCEEDS 93%
TREATMENT B SUCCEEDS 87%

SO: TREATMENT A IS WORSE IN AGGREGATE, BUT TREATMENT A IS BETTER WITHIN EACH GROUP EXAMINED SEPARATELY!!! HOW IS THIS POSSIBLE???

THIS IS CALLED “SIMPSON’S PARADOX”



Treatment A is more successful despite being used in situations where success is less common

“Berkson’s Paradox”

- Classic Example: You are a stamp collector who puts on display stamps that are either “pretty” or “rare”. Prettiness and rareness in stamps in the population are not related. You have 1000 stamps, 300 of which are pretty, 100 of which are rare, and 30 are both pretty and rare.

```
. tabi 30 70 \270 630 , col v
```

		PRETTY	NOT PRETTY	
		1	2	Total
RARE	1	30	70	100
		10.00	10.00	10.00
NOT RARE	2	270	630	900
		90.00	90.00	90.00
Total		300	700	1,000
		100.00	100.00	100.00

Cramér's V = 0.0000

NOTICE: 10% OF ALL PRETTY STAMPS ARE RARE; 10% OF NOT PRETTY STAMPS ARE RARE
THEREFORE NO RELATIONSHIP BETWEEN PRETTY AND RARENESS AMONG ALL STAMPS

- NOW: WHAT ABOUT AMONG THE STAMPS YOU DISPLAYED?
THAT IS, THE 370 STAMPS THAT ARE EITHER PRETTY OR RARE?

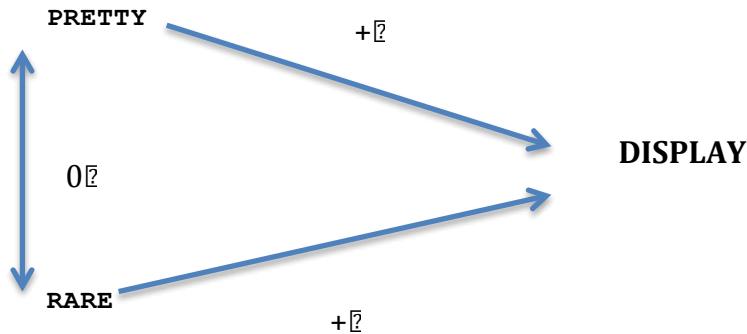
```
. tabi 30 70 \270 0 , col v
```

		PRETTY	NOT PRETTY	
	row	1	2	Total
RARE	1	30	70	100
		10.00	100.00	27.03
NOT RARE	2	270	0	270
		90.00	0.00	72.97
Total		300	70	370
		100.00	100.00	100.00

Cramér's v = -0.7937

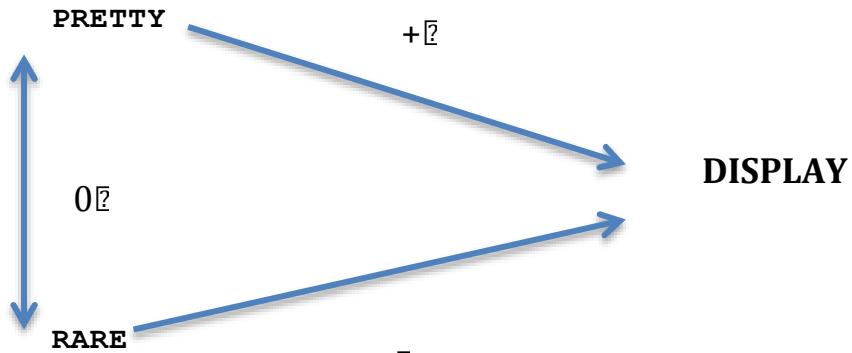
- Here we see that 10% of pretty stamps are (still) rare, but 100% of not-pretty stamps are rare (or else they would not have been on display!)
- So, among the displayed stamps, there is a **negative** relationship between pretty and rareness. By “controlling” for “display/not display”, we have induced a negative relationship between factors that are actually unrelated in the population!

CAUSAL DIAGRAM

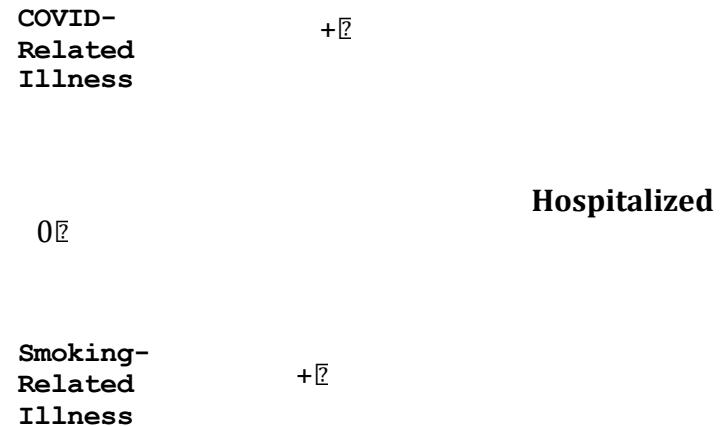


- But selecting on display ==‘yes’, pretty and rare are strongly negatively correlated
- In the causal language stemming from the work of Judea Pearl and followers, “DISPLAY” is said to be a “collider” and *one should never “control” for a collider* (or anything “downstream” from the collider as well)!
- This is one kind of selection bias, known to political scientists as “selecting on the dependent variable”

CAUSAL DIAGRAM



CAUSAL DIAGRAM



- Implications for political science and public policy? Substitute “COVID-19-Related Illness” and “Smoking-Related Illness” for “Pretty” and “Rare”, and substitute “Hospitalized” for “Display”. Assume these are the only illnesses leading to hospitalization
- *Among those hospitalized, if you don't have a smoking illness you must have a COVID-related illness. If you don't have a COVID-related illness you must have a smoking-related illness. So smoking and COVID are negatively related among hospitalizations.*
- Hence smoking helps prevent COVID???
- This was actually claimed and [debunked](#) in 2020 based on Berkson's Paradox in the journal *Significance*, a collaborative journal of the British Royal Statistical Association, the American Statistical Association, and the Statistical Society of Australia.