Gender Bias, Simpson's Paradox and Causal Inference

Lisa Goldberg

December 15, 2019

Berkeley Math Circle

Julia Hall Bowman Robinson (December 8, 1919–July 30, 1985) was an American mathematician noted for her contributions to the fields of computability theory and computational complexity theory—most notably in decision problems. Her work on Hilbert's 10th problem (now known as Matiyasevich's theorem or the MRDP theorem) played a crucial role in its ultimate resolution.

As a graduate student, Julia was employed as a teaching assistant with the Department of Mathematics and later as a statistics lab assistant by Jerzy Neyman in the Berkeley Statistical Laboratory, where her work resulted in her first published paper, "A Note on Exact Sequential Analysis."

A nepotism rule prevented Julia from teaching in the mathematics department since she was married to Professor Raphael M. Robinson. So she stayed in the statistics department despite wanting to teach calculus. Raphael retired in 1971. In 1976, after her election to the National Academy of Sciences, Julia was offered a professorship in the UC Berkeley mathematics department.

Graduate admissions at UC Berkeley in 1973

Was there gender bias in the 1973 graduate admissions process?

29% of admitted students were female.

34% of applicants were female.

	Men		Women	
	Applicants	Admitted	Applicants	Admitted
Total	8442	44%	4321	35%

The overall acceptance rate was 41%.

How might a statistician decide if the data indicate gender bias?

Observed Data

	Men	Women
Admitted	3738	1494
Denied	4704	2827
Total	8442	4321

Sex Bias in Graduate Admissions: Data from Berkeley

Measuring bias is harder than is usually assumed, and the evidence is sometimes contrary to expectation.

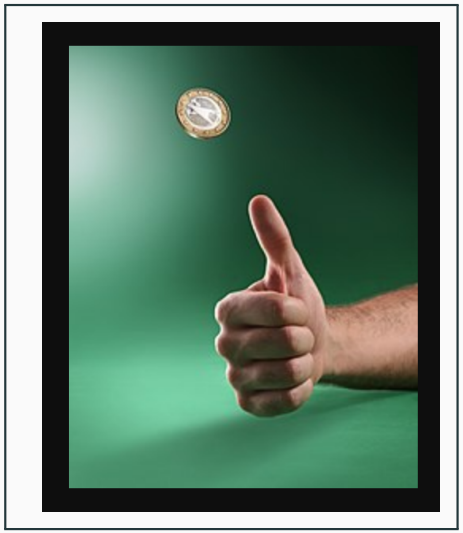
P. J. Bickel, E. A. Hammel, J. W. O'Connell

How might a statistician decide if the data indicate gender bias?

Use a test statistic: a quantity derived from sample data. It is important that the distribution of the test statistic be known (or approximately known) under assumptions.

A distribution is a collection of outcomes and their likelihoods.

An example of a distribution is a fair coin: outcomes are heads and tails; likelihoods are 50-50.



Bickel et al. used expectations based on the overall acceptance rate of approximately 41% to generate a test statistic

Expected Data

	Men	Women
Admitted	3460.7	1771.3
Denied	4981.3	2549.7
Total	8442	4321

The test statistic is called "Pearson chi-squared"

Expected Data

	Men	Women
Admitted	3460.7	1771.3
Denied	4981.3	2549.7

Observed minus Expected Data

	Men	Women
Admitted	277.3	-277.3
Denied	-277.3	277.3

$$\chi^2 = \sum_{n=1}^{4} \frac{(O_n - E_n)^2}{E_n} = 110.8$$

Under assumptions, the probability that χ^2 is 110.8 or greater by pure chance is 6.5×10^{-26}...

...suggesting gender bias in UC Berkeley admission process.

Which departments were guilty?

UC Berkeley's graduate admissions processes are conducted by individual departments...

... so a dean set out to determine the source of the bias...

...but did not find it.

Number of Departments	Result
16	No women applied or no one was rejected
4	Biased toward men
6	Biased toward women
75	No bias

Simpson's paradox, or the Yule-Simpson effect

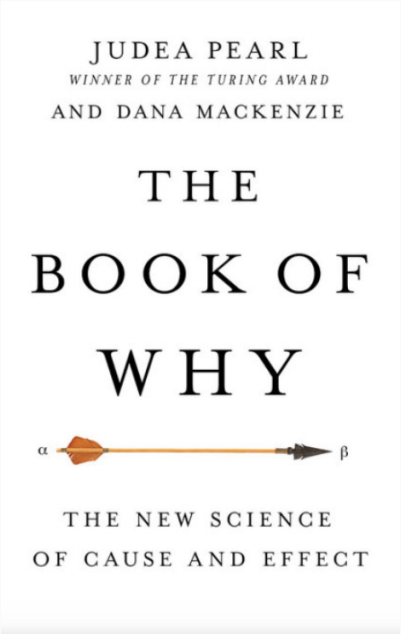
When looking at the statistical scores of groups, these scores may change, depending on whether the groups are looked at one by one, or if they are combined into a larger group.

Causal inference and Simpson's paradox

The Book of Why

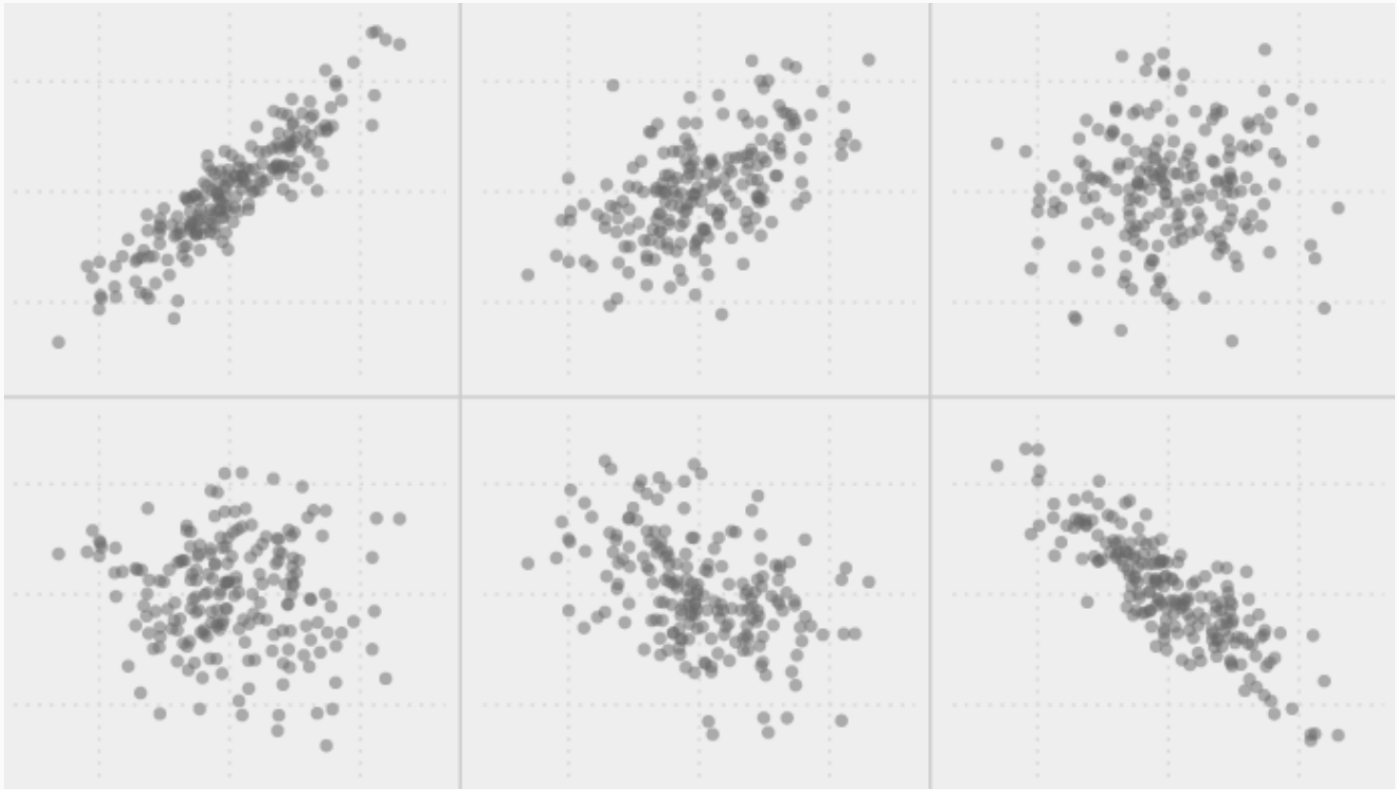
In 2018, Judea Pearl and Dana Mackenzie published *The Book of Why*, a historically-grounded, accessible, colorful treatment of causal inference and statistics.

The book is about a framework for extracting cause-and-effect relationships from data.



Why do we need a book about this?

Correlation is the work horse of statistics. It measures the tendency of two random quantities to move together.



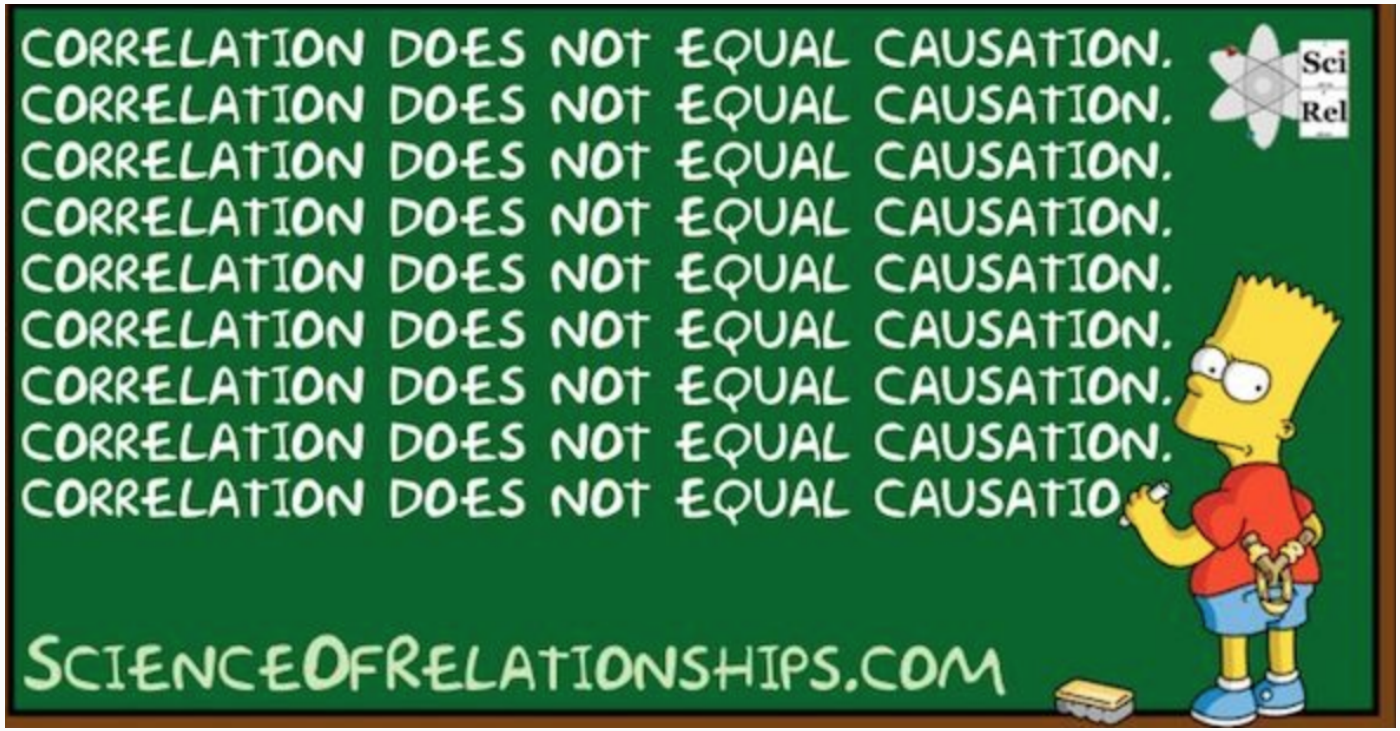
Why do we need a book about this?

But correlation is symmetric in its arguments...

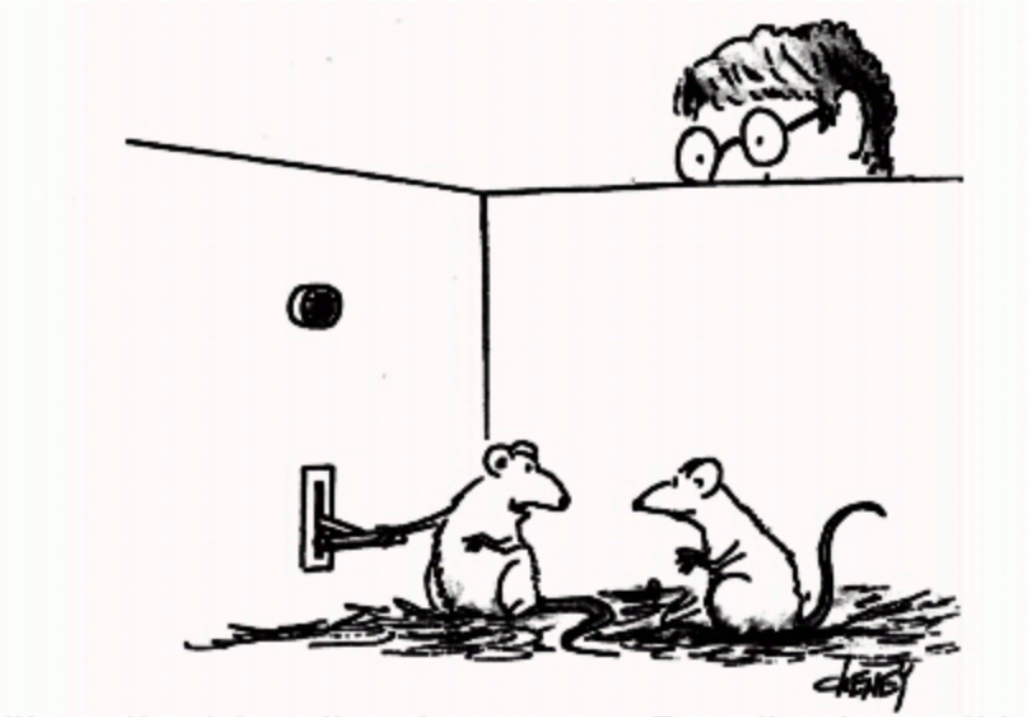
$$\rho(X, Y) \approx \frac{\sum_n (X_n - \overline{X})(Y_n - \overline{Y})}{\left(\sum_n (X_n - \overline{X})^2 \sum_n (Y_n - \overline{Y})^2\right)^{1/2}}$$

...so on its own, it can't imply $X \to Y$ or $Y \to X$.

Why do we need a book about this?



Still, researchers infer cause and effect from correlation all the time



It's a rather interesting phenomenon. Every time I press this lever, that post-graduate student breathes a sigh of relief.

Simpson's reversal

In elementary school, we learn that summing numerators and denominators is *not* the way to add fractions. In fact, it is possible that:

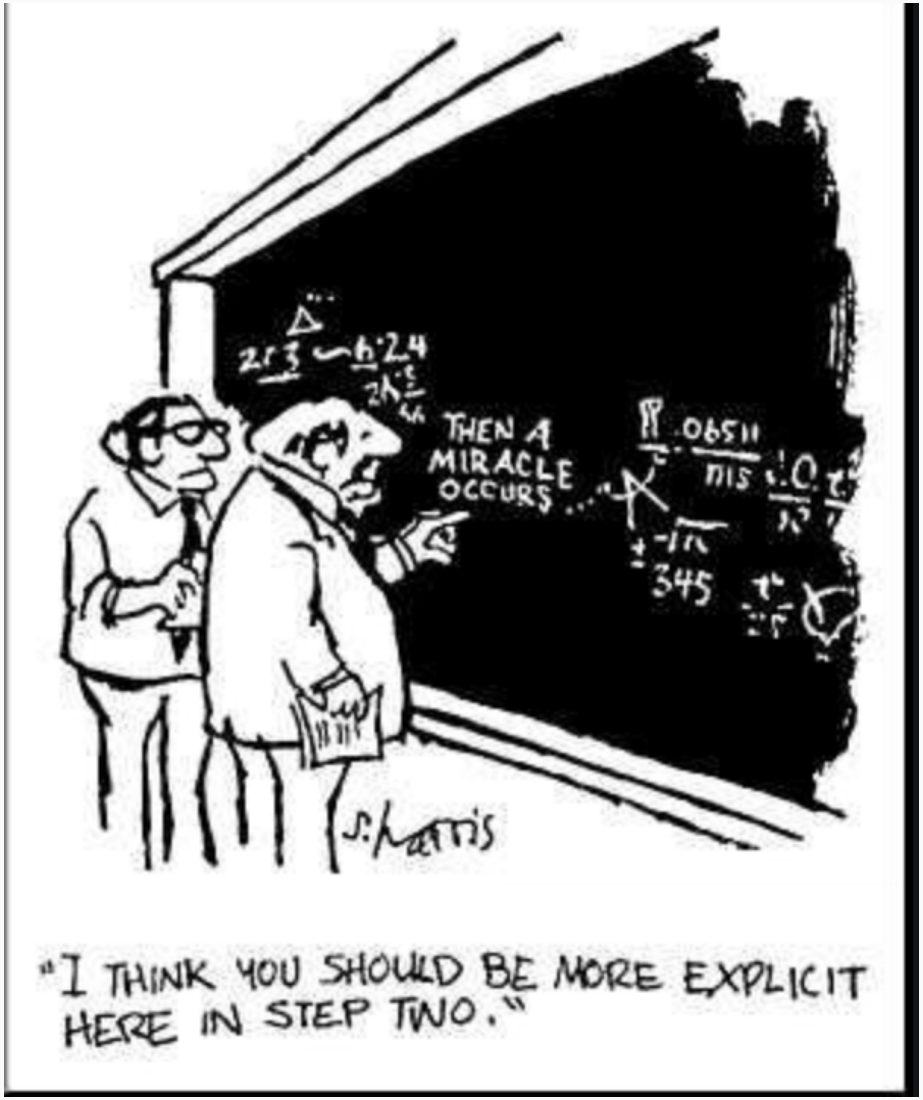
$$a/b > c/d \quad \text{and} \quad e/f > g/h$$

while

$$(a+e)/(b+f) < (c+g)/(d+h).$$

I'll illustrate with a simple example.

A new miracle drug



Does the new miracle drug prevent heart attacks?

	Treatment	Control
Heart attack	11	13
Healthy	49	47
Total	60	60

Does the new miracle drug prevent heart attacks?

All		Treatment	Control
	Heart attack	11	13
	Healthy	49	47
Group 1			
	Heart attack	3	1
	Healthy	37	19
Group 2			
	Heart attack	8	12
	Healthy	12	28

Does a new miracle drug prevent heart attacks?

All		Treatment	Control
	Heart attack	11	13
	Healthy	49	47
	Percent healthy	82	78
Group 1			
	Heart attack	3	1
	Healthy	37	19
	Percent healthy	93	95
Group 2			
	Heart attack	8	12
	Healthy	12	28
	Percent healthy	60	70

Simpson's reversal: percent healthy rates

$$\text{control}(1) > \text{treatment}(1) \quad \text{and} \quad \text{control}(2) > \text{treatment}(2)$$

$$19/20 > 37/40 \quad \text{and} \quad 28/40 > 12/20$$

while

$$\text{control}(\text{total}) < \text{treatment}(\text{total})$$

$$(19 + 28)/(20 + 40) = 47/60 < 49/60 = (37 + 12)/(40 + 20).$$

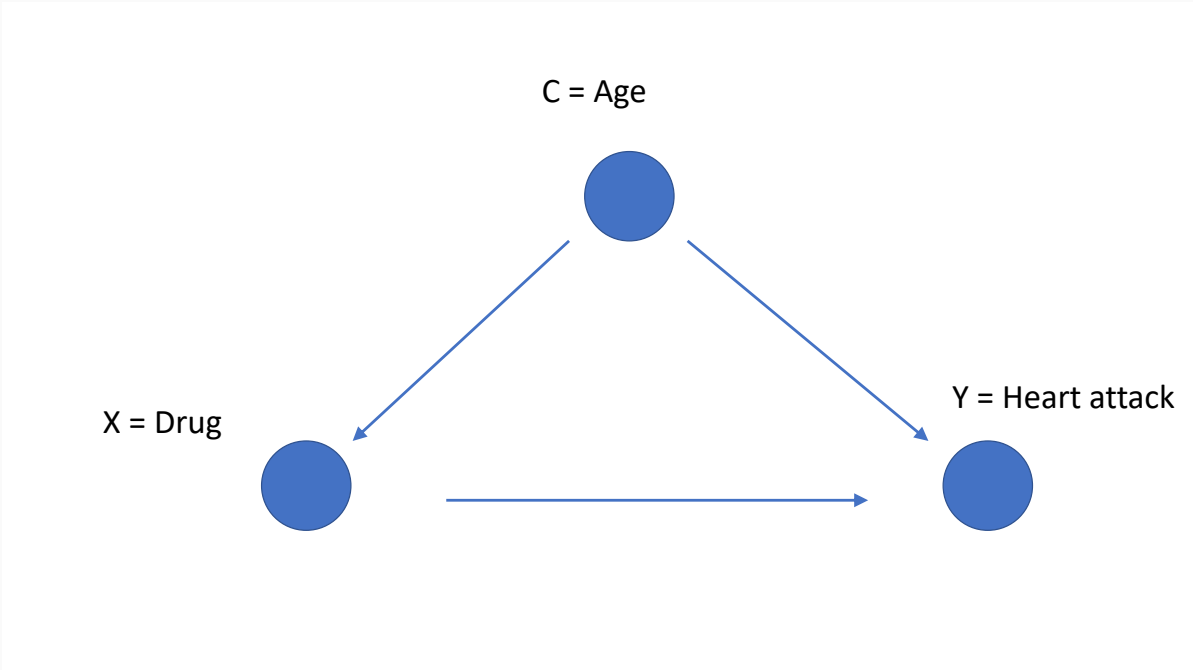
Is treatment recommended?

The data seem to show that treatment was effective overall but damaging to each subgroup.

Pearl and others argue that with with information about the nature of the groups, a causal model can tell us whether to trust the aggregated or disaggregated data.

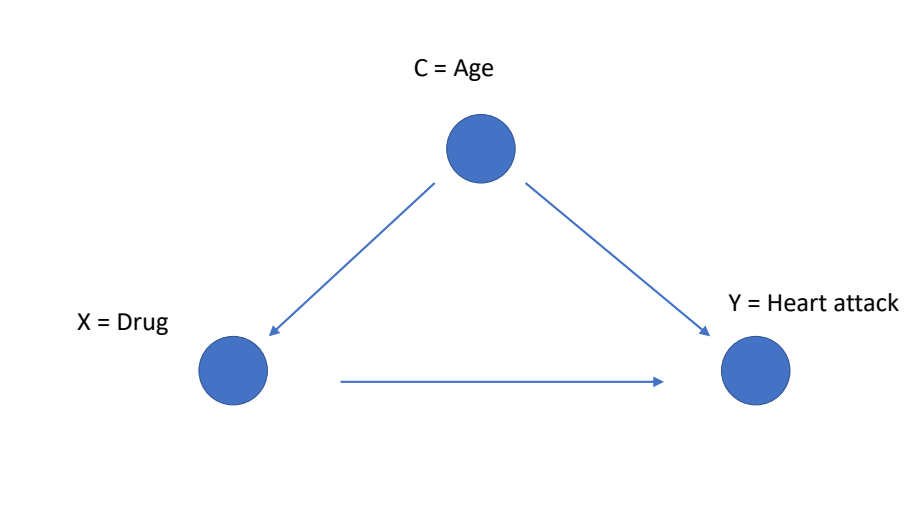
Age

Suppose Group 1 contains younger patients and Group 2 contains older patients Both the treatment and the outcome depend on age (suppose younger patients are more open to trying the drug).



Age

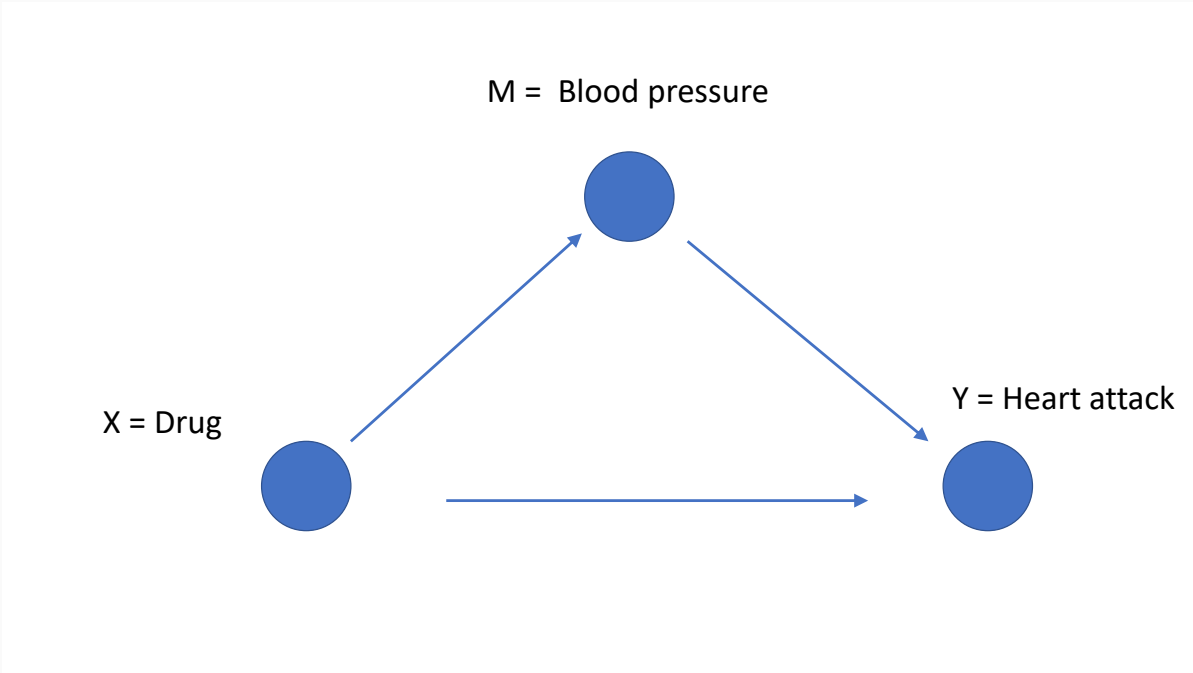
All	Treatment	Control
	82	78
Younger		
	93	95
Older		
	60	70



Conditioning on age, a confounder, is necessary, so the subset-specific results provide the proper recommendation: no treatment.

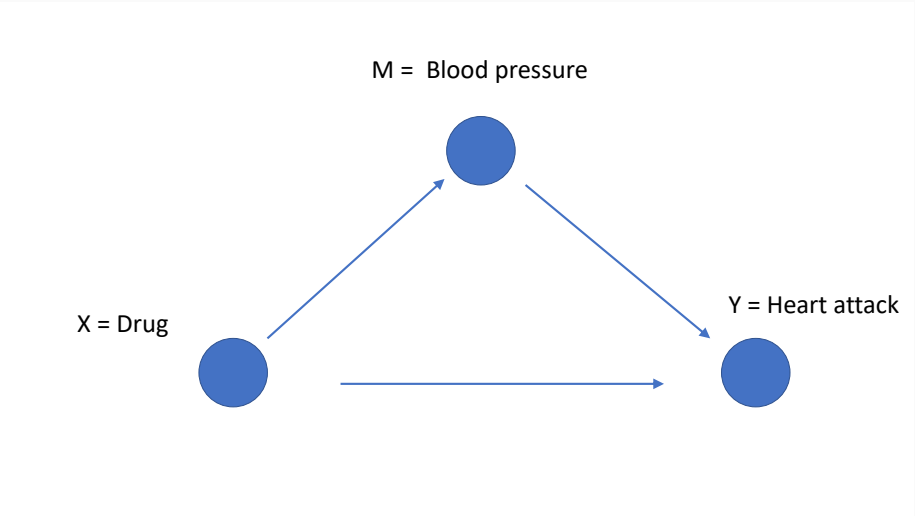
Blood pressure

Suppose the drug operates, in part, by lowering blood pressure, which is a mediator of the drug's effect.



Blood pressure

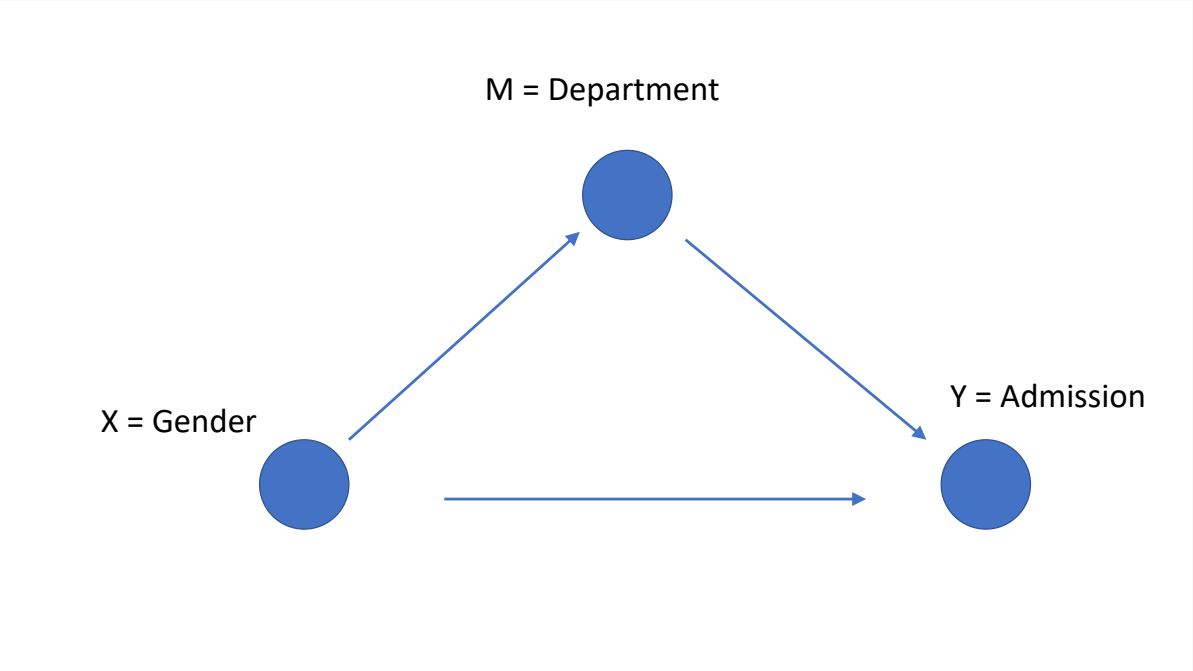
All	Treatment	Control
	82	78
Lower		
	93	95
Higher		
	60	70



Conditioning on blood pressure would disable one of the drug's causal paths. The aggregate results provide the proper recommendation: treatment.

Graduate admissions at UC Berkeley in 1973

Pearl's assessment of the 1973 graduate admission conundrum



We have seen before that conditioning on a mediator is incorrect if we want to estimate the total effect of one variable on another. But in a case of discrimination, according to the court, it is not the total effect but the direct effect that matters.

—Judea Pearl

Bickel, Hammell and O'Connell's assessment

In their study, these authors examine the admissions data in detail.

...and find that an important assumption underlying the statistical test they applied is not satisfied for the aggregate applicant pool.

Assumptions underlying the Pearson chi-squared test

Assumption 1:

In any given discipline male and female applicants do not differ in respect of their intelligence, skill, qualifications, promise, or other attribute deemed legitimately pertinent to their acceptance as student.

> *It is precisely this assumption that makes the study of "sex bias" meaningful, for if we did not hold it any differences in acceptance of applicants by sex could be attributed to differences in their qualifications, promise as scholars, and so on.*
>
> —P.J. Bickel, E.A. Hammell, J.W. O'Connell

What conclusions did Bickel, Hammell and O'Connell draw?

Assumption 2: The sex ratios of applicants to the various fields of graduate study are not importantly associated (or correlated) with any other factors in admission.

Records from the largest departments

Department	Men		Women	
	Applicants	Admitted	Applicants	Admitted
	825	62%	108	82%
	560	63%	25	68%
	325	37%	593	34%
	417	33%	375	35%
	191	28%	393	24%
	373	6%	341	7%

Records from the largest departments

Acceptance rates were relatively high in the largest departments, and they were higher for women than for men...

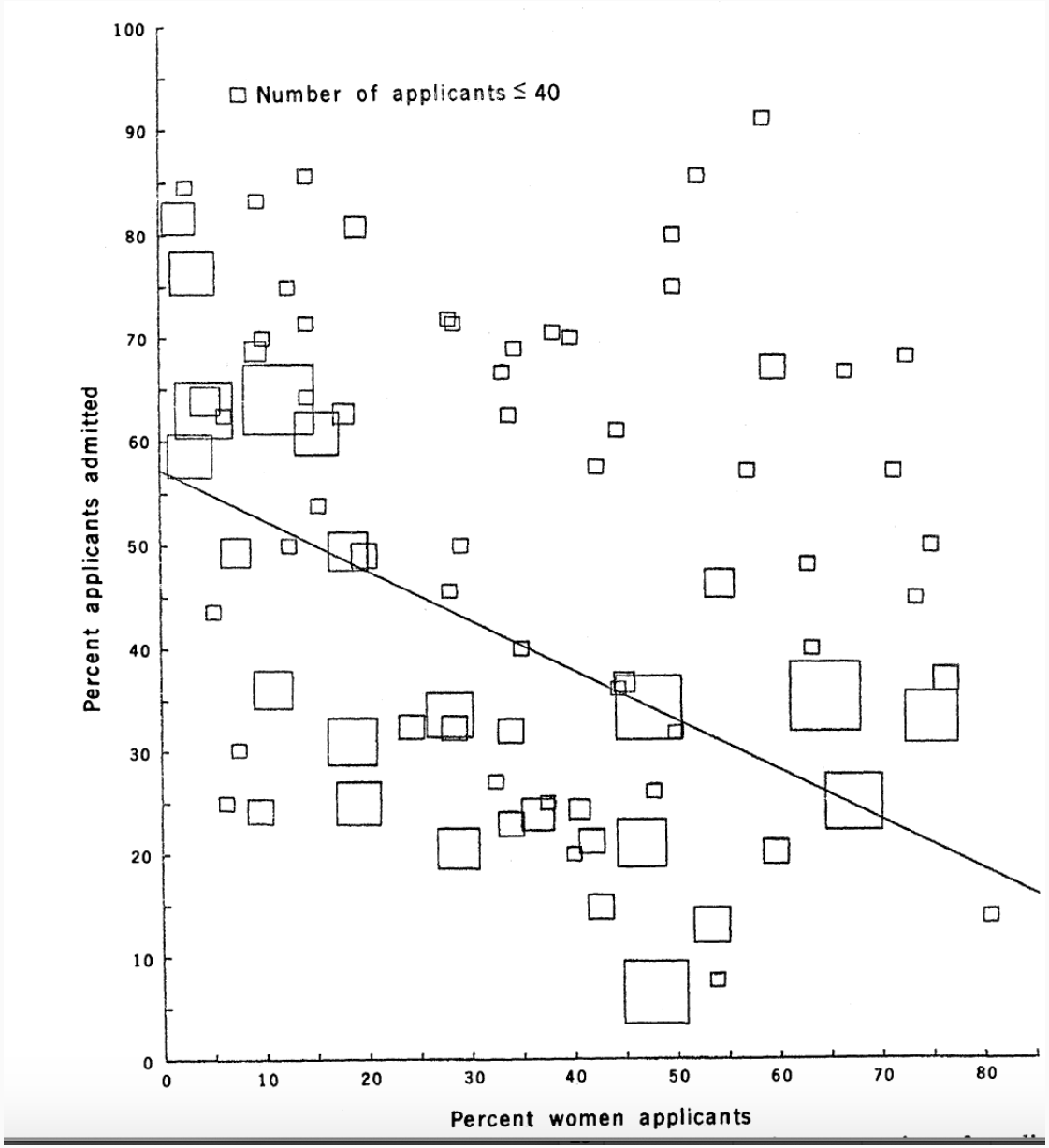
Department	Men		Women	
	Applicants	Admitted	Applicants	Admitted
	825	62%	108	82%
	560	63%	25	68%
	325	37%	593	34%
	417	33%	375	35%
	191	28%	393	24%
	373	6%	341	7%

Records from the largest departments

...but women were severely underrepresented in the applicant pools.

Department	Men		Women	
	Applicants	Admitted	Applicants	Admitted
	825	62%	108	82%
	560	63%	25	68%
	325	37%	593	34%
	417	33%	375	35%
	191	28%	393	24%
	373	6%	341	7%

Relationships among acceptance rates, percentage of female applicants and size of applicant pool.



A statistical test is no more valid than its assumptions

Assumption 2: The sex ratios of applicants to the various fields of graduate study are not importantly associated with any other factors in admission.

The demonstrated falsity of this assumption invalidates the results of the Pearson chi-squared test on the aggregate applicant pool.

After further analysis taking account of the tendency of women to apply to departments with lower acceptance rates, Bickel, Hammell and O'Connell concluded that there was no evidence of bias against women. However...

Final words in the 1975 paper

Women are shunted by their socialization and education toward fields of graduate study that are generally more crowded, less productive of completed degrees, and less well funded, and that frequently offer poorer professional employment prospects.

—P.J. Bickel, E.A. Hammell, J.W. O'Connell

Final words from me

Thoughtfully applied, mathematical tools can provide insight into complex, societal problems.

But these societal problems will be solved only when individuals take responsibility.

The Julia Robinson Mathematics Festival

Julia lives on to inspire young mathematicians

In 2007, Nancy Blachman founded the Julia Robinson Mathematics Festival (JRMF), which sponsors locally organized mathematics events targeting K12 students. The events are designed to introduce students to the richness and beauty of mathematics in a collaborative and non-competitive forum.

Thank you Berkeley Math Circle

and thank you Zvezda

References

Bickel, P. J., Hammel, E. A. & O'Connell, J. (1975), 'Sex bias in graduate admissions: Data from Berkeley', *Science* **187**, 398–403.

Moore, C. C. (2007), *Mathematics at Berkeley: A History*, A.K. Peters, Ltd.

Pearl, J. & Mackenzie, D. (2018), *The Book of Why*, Basic Books.

Simpson, E. H. (1951), 'The interpretation of contingency tables', *Journal of the Royal Statistical Society, Series B* **13**, 238–241.

Yule, U. (1903), 'Notes on the theory of association of attributes in statistics', *Biometrika* **2**(2), 121–134.