Introduction to data science & artificial intelligence (INF7100)

Arthur Charpentier

#121 Simpson's Paradox & Ecological Fallacy

été 2020

Graduate admissions data, Berkeley, 1973

Graduate admissions data from Berkeley, 1973

- men : 8442 applications, 44% admission rate
- women : 4321 applications, 35% admission rate

Discrimination towards women ?

		A	B	C	D	E	F
M	applied	825	560	325	417	191	373
	admitted	62%	63%	37%	33%	28%	6%
F	applied	108	25	593	375	393	341
	admitted	82%	68%	34%	35%	24%	7%

see Bickel *et al.* (1975, Sex bias in graduate admissions)

(Fake) Hospital Data

	hosp. A	hosp. B
total	1000	1000
survivors	800	900
deads	200	100
rate (%)	80%	90%

	healthy	
	hosp. A	hosp. B
total	600	900
survivors	590	870
deads	10	30
rate (%)	98%	97%

	sick	
	hosp. A	hosp. B
total	400	100
survivors	210	30
deads	190	70
rate (%)	53%	30%

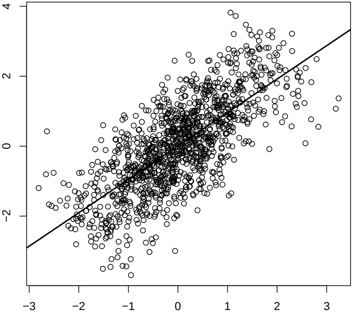
Mathematics of Simpson's Paradox

Heuristically, it is possible to have

$$\frac{a}{A} \leq \frac{b}{B} \text{ and } \frac{c}{C} \leq \frac{d}{D}$$

and at the same time

$$\frac{a+c}{A+C} \geq \frac{b+d}{B+D}$$



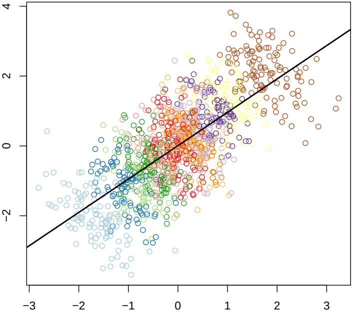
Mathematics of Simpson's Paradox

Heuristically, it is possible to have

$$\frac{a}{A} \leq \frac{b}{B} \text{ and } \frac{c}{C} \leq \frac{d}{D}$$

and at the same time

$$\frac{a+c}{A+C} \geq \frac{b+d}{B+D}$$



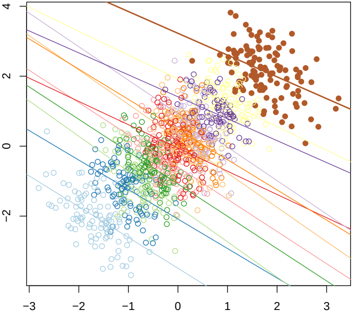
Mathematics of Simpson's Paradox

Heuristically, it is possible to have

$$\frac{a}{A} \leq \frac{b}{B} \text{ and } \frac{c}{C} \leq \frac{d}{D}$$

and at the same time

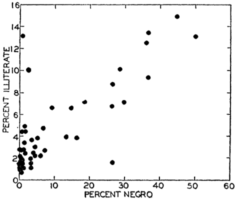
$$\frac{a+c}{A+C} \geq \frac{b+d}{B+D}$$

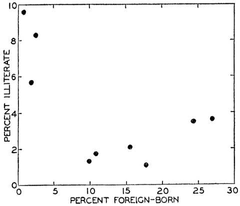


Ecological Fallacy

An ecological fallacy is a formal fallacy in the interpretation of statistical data that occurs when inferences about the nature of individuals are deduced from inferences about the group to which those individuals belong, via wikipedia)

See Robinson's Ecological Correlations and the Behavior of Individuals the individual correlation depends upon the internal frequencies of the within-areas individual correlations, while the ecological correlation depends upon the marginal frequencies of the within-areas individual correlation





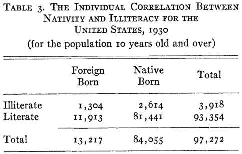
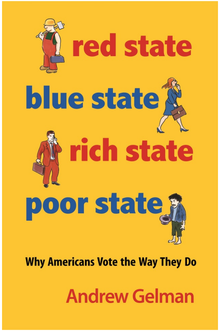


Table 3. The Individual Correlation Between Nativity and Illiteracy for the United States, 1930
(for the population 10 years old and over)

	Foreign Born	Native Born	Total
Illiterate	1,304	2,614	3,918
Literate	11,913	81,441	93,354
Total	13,217	84,055	97,272

Ecological Fallacy

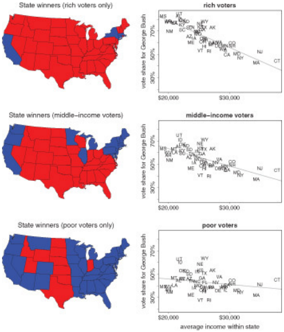
Very important concept in political science

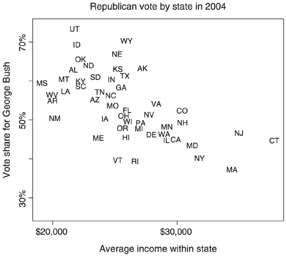


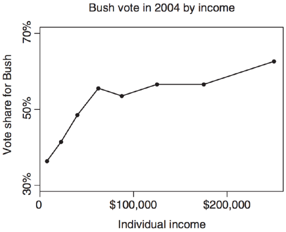
red state
blue state
rich state
poor state

Why Americans Vote the Way They Do

Andrew Gelman


State winners (rich voters only)
rich voters
State winners (middle–income voters)
middle–income voters
State winners (poor voters only)
poor voters
vote share for George Bush
70%
50%
30%
$20,000
$30,000
average income within state





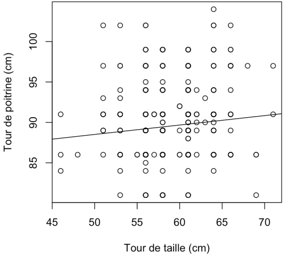
Gelman's Red State, Blue State, Rich State, Poor State: Why Americans Vote the Way They Do)

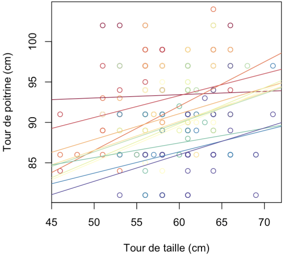
Playboy: Individual vs. Temporal Data

Are bust/chest and waist correlated measures ?

Dataset $n = 659$ observations ($\sim$ 55 years) of Playboy's playmate (inspired by Shapely centrefolds. Are women changing or is Playboy?).

- x_i: waist (cm)
- y_i: bust (cm)
- t_i: date



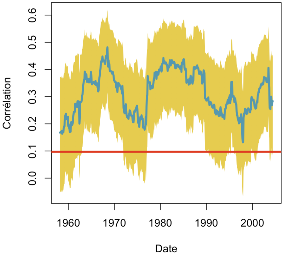


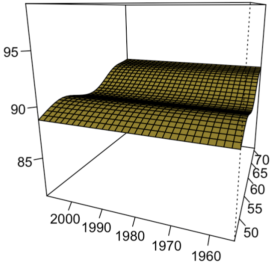
PLAYBOY
ENTERTAINMENT FOR MEN
COLLECTOR'S EDITION
www.playboy.com • NOVEMBER 2009
THE DEVIL IN MARGE SIMPSON
VICTORIA'S SECRET STUNNER ALINA PUSCAU
TERRIFYING NEW WORK BY STEPHEN KING
OUR ANGEL A TRIBUTE TO FARRAH FAWCETT
BADLANDS A PLAYBOY DISPATCH FROM THE TEXAS BORDER
THE MADNESS OF TRACY MORGAN
THE INTERVIEW BENICIO DEL TORO

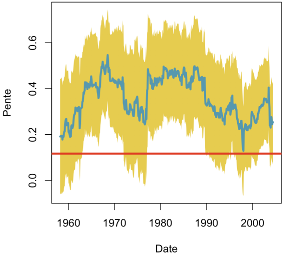
Playboy: Individual vs. Temporal Data

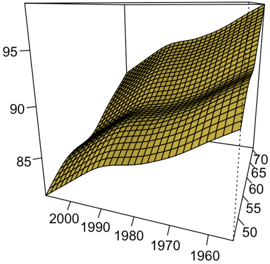
over 55 years
$\mathrm{cor}(x_i, y_i) \simeq 0.1$

underestimation !



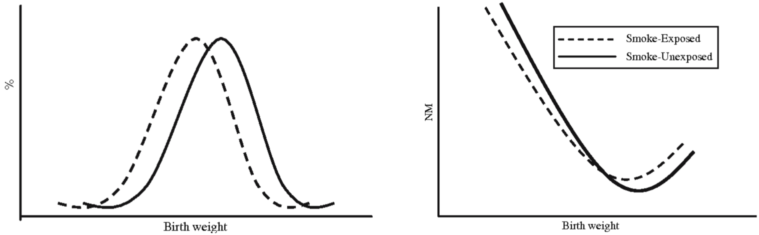






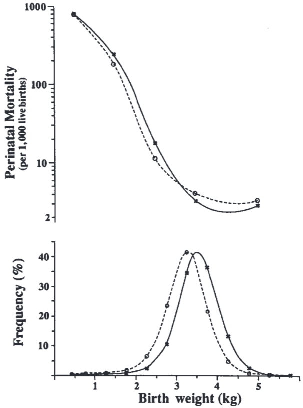
Birth Weight Paradox

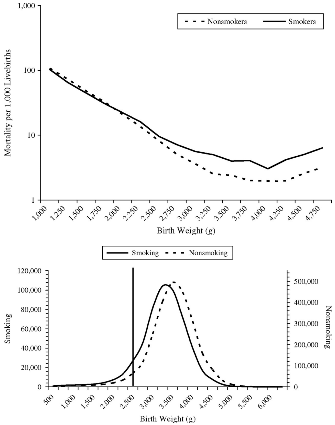
The low birth-weight paradox is an apparently paradoxical observation relating to the birth weights and mortality rate of children born to tobacco smoking mothers. Low birth-weight children born to smoking mothers have a **lower** infant mortality rate than the low birth weight children of non-smokers



via From causal diagrams to birth weight-specific curves of infant mortality and wikipedia

Birth Weight Paradox





see On the importance - and the unimportance - of birthweight, The Birth Weight “Paradox” Uncovered? and Big data : passer d’une analyse de corrélation à une interprétation causale