Structural equation modeling from Scientific Software International home of LISREL Student editions, documentation, examples, etc

Graphical Models, Causal Diagrams

Original Epi exposition. Greenland S., Pearl J., and Robins J.M. Causal diagrams for epidemiologic research. Epidemiology, 10(1):37-48, 1999.

Richardson and Robbins attempts at unification. Single World Intervention Graphs: A Primer Longer version

Graphical Markov Models: Overview Nanny Wermuth and D.R. Cox

C. Shalizi. Advanced Data Analysis from an Elementary Point of View, 2017; Chapter 24 (except 24.2)

2. Interpreting Associations: Spurious Correlation and Simpson's Paradox

Lecture topics

Class handout: Third Variables (page 1)

1. Spurious Correlation: some historical notes; partial and part correlations. (class slides)
2. Simpson's paradox wiki page (dichotomous outcome slide)

Research Examples

From Week 0 intro: Secret to Winning a Nobel Prize? Eat More Chocolate (Time) Publication: Chocolate Consumption, Cognitive Function, and Nobel Laureates Franz H. Messerli, M.D. N Engl J Med 2012; 367:1562-1564 October 18, 2012

Correlation study. New study finds sweary people are more honest Publication: Frankly, We Do Give a Damn: The Relationship Between Profanity and Honesty, Social Psychological and Personality Science.

Size does matter. Bigger is smarter: Overall, not relative, brain size predicts intelligence. Publication: Deaner RO, Isler K, Burkart J, van Schaik C: Overall Brain Size, and Not Encephalization Quotient, Best Predicts Cognitive Ability across Non-Human Primates. Brain Behav Evol 2007;70:115-124 (DOI: 10.1159/000102973)

Spurious Correlation

perennial favorite Spurious Correlation examples

Correlations Genuine and Spurious in Pearson and Yule, John Aldrich *Statistical Science*, Vol. 10, No. 4. (Nov., 1995), pp. 364-376. Jstor link

Spurious Correlation: A Causal Interpretation. Herbert A. Simon *Journal of the American Statistical Association*, Vol. 49, No. 267. (Sep., 1954), pp. 467-479. Jstor link

Simpson's Paradox

Kidney stone example Confounding and Simpson's paradox, BMJ, vol309, 1480-1, 1994

UC Berkeley admissions, Racial bias in Death Penalty in wiki page.

R Implentations and Resources.

Spurious correlation?

R-Package ppcor October 29, 2012 Title Partial and Semi-partial (Part) correlation

Simpson's Paradox.

R-package `Simpsons`. Frontiers in Psychology. 2013; 4: 513. Simpson's paradox in psychological science: a practical guide

Week 5 Review Questions

Path Analysis and Friends

Question 1. Freedman, Blau-Duncan example in class handout.

Freedman links "Stat Models for Causation" (pp3-4) or Freedman text Ch6 (revised)

Replicate class handout computations for the path analysis

Plus questions from Freedman text: scan of pp.80-81 at (pp86-7 revised ed) http://web.stanford.edu/~rag/stat209/DAFtextp8081.pdf (includes standardization material Hookes Law on week 4 class handout).

Freedman pp80-1 (set A) prob 1 prob 5 prob 6 prob 8

pdf scan also includes freedman Set E, p.97 prob 4(a,b) (p.103 revised)

Solution for question 1

Question 2. Causal Models of Publishing Productivity

freedman p.101 prob 5 (page 107 in revised version)

This Homework problem considers one of the path analysis models from "Causal Models of Publishing Productivity in Psychology", Rogers & Maranto, J. Applied Psychology, 1989, 74(4), 636-649.

direct link to paper http://content.apa.org/journals/apl/74/4/636.pdf

The path analysis conducted by the authors from a sample of 86 men and 76 women is shown in p.101 of Freedman's text and on page 647 of the publication; that page also exists at http://www-stat.stanford.edu/~rag/stat209/pathpage647.pdf

You do have the correlation matrix from adding Table 7 fits and residuals. But here all the problem asks you to do is look at and consider the usefulness of this analysis. Note they don't display the disturbance paths so we don't get a look at Rsq values.

What are the predictors of Pubs (direct effects) in this picture?

What are the predictors of Cites (direct effects) in this picture?

The diagram provides estimates of supposed causal effects ("causal model of publishing" is the article title); it displays regression coeffs , with coefficient estimates shown on the edges.

Consider a "productive researcher" to be defined in terms of the number of publications and the number of cites. The good news is that ability "affects" pubs and cites with a positive coefficient in each case. Therefore, higher ability leads to a more "productive researcher", according to the causal path gospel. Some bad news is that sex is a predictor of pubs with a large coefficient value. However, it is likely that there are confounding variables between sex and pubs.

Solution for question 2

Question 3. Longitudinal path analysis (based on the Goldstein example)

Apply the path analysis model taken from Goldstein (1979) (in class handouts week5,also Rogosa eq 2 1988, "casual models...) to verify results for path coefficients in eq 3 of Rogosa (1988) (also in handouts).

Data are given in http://statweb.stanford.edu/~rag/stat209/casualdat using the top frame of 40 observations for variables (perfectly measured) Xi(1) Xi(3) Xi(5) and taking the times of observation to be 1 3 5 respectively.

These data are in wide form--each row is a subject.

You can verify, if you like, that each subject's data lies on a straight-line (constant rate of change)

Try `pairs` on the three measurements to see the scatter plots over persons.

Obtain values for the path coefficients and the muliple correlations for the regression fits.

Can you obtain standard errors for the path coefficients for this small sample?

Any interpretations of the results from the path analysis?

Solution for question 3

Question 4. ENRICHMENT ITEM, Structural Equation Models, Method-of-moments for two-variable, two-indicator model

Problem 4 is an "enrichment" item, and you may want to look at the solution which is linked.

For latent variable models with multiple indicators How does structural equation model (latent vars) methods provide a correction for measurement error?

Method-of-moments for two-variable, two-indicator model

For the Structural Equation Models handout from Joreskog book, which is linked in the week 5 lecture materials (class handout) but we did not take up in detail in class, obtain parameter estimates for the no-correlated error version (9 parameters, top covariance matrix) in terms of the sample variance and covariances among the four indicators (y_ij).

Brute force substitution will get you a non-optimal estimate, suffices for instructional purposes.

Solution for question 4

Spurious Correlation

Question 5. Spurious correlation Consider the spurious correlation (common cause type) discussed class week 5. Additional examples from class page links in Simon (1954) or Aldrich (1995, sec 7 "illusory")

Correlation and Causation

a comment

Stephen Stigler

ABSTRACT Some purely methodological comments are made on the pitfalls and difficulties in making causal inferences from observational data, including in studies of disparity in medicine. The ideas of spurious correlation and measurement error are discussed with an eye towards their impact upon inferences about causality, and cautions are offered about over-reliance upon testing hypotheses.

STATISTICIANS HAVE LONG STRUGGLED to deduce causal relationships from correlation; that is, to determine a mechanistic relationship from purely empirical evidence of association. That is also the essential goal of the methods that are being discussed here. There are other important issues to be sure, such as whether or not the deduced causal relationship is illegal or immoral, but statistically that is secondary to the study of the nature of the relationship. The answer to the strict question "Can cause be deduced from correlation?" is generally "no." But necessity being the mother of invention, we do it anyway, by weakening the question to one permitting a more positive answer: "Under certain restrictive assumptions, can we conclude causation from correlation, beyond a reasonable doubt?" In that form a great deal of methodological progress has been made, including by one of this year's Nobel Prize–winning economists, Clive Granger. But the answer to the strict question remains "no," and it may be worthwhile recalling why that is so.

Department of Statistics, University of Chicago, Room 102, 1118 E. 58th Street, Chicago, IL 60637. E-mail: stigler@galton.uchicago.edu.

Perspectives in Biology and Medicine, volume 48, number 1 supplement (winter 2005):S88–S94

WEEK 2
spurious

There are a number of pitfalls for the unwary in deducing causation from correlation, but I will only give a few examples. The first is historical, and it involves the origin of the major worry implicitly addressed by Ian Ayres and most other most modern treatments of this question: spurious correlation. Karl Pearson discovered spurious correlation some time around 1896, during a study in craniometry (Pearson 1897; Pearson and Bramley-Moore 1899). In the particular investigation under consideration, Pearson studied measurements of a large collection of skulls from the Paris Catacombs, with the goal of understanding the interrelationships among the measurements. For each skull, his assistant measured the length and the breadth, and computed for different sets of skulls the correlation coefficient between these measures as an indication of their interrelationship. If constancy of *shape* prevailed for skulls of different sizes, then a positive correlation of length and breadth would be expected. If the *volume* contained within the skull were more or less constant (so that long skulls would be narrow and short skulls wide), then a negative correlation would be expected.

The correlation from Pearson's Parisian skull data turned out to be significantly greater than zero, about 0.2. But before Pearson could use his own (only medium sized) skull to speculate on the implications of this, the discovery was deflated by his noticing that if the skulls were divided into male and female, the correlation disappeared. Pearson recognized the general nature of this phenomenon and brought it to the attention of the world. When two measurements are correlated, this may be because they are both related to a third factor that has been omitted from the analysis. In Pearson's case, skull length and skull breadth were essentially uncorrelated if the factor "sex" were incorporated in the analysis. If "sex" were omitted, they were positively correlated. Figures 1 to 3 illustrate this phenomenon.

A second pitfall is measurement error. It is common in discrimination studies, including those Ayres reviews, to use regression methods to attempt to "correct" or "control" for group differences that may legitimately influence the outcome variable, even absent any discriminatory practices. For example, in a study of salary differences between two groups, it might be judged appropriate to "correct" for education, as reflected by years of formal schooling, by regressing salary on years of formal education. But as those of us who spend time in the classroom watching sleeping students know all too well, *formal* education does not equal *effective* education. If the salary judgment is made on the basis of a collection of assessments and interviews reflecting effective education, and the regression is performed using formal education as an imperfect measure of effective education, then if the two groups have different mean education levels a severe bias is introduced. The slopes of the two regression lines are attenuated (made flatter, moved closer to zero slope), and the less-educated group will appear to have been discriminated against. Statistical "proof" of salary discrimination could appear where in truth there is none. Figure 4 illustrates this phenomenon.

These two pitfalls may appear in medical studies in quite a variety of ways.

Larry Husten, Contributor
I'm a medical journalist covering cardiology news.

PHARMA & HEALTHCARE | 10/10/2012 @ 5:02PM | 13,808 views

Chocolate And Nobel Prizes Linked In Study

You don't have to be a genius to like chocolate, but geniuses are more likely to eat lots of chocolate, at least according to a new paper published in the August *New England Journal of Medicine*. Franz Messerli reports a highly significant correlation between a nation's per capita chocolate consumption and the rate at which its citizens win Nobel Prizes.

Chocolate sampler (Peter Dazeley/Getty Images)

Building on research raising the possibility that the flavanols in chocolate may enhance cognitive performance, Messerli "wondered

Larry Husten
Contributor

In addition to writing CardioBrief, I cover cardiology news for CardioExchange, a social media website for cardiologists published by the New England Journal of Medicine. I was the editor of TheHeart.Org from its inception in 1999 until December 2008. Following the purchase of TheHeart.Org by WebMD in 2005 I became the editorial director of WebMD professional news, encompassing TheHeart.Org and Medscape Medical News. Prior to joining TheHeart.Org I was a freelance medical journalist and wrote for a wide variety of medical and computer publications. In 1994-1995 I was a Knight Science Journalism Fellow at MIT. I have a PHD in English from SUNY Buffalo and I drove a taxicab in New York City before embarking on a career in medical journalism. You can follow me on Twitter at: @cardiobrief.

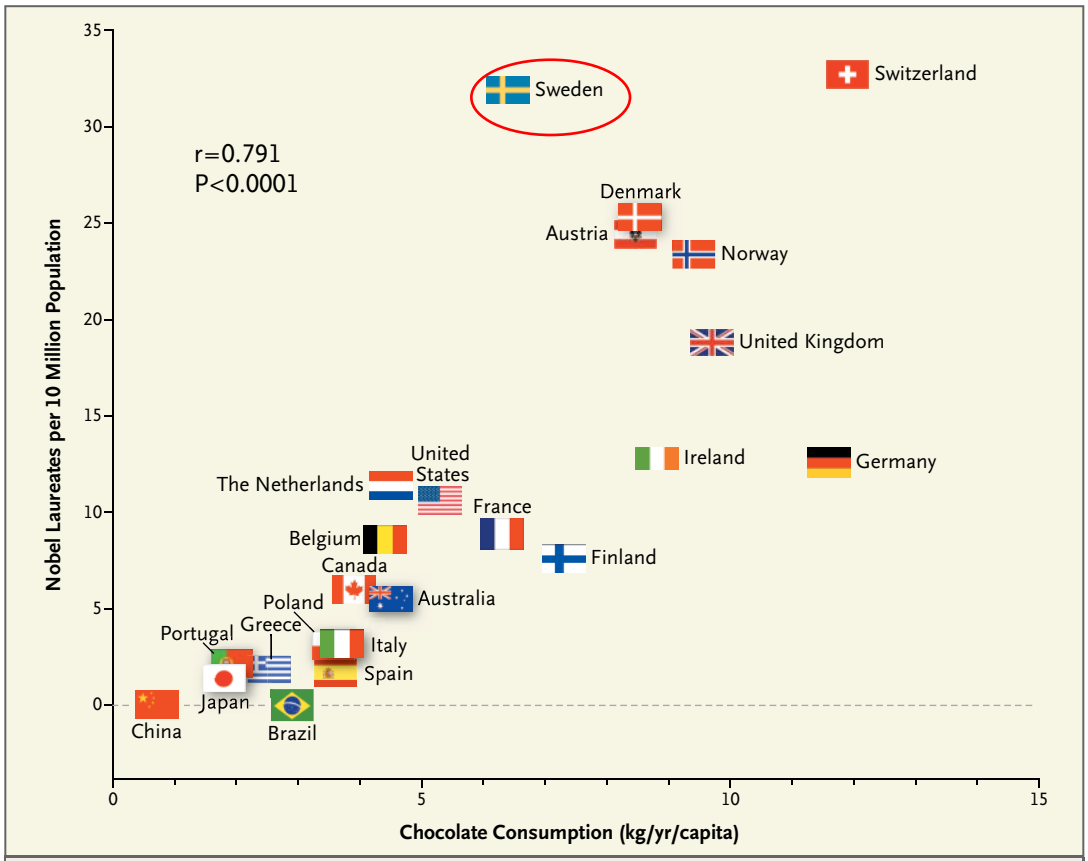


Figure 1. Correlation between Countries' Annual Per Capita Chocolate Consumption and the Number of Nobel Laureates per 10 Million Population.

DISCUSSION

The principal finding of this study is a surprisingly powerful correlation between chocolate intake per capita and the number of Nobel laureates in various countries. Of course, a correlation between X and Y does not prove causation but indicates that either X influences Y, Y influences X, or X and Y are influenced by a common underlying mechanism. However, since chocolate consumption has been documented to improve cognitive function, it seems most likely that in a dose-dependent way, chocolate intake provides the abundant fertile ground needed for the sprouting of Nobel laureates. Obviously, these findings are hypothesis-generating only and will have to be tested in a prospective, randomized trial.

The only possible outlier in Figure 1 seems to be Sweden. Given its per capita chocolate consumption of 6.4 kg per year, we would predict that Sweden should have produced a total of about 14 Nobel laureates, yet we observe 32. Considering that in this instance the observed number exceeds the expected number by a factor of more than 2, one cannot quite escape the notion that either the Nobel Committee in Stockholm has some inherent patriotic bias when assessing the candidates for these awards or, perhaps, that the Swedes are particularly sensitive to chocolate, and even minuscule amounts greatly enhance their cognition.

A second hypothesis, reverse causation — that is, that enhanced cognitive performance could stimulate countrywide chocolate consumption — must also be considered. It is conceivable that persons with superior cognitive function (i.e., the cognoscenti) are more aware of the health benefits of the flavanols in dark chocolate and are therefore prone to increasing their consumption. That receiving the Nobel Prize would in itself increase chocolate intake countrywide seems unlikely, although perhaps celebratory events associated with this unique

New study finds sweary people are more honest

Posted 14 days ago by Darin Graham in news Like

UPVOTE

Picture: JASON MERRITT/GETTY IMAGES

For those of us frequently told off for using foul language - a new defence might work:

> **"I was just being honest."**

People who swear may be more trustworthy, according to researchers.

In the three-part study, published in the *Social Psychological and Personality Science* journal - an international team led by Gilad Feldman of Maastricht University in the Netherlands analysed swearing in society.

First, the team studied 276 people to find out how they curse. They asked participants to list their favourite swear words and to 'self-report' their everyday use of profanity. Researchers also asked the subjects to note down the emotions they associate with those swear words - anger, exasperation or fear for example.

The foul-mouthed test subjects were also asked to fill in a psychological survey to gauge their honesty which helped rank how likely they were to lie.

The team found that those who lied less wrote down a higher number of frequently used swear words.

Article

Frankly, We Do Give a Damn: The Relationship Between Profanity and Honesty

Social Psychological and Personality Science
2017, Vol. 8(7) 816-826

DOI: 10.1177/1948550616681055
journals.sagepub.com/home/spp

Gilad Feldman[1], Huiwen Lian[2], Michal Kosinski[3], and David Stillwell[4]

Abstract

There are two conflicting perspectives regarding the relationship between profanity and dishonesty. These two forms of norm-violating behavior share common causes and are often considered to be positively related. On the other hand, however, profanity is often used to express one's genuine feelings and could therefore be negatively related to dishonesty. In three studies, we explored the relationship between profanity and honesty. We examined profanity and honesty first with profanity behavior and lying on a scale in the lab (Study 1; $N = 276$), then with a linguistic analysis of real-life social interactions on Facebook (Study 2; $N = 73{,}789$), and finally with profanity and integrity indexes for the aggregate level of U.S. states (Study 3; $N = 50$ states). We found a consistent positive relationship between profanity and honesty; profanity was associated with less lying and deception at the individual level and with higher integrity at the society level.

Keywords

profanity, honesty, cursing, integrity

> Frankly my dear, I don't give a damn.
>
> Gone with the Wind (1939)

Profane as it is, this memorable line by the character Rhett Butler in the film *Gone with the Wind* profoundly conveys Butler's honest thoughts and feelings. However, it was the use of this profane word that led to a US$5,000 fine against the film's production for violating the Motion Picture Production Code. This example reveals the conflicting attitudes that most societies hold toward profanity, reflected in a heated debate taking place in online forums and media in recent years—with passionate views on both sides. For example, the website *debate.org*, which conducts online polls and elicits general public opinions on popular online debates, has many comments on the issue, with a 50–50 tie between the two views (Are people who swear more honest?, 2015). This public debate reflects an interesting question and mirrors the academic discussion regarding the nature of profanity. On the one hand, profane individuals are widely perceived as violating moral and social codes and thus deemed untrustworthy and potentially antisocial and dishonest (Jay, 2009). On the other hand, profane language is considered as more authentic and unfiltered, thus making its users appear more honest and genuine (Jay, 2000). These opposing views on profanity raise the question of whether profane individuals tend to be more or less dishonest.

Profanity

Profanity refers to obscene language including taboo and swear words, which in regular social settings are considered inappropriate and in some situations unacceptable. It often includes sexual references, blasphemy, objects eliciting disgust, ethnic–racial–gender slurs, vulgar terms, or offensive slang (Mabry, 1974). The interest in understanding the psychological roots of the use of profanity dates back to as far as the early 20th century (Patrick, 1901), yet the literature in this domain is scattered across different scientific fields with only recent attempts to connect the findings into a unified framework (Jay, 2009).

The reasons for using profanity depend on the person and the situation, yet profanity is commonly related to the expression of emotions such as anger, frustration, or surprise (Jay & Janschewitz, 2008). The spontaneous use of profanity is usually the unfiltered genuine expression of emotions, with the most extreme type being the bursts of profanity (i.e., *coprolalia*) accompanying the Tourette syndrome (Cavanna & Rickards, 2013). The more controllable use of profanity often helps to convey world views or internal states or is used to insult an object, a view, or a person (Jay, 2009). Speech involving

[1] Department of Work and Social Psychology, Maastricht University, Maastricht, the Netherlands
[2] Department of Management, Hong Kong University of Science and Technology, Clearwater Bay, Kowloon, Hong Kong
[3] Graduate School of Business, Stanford University, Stanford, CA, USA
[4] Judge School of Business, University of Cambridge, Cambridge, United Kingdom

Corresponding Author:
Gilad Feldman, Department of Work and Social Psychology, Maastricht University, Maastricht 6200MD, the Netherlands.
Email: gilad.feldman@maastrichtuniversity.nl

Table 1. Study 1: Means, Standard Deviations, and Correlations for Variables.

Variables	Mean	*SD*	1	2	3	4	5
1. Honesty	7.63	3.00	(.79)				
2. Profanity self-report	6.51	2.56	.34***	(.84)			
3. Profanity behavioral 1	4.09	2.61	.20**	.46***	(—)		
4. Profanity behavioral 2	1.60	1.62	.13*	.41***	.45***	(—)	
5. Age	40.71	12.75	−.13*	−.34***	−.05	−.08	(—)
6. Gender	1.62	0.49	−.06	−.03	−.07	−.04	.08

Note. $N = 276$. Gender coding: 1 = *male*, 2 = *female*. Scale α coefficients are on the diagonal. Profanity behavioral 1 = number of most frequently used curse words written; profanity behavioral 2 = number of most liked curse words written.
*$p < .05$. **$p < .01$. ***$p < .001$.

171 females). The exclusion of participants had no significant impact on the reported effect sizes or p values below. Participants self-reported profanity use in everyday life: given the opportunity to use profanity, rated reasons for the use of profanity, and answered a lie scale.

Measures

Profanity use behavioral measure. In 2 items, participants were asked to list their most commonly used and favorite profanity words: "Please list the curse words you [1 – use; 2 – like] the most (feel free, don't hold back)." By giving participants an opportunity to curse freely, we expected that the daily usage and enjoyment of profanity would be reflected in the total number of curse words written. Participants' written profanity was counted and coded by the first author and a coder unrelated to the project, who was unaware of the study hypotheses and data structure. The interrater reliability was .91 (95% confidence interval [CI] [.87, .94]) for most commonly used curse words and .93 (95% CI [.91, .97]) for favorite curse words, indicating a very high level of agreement.

Profanity self-reported use. To supplement the behavioral measures, we also added self-reported use of profanity. Participants self-reported their everyday use of profanity (Rassin & Muris, 2005) using 3 items: "How often do you curse (swear/use bad language)" (1) "verbally in person (face to face)," (2) "in private (no one around)," and (3) "in writing (e.g., texting/messaging/posting online/emailing"; 1 = *never*, 2 = *once a year or less*, 3 = *several times a year*, 4 = *once a month*, 5 = *2–3 times a month*, 6 = *once a week*, 7 = *2–3 times a week*, 8 = *4–6 times a week*, 9 = *daily*, 10 = *a few times a day*; $\alpha = .84$).

Reasons for profanity use. Following Rassin and Muris (2005), we also asked participants to rate reasons for their use of profanity (0 = *never a reason for me to swear*; 5 = *very often a reason for me to swear*) and asked questions regarding the general perceived reasons for using profanity (0 = *not at all*; 5 = *to a very large extent*; see Online Supplemental Materials).

Honesty. Honesty was measured using the Lie subscale of the Eysenck Personality Questionnaire Revised short scale (Eysenck, Eysenck, & Barrett, 1985). The Lie subscale is one of the most common measures for assessing individual differences in lying for socially desirable responding (Paulhus, 1991). The Lie scale includes 12 items, such as "If you say you will do something, do you always keep your promise no matter how inconvenient it might be?" and "Are all your habits good and desirable ones?" (dichotomous Yes/No scale). In these examples, positive answers are considered unrealistic and therefore most likely a lie ($\alpha = .79$). The Lie scale was reversed for the honesty measure.

Results

The means, standard deviations, and correlations for the honesty and profanity measures are detailed in Table 1. Honesty was positively correlated with all profanity measures, meaning that participants lied less on the Lie scale if they wrote down a higher number of frequently used ($r = .20$, $p = .001$; CI [.08, .31]) and liked curse words ($r = .13$, $p = .032$; CI [.01, .24]) or self-reported higher profanity use in their everyday lives ($r = .34$, $p < .001$; CI [.23, .44]), even when controlling for age and gender (Behavioral 1: partial $r = .20$, $p = .001$; CI [.08, .31]; Behavior 2: partial $r = .12$, $p = .049$; CI [.001, .24]; self-report: partial $r = .32$, $p < .001$; CI [.21, .42]).

We asked participants to rate their reasons for use of profanity. The reasons that received the highest ratings were the expression of negative emotions ($M = 4.09$, $SD = 1.33$), habit ($M = 3.08$, $SD = 1.82$), and an expression of true self ($M = 2.17$, $SD = 1.73$). Participants also indicated that in their personal experience, profanity was used for being more honest about their feelings ($M = 2.69$, $SD = 1.72$) and dealing with their negative emotions ($M = 2.57$, $SD = 1.64$). Profanity received a lower rating as a tool for insulting others ($M = 1.41$, $SD = 1.53$) as well as for being perceived as intimidating or insulting ($M = 1.12$, $SD = 1.36$). This supports the view that people regard profanity more as a tool for the expression of their genuine emotions rather than being antisocial and harmful.

Study 2—Naturalistic Deceptive Behavior on Facebook

Study 1 provided initial support for a positive relationship between profanity use and honesty, with the limitations of lab

Mail Online

Blondes have more funds: How reaching for the bleach could see you earn £600 MORE than brunette colleagues

By Daily Mail Reporter
Last updated at 12:33 PM on 30th December 2010

Despite having an 'airhead' reputation, girls with lighter locks bring home around £600 a year more than brunettes or red-heads, a Superdrug study has revealed.

The report found the average blonde takes home £23,150 - or £1,408 per month - compared to £22,586 for brunettes and £22,327 for red-heads.

Many fair-headed girls also admitted they didn't mind their ditzy image, with some even playing up to it.

The Internet Blowhard's Favorite Phrase

Why do people love to say that correlation does not imply causation?

By Daniel Engber | Posted Tuesday, Oct. 2, 2012, at 8:33 AM ET

Karl Pearson, English mathematician and eugenicist, in 1912
Photo by Wikimedia Commons.

Depressed people send more email. They spend more time on Gchat. Researchers at the Missouri University of Science and Technology recently assessed some college students for signs of melancholia then tracked their behavior online. "We identified several features of Internet usage that correlated with depression," they said. Sad people use IM and file-share. They play video games. They surf the Web in their own, sad way.

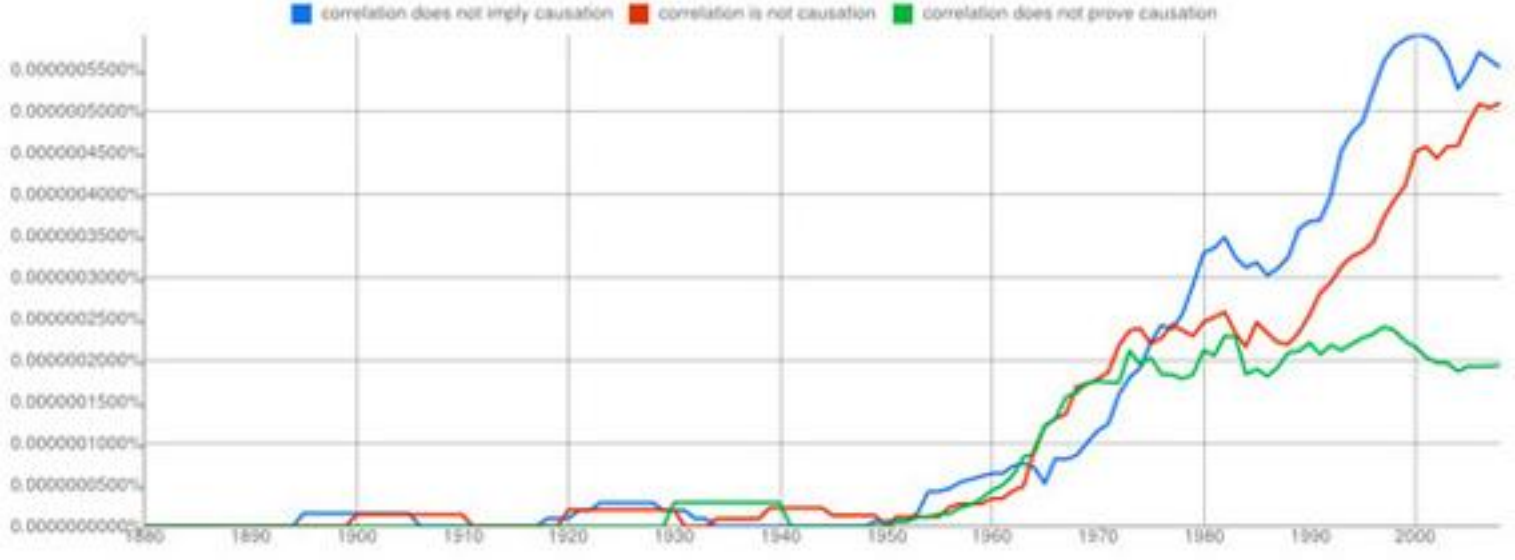
Not everyone found the news believable. "Facepalm. Correlation doesn't imply causation," wrote one unhappy Internet user. "That's pretty much how I read this too... correlation is NOT causation," agreed a Huffington Post superuser, seemingly distraught. "I was surprised not to find a discussion of correlation vs. causation," cried someone at Hacker News. "Correlation does not mean causation," a reader moaned at Slashdot. "There are so many variables here that it isn't funny."

And thus a deeper correlation was revealed, a link more telling than any that the Missouri team had shown. I mean the affinity between the online commenter and his favorite phrase—the statistical cliché that closes threads and ends debates, the freshman platitude turned final shutdown. "Repeat after me," a poster types into his window, and then he sighs, and then he types out his sigh, *s-i-g-h*, into the comment for good measure. Does he have to write it on the blackboard? *Correlation does not imply causation.* Your hype is busted. Your study debunked. End of conversation. Thank you and good night.

Graph these **case-sensitive** comma-separated phrases: correlation does not imply causation,correlation is not causation,cor

between 1880 and 2008 from the corpus English with smoothing of 5.

Search lots of books



Search in Google Books:

1880 - 1966	1967 - 1993	1994 - 1996	1997 - 2001	2002 - 2008	correlation does not prove causation (English)
1880 - 1970	1971 - 2000	2001 - 2003	2004 - 2006	2007 - 2008	correlation does not imply causation (English)
1880 - 1968	1969 - 2000	2001 - 2003	2004 - 2005	2006 - 2008	correlation is not causation (English)

is included in the regression equation that explains Y. "Statistical significance" is indicated by an asterisk, and three asterisks signal a high degree of significance. The idea is that a statistically significant coefficient differs from 0, so that X has a causal influence on Y. By contrast, an insignificant coefficient is zero: then X does not exert a causal influence on Y.

The reasoning is seldom made explicit, and difficulties are frequently overlooked. Stringent assumptions are needed to determine significance from the data. Even if significance can be determined and the null hypothesis rejected or accepted, there is a much deeper problem. To make causal inferences, it must in essence be assumed that equations are invariant under proposed interventions. Verifying such assumptions—without making the interventions—is quite problematic. On the other hand, if the coefficients and error terms change when the right hand side variables are manipulated rather than being passively observed, then the equation has only a limited utility for predicting the results of interventions. These difficulties are well known in principle, but are seldom dealt with by investigators doing applied work in the social and life sciences. Despite the problems, and the disclaimer in the footnote, Yule's regression approach has become widely used in the social sciences and epidemiology.

Some formal models for causation are available, starting with Neyman (1923). See Hodges and Lehmann (1964, sec. 9.4), Rubin (1974), or Holland (1988). More recent developments will be found in Pearl (1995, 2000) or Angrist, Imbens and Rubin (1996). For critical discussion from various perspectives, see Goldthorpe (1998, 2001), Humphreys and Freedman (1996, 1999), Abbott (1997), McKim and Turner (1997), Manski (1995), Lieberson (1985), Lucas (1976), Liu (1960), or Freedman (1987, 1991, 1995). Ní Bhrolcháin (2001) presents some fascinating case studies. The role of invariance is considered in Heckman (2000) and Freedman (2002). The history is reviewed by Stigler (1986) and Desrosières (1993).

5. REGRESSION MODELS IN EPIDEMIOLOGY

Regression models (and variations like the Cox model) are widely used in epidemiology. The models seem to give answers, and create at least the appearance of methodological rigor. This section discusses one example, which is fairly typical of such applications and provides an interesting contrast to Snow on cholera. Snow used primitive statistical techniques, but his study designs were extraordinarily well thought out, and he made a huge effort to collect the relevant data. By contrast, many empirical papers published today, even in the leading journals, lack a sharply-focused research question; or the study design connects the hypotheses to the data collection only in a very loose way. Investigators often try to use statistical models not only to control for confounding, but also to correct basic deficiencies in the design or the data. Our example will illustrate some of these points.

Kanarek et al. (1980) asked whether asbestos in the drinking water causes cancer. They studied 722 census tracts in the San Francisco Bay Area. (A census tract is a small geographical region, with several thousand inhabitants.) The investigators measured asbestos concentration in the water for each tract. Perhaps surprisingly, there is enormous variation; less surprisingly, higher concentrations are found in poorer tracts. Kanarek et al. compared the "observed" number of cancers by site with the expected number, by sex, race, and tract. The "expected" number is obtained by applying age-specific national rates to the population of the tract, age-group by age-group; males and females are done separately, and only whites are considered. (There are about 100 sites for which age-specific national data are available; comparison of observed to expected numbers is an

example of "indirect standardization.")

Regression is used to adjust for income, education, marital status, and occupational exposure. The equation is not specified in great detail, but is of the form

$$\log \frac{\text{Obs.}}{\text{Exp.}} = A_0 + A_1 \text{ asbestos concentration} + A_2 \text{ income} + A_3 \text{ education} + A_4 \text{ married} + A_5 \text{ asbestos workers} + \text{error}.$$

Here, "income" is the median figure for persons in the tract, and "education" is the median number of years of schooling; data are available from the census. These variables adjust to some extent for socio-economic differences between tracts: usually, rates of disease go down as income and education go up. The next variable in the equation is the fraction of persons in the tract who are married; such persons are typically less subject to disease than the unmarried. Finally, there is the number of "asbestos workers" in the tract; these persons may have unusually high rates of cancer, due to exposure on the job. Thus, the variables on the right hand side of the equation are potential confounders, and the equation tries to adjust for their effects. The estimate of A_1 for lung cancer in males is "highly statistically significant," with $P < .001$. A highly significant coefficient like this might be taken as strong evidence of causation, but there are serious difficulties.

Confounding. No adjustment is made for smoking habit, which was not measured in this study. Smoking is strongly but imperfectly associated with socio-economic status, and hence with asbestos concentration in the water. Furthermore, smoking has a substantial effect on cancer rates. Thus, smoking is a confounder. The equation does not correct for the effects of smoking, and the P-value does not take this confounding into account.

FIGURE 2. Smoking as an unmeasured confounder. The non-causal association between asbestos in the water and lung cancer is explained by the associations with smoking.

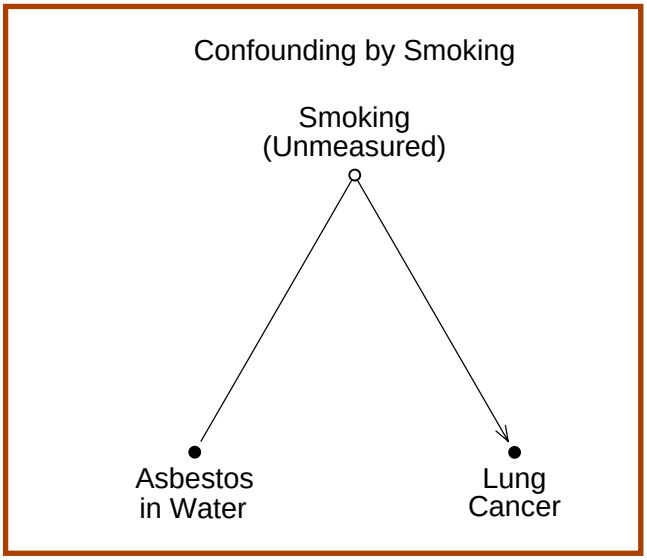


Figure 2 illustrates an alternative explanation for the data. (i) Smoking (an unmeasured confounder) is associated with the concentration of asbestos fibers in the water. The association is signaled by the straight line joining the two variables. (ii) Smoking has a strong, direct effect on

lung cancer, indicated by the arrow in the figure. Associations (i) and (ii) explain the association between asbestos fibers in the water and lung cancer rates; this observed association is not causal. To recapitulate, a confounder is associated with the putative cause and with its effect; the confounder may explain part or all of an observed association. In epidemiology, unmeasured or poorly measured confounders are the rule rather than the exception. (Technically, the relationships in Figure 2 must persist even after conditioning on the measured covariates.)

Model specification. The choice of variables and functional form is somewhat arbitrary, although not completely unreasonable. The authors say that their equation is suggested by mathematical models for cancer, but the connection is rather loose; nor have the cancer models themselves been validated (Freedman and Navidi, 1989, 1990).

Statistical assumptions. To compute the P-value, it is tacitly assumed that errors are statistically independent from tract to tract, and identically distributed. This assumption may be convenient, but it lacks an empirical basis.

The search for significance. Even if we set the fundamental difficulties aside, the authors have made several hundred tests on the equations they report, without counting any preliminary data analysis that may have been done. The P-values are not adjusted for the effects of the search, which may be substantial (Dijkstra, 1988; Freedman, 1983).

Weak effects. The effect being studied is weak: a 100-fold increase in asbestos fiber concentration is associated with perhaps a 5% increase in lung cancer rates. What is unusual about the present example is only the strength of the unmeasured confounder, and the weakness of the effect under investigation.

Epidemiology is best suited to the investigation of strong effects, which are hard to explain away by confounding (Cornfield et al., 1959, p. 199). As attention shifts to the weaker and less consistent effects that may be associated with low doses, difficulties will increase. Long delays between the beginning of exposure and the onset of disease are a further complication. Toxicology may be of some value but presents difficulties of its own (Freedman, Gold, and Lin, 1996; Freedman and Zeisel, 1988). The limitations of epidemiology are discussed by Taubes (1995). For detailed case studies, see Vandenbroucke and Pardoel (1989), Taubes (1998), or Freedman and Petitti (2001). Other examples will be given in section 7.

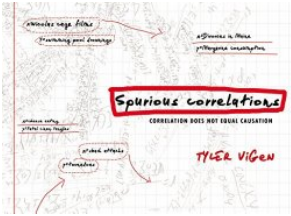
6. SOME GENERAL CONSIDERATIONS

Model specification. A model is specified by choosing (i) the explanatory variables to put on the right hand side, (ii) the functional form of the equation, and (iii) the assumptions about error terms. Explanatory variables are also called "covariates," or "independent variables"; the latter term does not connote statistical independence. The functional form may be linear, or log linear, or something even more exotic. Errors may be assumed independent or autoregressive; or some other low-order covariance matrix may be assumed, with a few parameters to estimate from the data.

Epidemiologists often have binary response variables: for instance, disease is coded as "1" and health as "0." A "logit" specification is common in such circumstances. Conditional on the covariates, subjects are assumed to be independent. If Y_i is the response for subject i while X_i is a $1 \times p$ vector of covariates, the logit specification is

$$\log \frac{\text{Prob}\{Y_i = 1\}}{\text{Prob}\{Y_i = 0\}} = X_i \beta.$$

Spurious correlations

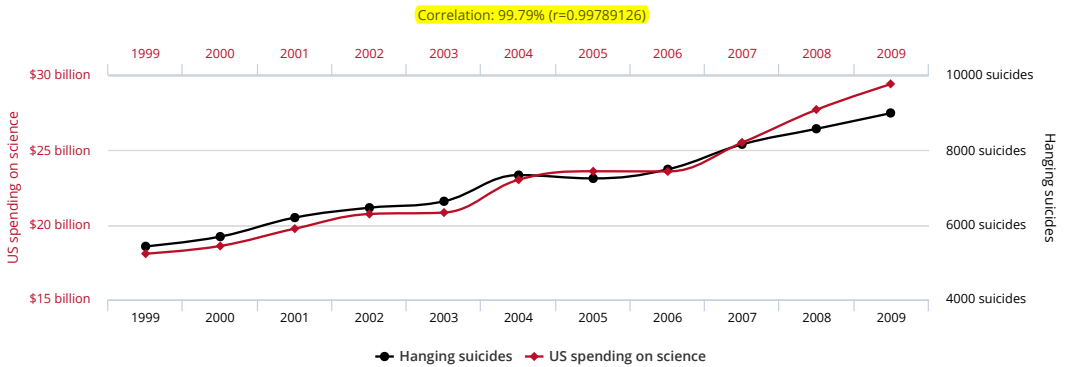


Now a ridiculous book!

- Spurious charts
- Fascinating factoids
- Commentary in the footnotes

Amazon | Barnes & Noble | Indie Bound

US spending on science, space, and technology
correlates with
Suicides by hanging, strangulation and suffocation



Data sources: U.S. Office of Management and Budget and Centers for Disease Control & Prevention

---

spurious correlations

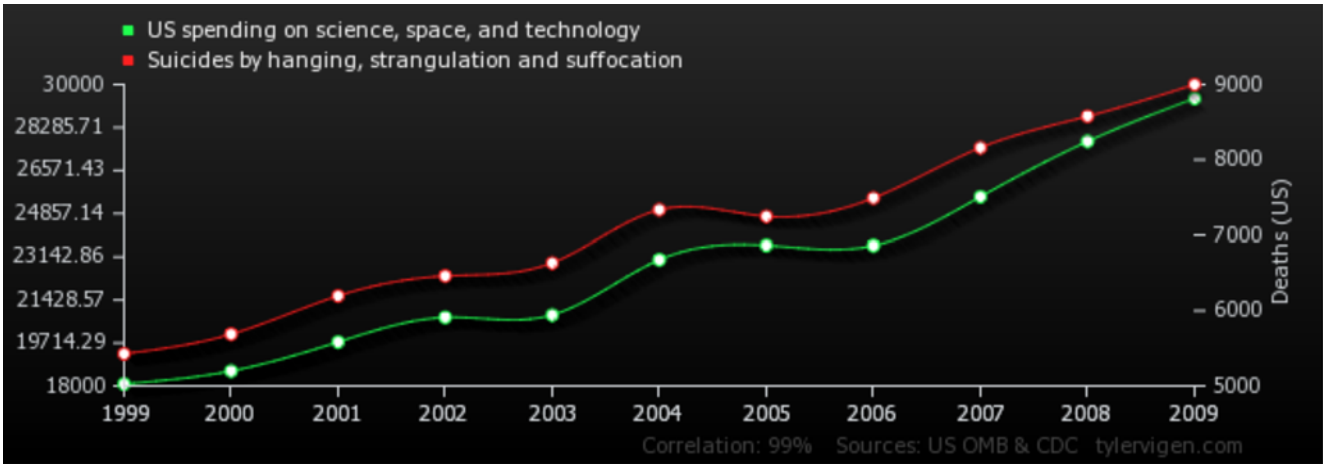
Discover a new correlation

Follow @TylerVigen Follow

US spending on science, space, and technology

correlates with

Suicides by hanging, strangulation and suffocation



	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009
US spending on science, space, and technology *Millions of todays dollars (US OMB)*	18,079	18,594	19,753	20,734	20,831	23,029	23,597	23,584	25,525	27,731	29,449
Suicides by hanging, strangulation and suffocation *Deaths (US) (CDC)*	5,427	5,688	6,198	6,462	6,635	7,336	7,248	7,491	8,161	8,578	9,000

Correlation: 0.992082

Permalink - Mark as interesting (5,147) - Not interesting (2,370)

Number people who drowned by falling into a swimming-pool

correlates with

Number of films Nicolas Cage appeared in

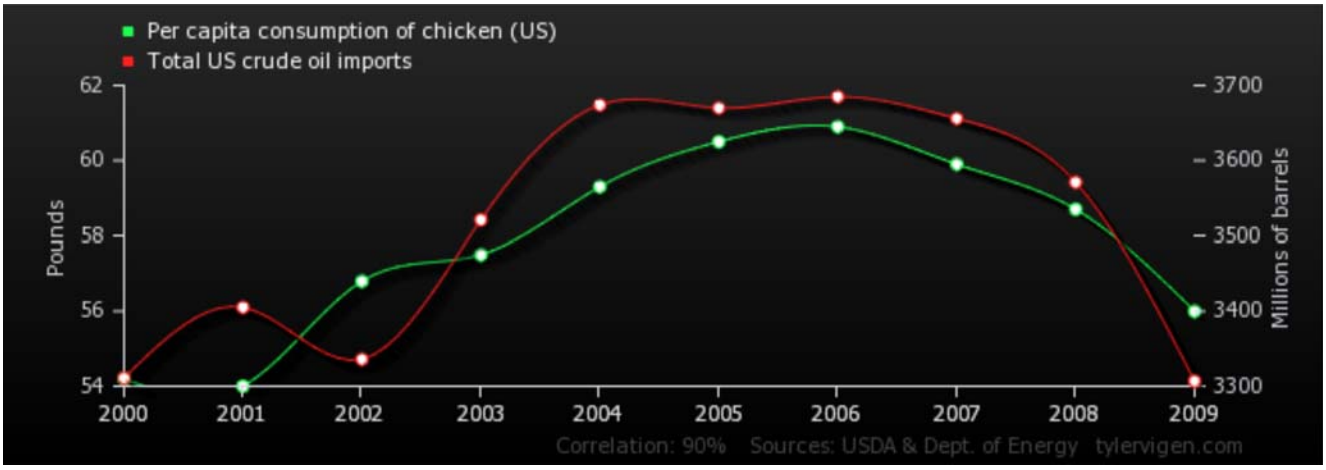
Correlation: 0.884883

Permalink - Mark as interesting (436) - Not interesting (329)

Per capita consumption of chicken (US)

correlates with

Total US crude oil imports



	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009
Per capita consumption of chicken (US) *Pounds (USDA)*	54.2	54	56.8	57.5	59.3	60.5	60.9	59.9	58.7	56
Total US crude oil imports *Millions of barrels (Dept. of Energy)*	3,311	3,405	3,336	3,521	3,674	3,670	3,685	3,656	3,571	3,307

Correlation: 0.899899

Permalink - Mark as interesting (430) - Not interesting (262)

Per capita consumption of sour cream (US)

correlates with

Motorcycle riders killed in noncollision transport accident

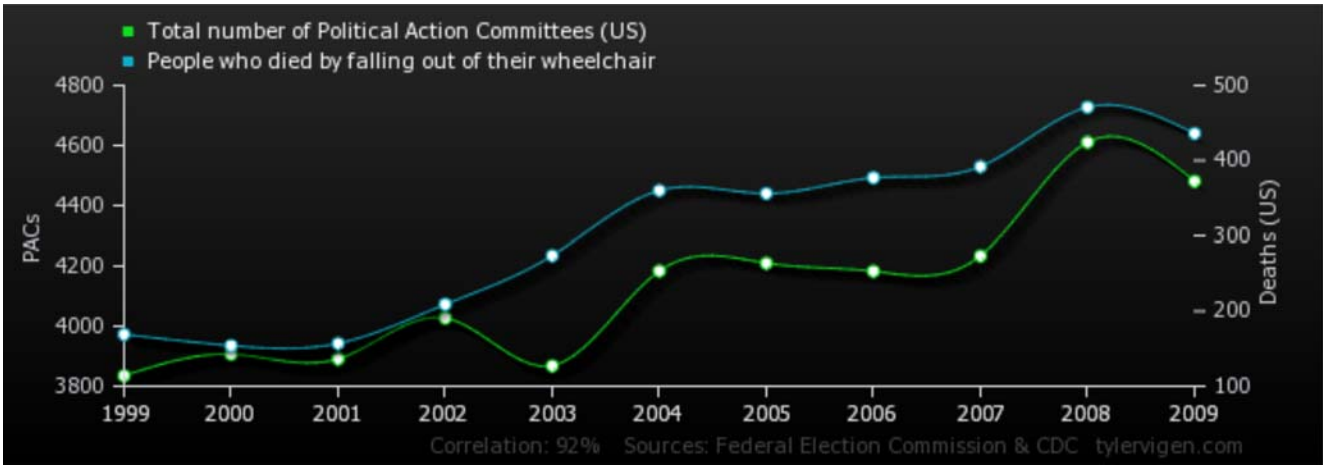
Correlation: 0.890472

Permalink - Mark as interesting (396) - Not interesting (204)

Total number of Political Action Committees (US)

correlates with

People who died by falling out of their wheelchair



	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009
Total number of Political Action Committees (US) *PACs (Federal Election Commission)*	3,835	3,907	3,891	4,027	3,868	4,184	4,210	4,183	4,234	4,611	4,481
People who died by falling out of their wheelchair *Deaths (US) (CDC)*	169	154	157	209	274	360	356	377	392	471	436

Correlation: 0.915876

Permalink - Mark as interesting (368) - Not interesting (221)

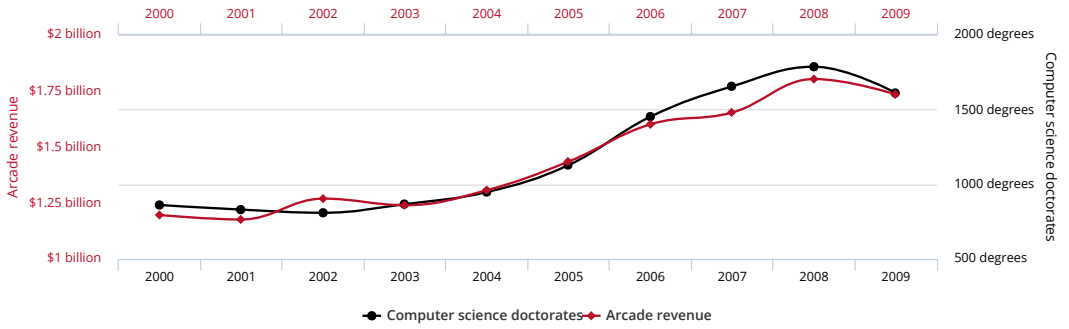
Number people who drowned while in a swimming-pool

correlates with

Power generated by nuclear power plants (US)

Total revenue generated by arcades
correlates with
Computer science doctorates awarded in the US

Correlation: 98.51% (r=0.985065)

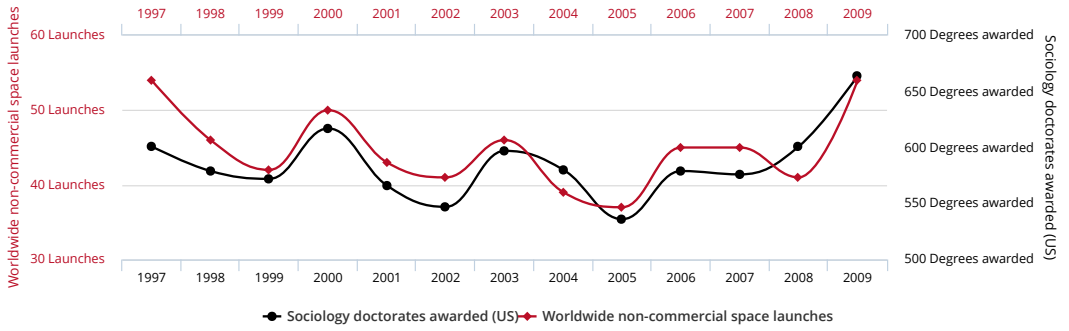


tylervigen.com

Data sources: U.S. Census Bureau and National Science Foundation

Worldwide non-commercial space launches
correlates with
Sociology doctorates awarded (US)

Correlation: 78.92% (r=0.78915)



tylervigen.com

Data sources: Federal Aviation Administration and National Science Foundation

Example 8.5 shows that observational studies can provide information about causality, but must be interpreted cautiously. Researchers generally agree that a causal interpretation of an observed association requires extra support—for instance, that the association be observed consistently in observational studies conducted under various conditions and taking various extraneous factors into account, and also, ideally, that the causal link be supported by experimental evidence. We do not mean to say that an observed association *cannot* be causally interpreted, but only that such interpretation requires particular caution.

Spurious Association

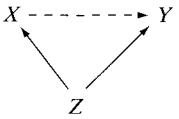
le 8.6

Ultrasound. It is quite common for a physician to use ultrasound examination of the fetus of a pregnant woman. However, when ultrasound technology was first used there were concerns that the procedure might be harmful to the baby. An early study seemed to bear this out: on average, babies exposed to ultrasound in

the womb were lighter at birth than were babies not exposed to ultrasound.[8] Later, a study was done in which some women were randomly chosen to have ultrasounds and other were not given ultrasounds. This study found no difference in birthweight between the two groups.[9] It seems that the reason a difference appeared in the first study was that ultrasound was being used mostly for women who were experiencing problem pregnancies. The complications with the pregnancy were leading to low birthweight, not the use of ultrasound. ■

Tom Cruise has one

Figure 8.2 gives a schematic representation of the situation in Example 8.6. Changes in X (having an ultrasound examination) are associated with changes in Y (lower birthweight). However, X and Y are both dependent on a third variable Z (whether or not there are problems with the pregnancy), which is the variable that is driving the relationship. Changes in X and changes in Y are a common response to the third variable Z. We say that the association between X and Y is **spurious**: When we control for the "lurking variable" Z, the link between X and Y disappears. In the case of Example 8.6, it is not having an ultrasound that influences birthweight; what matters is whether or not there were problems with the pregnancy.



The association between X and Y is spurious; controlling for the lurking variable Z eliminates the X-Y link.

Figure 8.2 Schematic representation of spurious association

Freedman paper; asbestos, lung cancer (smoking)

association, confounding, spurious correlation

Confounding

Many observational studies are aimed at discovering some kind of causal relationship. Such discovery can be very difficult because of extraneous variables that enter in an uncontrolled (and perhaps unknown) way. The investigator must be guided by the maxim:

Association is not causation.

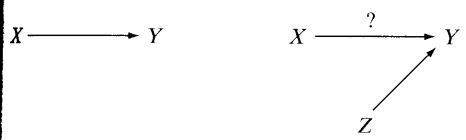
For instance, it is known that some populations whose diets are high in fiber enjoy a reduced incidence of colon cancer. But this observation does not in itself show that it is the high-fiber diet, rather than some other factor, that provides the protection against colon cancer.

The following example shows how uncontrolled extraneous variables can cloud an observational study, and what kinds of steps can be taken to clarify the picture.

Example 8.4

Smoking and Birthweight. In a large observational study of pregnant women, it was found that the women who smoked cigarettes tended to have smaller babies than the nonsmokers.[2] (This study was mentioned in Example 8.2.) It is plausible that smoking could cause a reduction in birthweight, for instance by interfering with the flow of oxygen and nutrients across the placenta. But of course plausibility is not proof. In fact, the investigators found that the smokers differed from the nonsmokers with respect to many other variables. For instance, the smokers drank more whiskey than the nonsmokers. Alcohol consumption might plausibly be linked to a deficit in growth. ■

In Example 8.4 three variables are presented; let us refer to these as X = smoking, Y = birthweight, and Z = alcohol consumption. There is an association between X and Y, but is there a *causal* link between them? Or is there a causal link between Z and Y? Figure 8.1 gives a schematic representation of the situation. Changes in X are associated with changes in Y. However, changes in Z are also associated with changes in Y. We say that the effect that X has on Y is **confounded** with the effect that Z has on Y. In the context of Example 8.4, we say that the effect that smoking has on birthweight is confounded with the effect that alcohol consumption has on birthweight. In observational studies, confounding of effects is a common problem.



(a) (b) The effect of X on Y is confounded with the effect of Z on Y.

Figure 8.1 Schematic representation of causation (a) and of confounding (b)

Example 8.5

Smoking and Birthweight. The study presented in Example 8.4 uncovered many confounding variables. For example, the smokers drank more coffee than the nonsmokers. In addition—and this is especially puzzling—it was found that the smokers began to menstruate at younger ages than the nonsmokers. This phenomenon (early onset of menstruation) could not possibly have been *caused* by smoking, because it occurred (in almost all instances) *before* the woman began to

THIRD-VARIABLES WEEK 2 STAT 209 [redacted]

Partial, part correlations (spurious associations)

Consider X_1, X_2, X_3 (may be measured w/ error)

$$r_{13\cdot 2} = \frac{r_{13} - r_{12}\, r_{23}}{\sqrt{(1-r_{12}^2)(1-r_{23}^2)}}$$

adjusted vars

$$r_{13\cdot 2} = r_{(1\cdot 2)(3\cdot 2)}$$

also $r_{12\cdot 345} = r_{(1\cdot 345)(2\cdot 345)}$ etc

From Stat 60

$H_0: \rho = 0$

t-statistic $\dfrac{r\sqrt{n-2}}{\sqrt{1-r^2}}$

t_{n-2} critical value

for $H_0: \rho_{12\cdot 3} = 0$

$\rho_{12\cdot 3} = \rho_{(1\cdot 3)(2\cdot 3)}$ test statistic $r_{12\cdot 3}\sqrt{n-3}/\sqrt{1-r}$

part correlations $r_{(1\cdot 2)3}$ $r_{1(3\cdot 2)}$

$$R^2_{y\cdot x_1 x_2} = r^2_{yx_1} + r^2_{y(x_2\cdot x_1)}$$

Dichotomous Data: Spurious Correlation, Confoundi

Partial, Part Correl.

<Neter 9.4 ~~pp 286-7~~> p. 276

partial correlation

X_1 X_2 X_3

[Radin X_1 nurturance (home nv)
X_2 ach motivation
X_3 achievement (S-Binet)]

Correlation (X_2 moderator var?)

$$r_{13.2} = \frac{r_{13} - r_{12} r_{23}}{\sqrt{(1 - r_{12}^2)(1 - r_{23}^2)}}$$

<9.417
<~~8.387~~

relation to standardized regression coefficient

$$= r_{(1.2)(3.2)}$$

adjusted variables (marks. log)

Part (semi-partial)

$r_{(1\cdot 2)3}$ $r_{1(3\cdot 2)}$

Relations to R^2

(relation to extra SS Nwk sec 7.3)

Y, X_1, X_2

$$(1 - R^2_{Y\cdot X_1 X_2}) = (1 - r^2_{YX_1})(1 - r^2_{YX_2\cdot X_1})$$

$$R^2_{Y\cdot X_1 X_2} = r^2_{YX_1} + r^2_{Y(X_2\cdot X_1)}$$

$$= r^2_{YX_2} + r^2_{Y(X_1\cdot X_2)}$$

Inferences for partial, part

Fishers z transf. $(n-4)$

cor.test usues n-3

Package 'ppcor'

December 3, 2015

Type Package

Title Partial and Semi-Partial (Part) Correlation

Version 1.1

Date 2015-11-19

Author Seongho Kim

Maintainer Seongho Kim <biostatistician.kim@gmail.com>

Depends R (>= 2.6.0), MASS

Description Calculates partial and semi-partial
(part) correlations along with p-value.

License GPL-2

NeedsCompilation no

Repository CRAN

Date/Publication 2015-12-03 13:05:14

R topics documented:

ppcor-package 1
pcor 3
pcor.test 4
spcor 6
spcor.test 7

Index **9**

`ppcor-package` *Partial and Semi-partial (Part) Correlation*

Description

Calculates parital and semi-partial (part) correlations along with p value.

Details

THIRD-VARIABLES WEEK 2 STAT 209 [redacted]

Partial, part correlations (spurious associations)

Consider X_1, X_2, X_3 (may be measured w/ error)

$$r_{13\cdot 2} = \frac{r_{13} - r_{12}\, r_{23}}{\sqrt{(1-r_{12}^2)(1-r_{23}^2)}}$$

adjusted vars

$r_{13\cdot 2} = r_{(1\cdot 2)(3\cdot 2)}$

also $r_{12\cdot 345} = r_{(1\cdot 345)(2\cdot 345)}$ etc

From Stat 60

$H_0: \rho = 0$

t-statistic $\frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$

t_{n-2} critical value

for $H_0: \rho_{12\cdot 3} = 0$

$\rho_{12\cdot 3} = \rho_{(1\cdot 3)(2\cdot 3)}$ test statistic $r_{12\cdot 3}\sqrt{n-3}/\sqrt{1-r^2}$

part correlations $r_{(1\cdot 2)3}$ $r_{1(3\cdot 2)}$

$$R^2_{y\cdot x_1 x_2} = r^2_{yx_1} + r^2_{y(x_2\cdot x_1)}$$

Dichotomous Data: Spurious Correlation, Confoundi

Simpson's paradox: conditional, marginal tables

Death penalty ex. odds ratios flip w/ 3rd variable

DP, DR, VR U.C. Berkeley grad admissions

```
SIMPSON'S PARADOX (marginal vs conditional odds ratios) DEATH PENALTY ex
> deathP = matrix(c(19,17, 141,149), nr = 2,
+ dimnames = list("Def" = c("Wh", "Blk"), "DP" = c("Y","N")))
> deathP   # unconditional, marginal table
     DP
Def    Y   N
  Wh  19 141
  Blk 17 149
> prop.table(deathP,1)
     DP
Def           Y         N
  Wh  0.1187500 0.8812500
  Blk 0.1024096 0.8975904
> # so where's the racial bias? Wh seems more likely to fry
> deathPWvic = matrix(c(19,11, 132,52), nr = 2,
+ dimnames = list("Def" = c("Wh", "Blk"), "DP" = c("Y","N")))
> deathPBvic = matrix(c(0,6, 9,97), nr = 2,
+ dimnames = list("Def" = c("Wh", "Blk"), "DP" = c("Y","N")))
> prop.table(deathPWvic,1)
     DP
Def           Y         N
  Wh  0.1258278 0.8741722
  Blk 0.1746032 0.8253968
> prop.table(deathPBvic,1)
     DP
Def           Y         N
  Wh  0.00000000 1.0000000
  Blk 0.05825243 0.9417476
> # for each level  of Victim race, Black Def more likely to receive DP
reversal by conditioning instance of Simpson's Paradox (e.g. marginal vs cond'l or)
```

Stat 141 ex

DP × Def association, dichotomous vars

2 cond'l tables

condition on race of victim

Dichotomous Outcomes

Dichotomous outcome (0,1)

W: $W=1$ $Y \geq c$

$W=0$ $Y < c$

see HW1 logistic regr.

measured outcome Y, "t-test" on group membership G (0,1)

$E(Y|G) = \beta_0 + \beta_1 G$

$\beta_1 = \mu_1 - \mu_0$

is outcome associated with Group membership?

(either rcft or observational self-selection)

note: point-biserial correlation (sample)

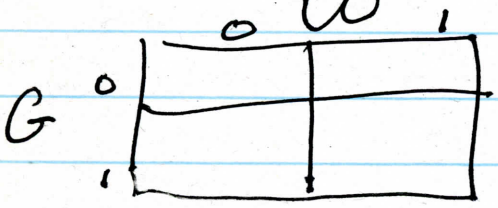
$r_{YG} = \frac{\bar{Y}_1 - \bar{Y}_0}{s_y} \sqrt{\frac{n_1 n_0}{n(n-1)}}$

$n = n_1 + n_0$

Group Effects w/ Dichotomous Outcome

2 x 2 tables

(independence no group effect hypoth)



X^2 deviation from independence

phi coeff $\sqrt{X^2/n}$

product moment correlation

other measures of association W, G

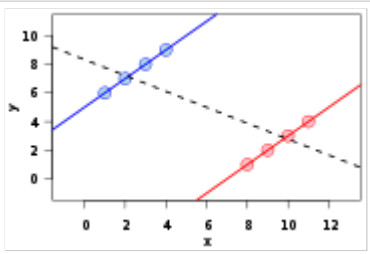
relative risk, odds ratio

Same cautions, difficulties in experiments vs observational studies (self selection)

Jamie Robbins

Simpson's paradox

From Wikipedia, the free encyclopedia



Simpson's paradox for continuous data: a positive trend appears for two separate groups (blue and red), a negative trend (black, dashed) appears when the data are combined.

In probability and statistics, **Simpson's paradox** (or the **Yule-Simpson effect**) is an apparent paradox in which the successes of groups seem reversed when the groups are combined. This result is often encountered in social and medical science statistics,[1] and occurs when frequency data are hastily given causal interpretation;[2] the paradox disappears when causal relations are derived systematically, through formal analysis.

Though mostly unknown to laymen, Simpson's Paradox is well known to statisticians, and is described in several introductory statistics books.[3][4] Many statisticians believe that the mainstream public should be apprised of counterintuitive results such as Simpson's paradox,[5] in particular to caution against the inference of causal relationships based on the association between two variables.[6]

Edward H. Simpson described the phenomenon in 1951,[7] along with Karl Pearson et al.,[8] and Udny Yule in 1903.[9] The name *Simpson's paradox* was coined by Colin R. Blyth in 1972.[10] Since Simpson did not discover this statistical paradox, some authors, instead, have used the impersonal names *reversal paradox* and *amalgamation paradox* in referring to what is now called *Simpson's Paradox* and the *Yule-Simpson effect*.[11]

Contents

- 1 Basic principle
- 2 Examples
  - 2.1 Kidney stone treatment
  - 2.2 Berkeley sex bias case
  - 2.3 2006 US school study
  - 2.4 Health care disparities
  - 2.5 Low birth weight paradox
  - 2.6 Batting averages
- 3 Description
  - 3.1 Vector interpretation
- 4 How likely is Simpson's paradox?
- 5 References
- 6 External links

Basic principle

Suppose 5 men and 5 women apply to two different departments in a school.

	Men	Women
Arts	3 out of 4 accepted (**75%**)	1 out of 1 accepted (**100%**)
Science	0 out of 1 accepted (**0%**)	1 out of 4 accepted (**25%**)
Totals	3 out of 5 accepted (**60%**)	2 out of 5 accepted (**40%**)

Although each department separately has a higher acceptance rate for women, the combined acceptance rate for men is much higher.

Examples

Kidney stone treatment

(Accepted 25 August 1994)

Confounding and Simpson's paradox

Steven A Julious, Mark A Mullee

Medical Statistics and Computing, University of Southampton, Southampton General Hospital, Southampton SO16 6YD
Steven A Julious, *statistician*
Mark A Mullee, *senior statistical programmer*

Correspondence to:
Mr Julious.

BMJ 1994;309:1480-1

A common problem when analysing clinical data is that of confounding. This occurs when the association between an exposure and an outcome is investigated but the exposure and outcome are strongly associated with a third variable. An extreme example of this is Simpson's paradox, in which this third factor reverses the effect first observed.[1] This phenomenon has long been recognised as a theoretical possibility but few real examples have been presented.

Examples

Charig *et al* undertook a historical comparison of success rates in removing kidney stones.[2] Open surgery (1972-80) had a success rate of 78% (273/350) while percutaneous nephrolithotomy (1980-5) had a success rate of 83% (289/350), an improvement over the use of open surgery. However, the success rates looked rather different when stone diameter was taken into account. This showed that, for stones of <2 cm, 93% (81/87) of cases of open surgery were successful compared with just 83% (234/270) of cases of percutaneous nephrolithotomy. Likewise, for stones of ≥ 2 cm, success rates of 73% (192/263) and 69% (55/80) were observed for open surgery and percutaneous nephrolithotomy respectively.

The main reason why the success rate reversed is because the probability of having open surgery or percutaneous nephrolithotomy varied according to the diameter of the stones. In observational (non-randomised) studies comparing treatments it is likely that the initial choice of treatment would have been influenced by patients' characteristics such as age or severity of condition; so any difference between treatments could be accounted for by these original factors. Such a situation may arise when a new treatment is being phased in over time. Randomised trials are therefore necessary to demonstrate any treatment effect.

In another example Hand reported that the proportion of male patients in a psychiatric hospital seemed to fall slightly over time, from 46·4% (343/739) in 1970 to 46·2% (238/515) in 1975.[3] When the results were broken down according to patients' age, however, it was observed that the proportion of male patients had increased; from 59·4% (255/429) to 60·5% (156/258) among those aged <65 and from 28·4% (88/310) to 31·9% (82/257) among those aged ≥ 65.

The table shows another example, a study of mortality and diabetes (data from the Poole diabetic cohort[4]). In the study only 29% of the patients with

How likely is Simpson's paradox?

If a $2 \times 2 \times 2$ table - such as in the kidney stone example - is selected at random (given certain conditions), the probability is approximately 1/60 that Simpson's paradox will occur purely by chance.[23]

References

1. ^ Clifford H. Wagner (February 1982). "Simpson's Paradox in Real Life". *The American Statistician* **36** (1): 46–48. doi:10.2307/2684093 (http://dx.doi.org/10.2307%2F2684093) .
2. ^ ***a b c d*** Judea Pearl. *Causality: Models, Reasoning, and Inference*, Cambridge University Press, 2000. ISBN 0-521-77362-8.
3. ^ ***a b*** David Freedman, Robert Pisani and Roger Purves. Statistics (3rd edition). W.W. Norton, 1998, p. 19. ISBN 0-393-97083-3.
4. ^ David S. Moore and D.S. George P. McCabe (February 2005). "*Introduction to the Practice of Statistics*" (5th edition). W.H. Freeman & Company. ISBN 071676282X.
5. ^ Robert L. Wardrop (February 1995). "Simpson's Paradox and the Hot Hand in Basketball". *The American Statistician*, **49 (1)**: pp. 24–28.
6. ^ Alan Agresti (2002). "Categorical Data Analysis" (Second edition). John Wiley and Sons. ISBN 0-471-36093-7
7. ^ Simpson, Edward H. (1951). "The Interpretation of Interaction in Contingency Tables". *Journal of the Royal Statistical Society, Ser. B* **13**: 238–241.
8. ^ Pearson, Karl; Lee, A.; Bramley-Moore, L. (1899). "Genetic (reproductive) selection: Inheritance of fertility in man". *Philosophical Translations of the Royal Statistical Society, Ser. A* **173**: 534–539.
9. ^ G. U. Yule (1903). "Notes on the Theory of Association of Attributes in Statistics". *Biometrika* **2**: 121–134. doi:10.1093/biomet/2.2.121 (http://dx.doi.org/10.1093%2Fbiomet%2F2.2.121) .
10. ^ Colin R. Blyth (June 1972). "On Simpson's Paradox and the Sure-Thing Principle". *Journal of the American Statistical Association* **67** (338): 364–366. doi:10.2307/2284382 (http://dx.doi.org/10.2307%2F2284382) .
11. ^ I. J. Good, Y. Mittal (June 1987). "The Amalgamation and Geometry of Two-by-Two Contingency Tables (http://links.jstor.org/sici?sici=0090-5364%28198706%2915%3A2%3C694%3ATAAGOT%3E2.0.CO%3B2-0) ". *The Annals of Statistics* **15** (2): 694–711. doi:10.1214/aos/1176350369 (http://dx.doi.org/10.1214%2Faos%2F1176350369) .
12. ^ C. R. Charig, D. R. Webb, S. R. Payne, O. E. Wickham (29 March 1986). "Comparison of treatment of renal calculi by operative surgery, percutaneous nephrolithotomy, and extracorporeal shock wave lithotripsy" (http://www.pubmedcentral.nih.gov/articlerender.fcgi?tool=pubmed&pubmedid=3083922) . *Br Med J (Clin Res Ed)* **292** (6524): 879–882. doi:10.1136/bmj.292.6524.879 (http://dx.doi.org/10.1136%2Fbmj.292.6524.879) . PMID 3083922 (http://www.ncbi.nlm.nih.gov/pubmed/3083922) . http://www.pubmedcentral.nih.gov/articlerender.fcgi?tool=pubmed&pubmedid=3083922.
13. ^ Steven A. Julious and Mark A. Mullee (December 3, 1994). "Confounding and Simpson's paradox" (http://bmj.bmjjournals.com/cgi/content/full/309/6967/1480) . *BMJ* **309** (6967): 1480–1481. PMID 7804052 (http://www.ncbi.nlm.nih.gov/pubmed/7804052) . http://bmj.bmjjournals.com/cgi/content/full/309/6967/1480.
14. ^ ***a b c*** P.J. Bickel, E.A. Hammel and J.W. O'Connell (1975). "Sex Bias in Graduate Admissions: Data From Berkeley (http://www.sciencemag.org/cgi/content/abstract/187/4175/398) ". *Science* **187** (4175): 398–404. doi:10.1126/science.187.4175.398 (http://dx.doi.org/10.1126%2Fscience.187.4175.398) . PMID 17835295 (http://www.ncbi.nlm.nih.gov/pubmed/17835295) ..
15. ^ H. Braun, F. Jenkins and W. Grigg, (2006) "Comparing Private Schools and Public Schools Using Hierarchical Linear Modeling (http://www.nytimes.com/packages/pdf/national/20060715report.pdf) , U.S. Department of Education, National Center for Education Statistics, Institute of Education Sciences, Washington, DC, United States Government Printing Office.
16. ^ Diana Jean Schemo. "Public Schools Perform Near Private Ones in Study (http://www.nytimes.com/2006/07/15/education/15report.html) . The New York Times, 15 July 2006. Retrieved on 25 July 2007.
17. ^ Asch DA, Armstrong KA. Aggregating and partitioning populations in health care disparities research: differences in perspective. J Clin Oncol. 2007;25:2117-21. (http://jco.ascopubs.org/cgi/reprint/25/15/2117)
18. ^ Wilcox, Allen (2006). "The Perils of Birth Weight — A Lesson from Directed Acyclic Graphs (http://aje.oxfordjournals.org/cgi/content/abstract/164/11/1121) ". *American Journal of Epidemiology*. 164(11):1121–1123.
19. ^ Ken Ross. "*A Mathematician at the Ballpark: Odds and Probabilities for Baseball Fans (Paperback)*" Pi Press, 2004. ISBN 0131479903. 12–13
20. ^ Statistics available from http://www.baseball-reference.com/ : Data for Derek Jeter (http://www.baseball-reference.com/j/jeterde01.shtml) , Data for David Justice (http://www.baseball-reference.com/j/justida01.shtml) .
21. ^ John Fox (1997). "Applied Regression Analysis, Linear Models, and Related Methods". Sage Publications. ISBN 080394540X. 136–137
22. ^ Jerzy Kocik (December 2001). Proofs without Words: Simpson's Paradox (http://www.jstor.org/pss/2691038) . *Mathematics Magazine*. 74 (5), p. 399.
23. ^ Marios G. Pavlides and Michael D. Perlman (August 2009). "How Likely is Simpson's Paradox?". *The American Statistician* **63** (3): 226–233. doi:10.1198/tast.2009.09007 (http://dx.doi.org/10.1198%2Ftast.2009.09007) .

External links

- Stanford Encyclopedia of Philosophy: "Simpson's Paradox (http://plato.stanford.edu/entries/paradox-simpson/) " -- by Gary Malinas.
- Earliest known uses of some of the words of mathematics: S (http://jeff560.tripod.com/s.html)

Simpson's Paradox

First published Mon Feb 2, 2004; substantive revision Thu Aug 6, 2009

An association between a pair of variables can consistently be inverted in each subpopulation of a population when the population is partitioned. For example, a medical treatment can be associated with a *higher* recovery rate for treated patients compared with the recovery rate for untreated patients; yet, treated male patients and treated female patients can each have *lower* recovery rates when compared with untreated male patients and untreated female patients. Conversely, higher recovery rates for treated patients in each subpopulation are consistent with a lower recovery rate in the total population when data are aggregated. The arithmetical structures that underlie facts like these invalidate a cluster of arguments that many people, at least initially, take to be intuitively valid. E.g., despite intuitions to the contrary, the following argument is invalid.

> The probability of male patients recovering following treatment is greater than the probability of their recovering following no treatment.
>
> The probability of female patients recovering following treatment is greater than the probability of their recovering following no treatment.
>
> Therefore, the probability of (male and female) patients recovering following treatment is greater than the probability of their recovering following no treatment.

Further, the arithmetical structures that invalidate such arguments pose deep problems for inferences from statistical regularities to conclusions about causal relations. Robust associations between variables can mask underlying causal structures that, when made explicit, expose the associations to be causally spurious. In the example above, higher recovery rates in each subpopulation are not sufficient to establish that a proposed treatment is causally effective in promoting recovery. Provided that the sample space is large enough to support causal inferences, different partitions of the population will exhibit different regularities that can appear to support incompatible conclusions about whether a treatment is causally effective. However, once the arithmetical structures that underlie arguments like the one above are made explicit, the structures provide a rich resource for providing causal models for actual and possible causal systems that are initially puzzling and can appear to be impossible. These include causal models for the evolution of traits such as altruism in a setting in which natural selection disadvantages

But would you want to matriculate?

We consider data on admissions for Fall 1973 graduate study at U.C. Berkeley in the six largest departments. These data among others were the subject of extensive litigation on gender discrimination a few years back.

The data on each applicant consists of the applicants gender (G), whether admitted (A) and major department (D).

	Whether admitted, male		Whether admitted, female	
Dept	Yes	No	Yes	No
a	512	313	89	19
b	353	207	17	8
c	120	205	202	391
d	138	279	131	244
e	53	138	94	299
f	22	351	24	317

Dichotomous Data: Spurious Correlation, Confoundi

Simpson's paradox: conditional, marginal tables
odds ratios flip w/ 3rd variable

Death penalty ex.
DP, DR, VR U.C. Berkeley grad admissions

Stat 141 ex

```
SIMPSON'S PARADOX (marginal vs conditional odds ratios) DEATH PENALTY ex
> deathP = matrix(c(19,17, 141,149), nr = 2,
+ dimnames = list("Def" = c("Wh", "Blk"), "DP" = c("Y","N")))
> deathP   # unconditional, marginal table
     DP
Def    Y   N
  Wh  19 141
  Blk 17 149
> prop.table(deathP,1)
     DP
Def           Y         N
  Wh  0.1187500 0.8812500
  Blk 0.1024096 0.8975904
> # so where's the racial bias? Wh seems more likely to fry
> deathPWvic = matrix(c(19,11, 132,52), nr = 2,
+ dimnames = list("Def" = c("Wh", "Blk"), "DP" = c("Y","N")))
> deathPBvic = matrix(c(0,6, 9,97), nr = 2,
+ dimnames = list("Def" = c("Wh", "Blk"), "DP" = c("Y","N")))
> prop.table(deathPWvic,1)
     DP
Def           Y         N
  Wh  0.1258278 0.8741722
  Blk 0.1746032 0.8253968
> prop.table(deathPBvic,1)
     DP
Def           Y         N
  Wh  0.00000000 1.0000000
  Blk 0.05825243 0.9417476
> # for each level  of Victim race, Black Def more likely to receive DP
reversal by conditioning instance of Simpson's Paradox (e.g. marginal vs cond'l or)
```

DP × Def association, dichotomous vars

2 cond'l tables

condition on race of victim

SIMPSON'S PARADOX (marginal vs conditional odds ratios) DEATH PENALTY ex

```
> deathP = matrix(c(19,17, 141,149), nr = 2,
+ dimnames = list("Def" = c("Wh", "Blk"), "DP" = c("Y","N")))
> deathP   # unconditional, marginal table
     DP
Def    Y   N
  Wh  19 141
  Blk 17 149
> prop.table(deathP,1)
     DP
Def           Y         N
  Wh  0.1187500 0.8812500
  Blk 0.1024096 0.8975904
> # so where's the racial bias? Wh seems more likely to fry
> deathPWvic = matrix(c(19,11, 132,52), nr = 2,
+ dimnames = list("Def" = c("Wh", "Blk"), "DP" = c("Y","N")))
> deathPBvic = matrix(c(0,6, 9,97), nr = 2,
+ dimnames = list("Def" = c("Wh", "Blk"), "DP" = c("Y","N")))
> prop.table(deathPWvic,1)
     DP
Def           Y         N
  Wh  0.1258278 0.8741722
  Blk 0.1746032 0.8253968
> prop.table(deathPBvic,1)
     DP
Def            Y         N
  Wh  0.00000000 1.0000000
  Blk 0.05825243 0.9417476
> # for each level  of Victim race, Black Def more likely to receive DP
reversal by conditioning instance of Simpson's Paradox (e.g. marginal vs cond'l or)
```

Stat141-- victim race stratifying var, flips association

Table 5.1 Death Penalty Verdict by Defendant's Race and Victim's Race

Defendant's Race	Victim's Race	Death Penalty		Percentage Yes
		Yes	No	
White	White	19	132	12.6
	Black	0	9	0.0
Black	White	11	52	17.5
	Black	6	97	5.8

Source: Reprinted with permission from Radelet (1981).

the levels of victim's race. About 12% of white defendants and about 10% of black defendants received the death penalty. *Ignoring* victim's race, the percentage of "yes" death penalty verdicts was lower for blacks than for whites.

For each combination of defendant's race and victim's race, Table 5.1 lists the percentage of subjects who received the death penalty. These are displayed in Figure 5.1. Consider the association between defendant's race and the death penalty verdict, controlling for victim's race. When the victim was white, the death penalty was imposed about 5 percentage points more often for black defendants than for white defendants. When the victim was black, the death penalty was imposed over 5 percentage points more often for black defendants than for white defendants. *Controlling* for victim's race, the percentage of "yes" death penalty verdicts was higher for blacks than for whites. The direction of the association is the reverse of that in the marginal table.

Why does the association between death penalty verdict and defendant's race change direction when we control victim's race? Let us study Table 5.3. For each pair of variables, it lists the marginal odds ratio and also the partial odds ratio at each level of the third variable. The

Table 5.2 Frequencies for Death Penalty Verdict and Defendant's Race

Defendant's Race	Death Penalty		Total
	Yes	No	
White	19	141	160
Black	17	149	166
Total	36	290	326

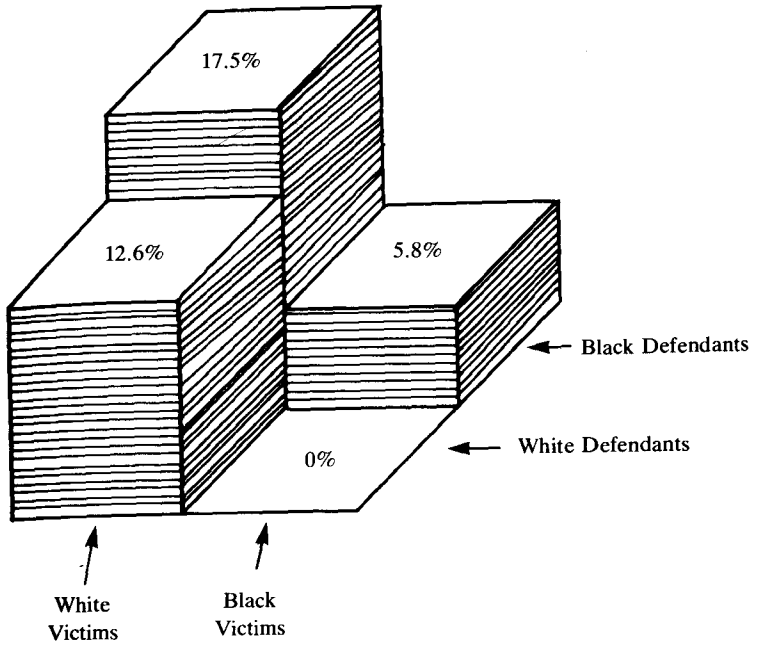


Figure 5.1 Percent receiving death penalty.

arginal odds ratios describe the association when the third variable is nored (i.e., when we sum the counts over the levels of the third variable obtain a marginal two-way table). The partial odds ratios describe the sociation when the third variable is controlled. Since one cell count in e three-dimensional table equals zero and since several of them are all, we added 0.5 to each cell count before computing these odds tios.

able 5.3 Odds Ratios for Death Penalty (*P*), Victim's Race (*V*), and efendant's Race (*D*)[a]

Association		Variables		
		P–D	*P–V*	*D–V*
Marginal		1.18	2.71	25.99
Partial	Level 1	0.67	2.80	22.04
	Level 2	0.79	3.29	25.90

[a] The value 0.5 was added to each cell frequency before calculation of odds ratios.

frontiers in PSYCHOLOGY

HYPOTHESIS AND THEORY ARTICLE
published: 12 August 2013
doi: 10.3389/fpsyg.2013.00513

Simpson's paradox in psychological science: a practical guide

Rogier A. Kievit[1,2]*, Willem E. Frankenhuis[3], Lourens J. Waldorp[1] and Denny Borsboom[1]

[1] Department of Psychological Methods, University of Amsterdam, Amsterdam, Netherlands
[2] Medical Research Council – Cognition and Brain Sciences Unit, Cambridge, UK
[3] Department of Developmental Psychology, Radboud University Nijmegen, Nijmegen, Netherlands

Edited by:
Joshua A. McGrane, The University of Western Australia, Australia

Reviewed by:
Mike W. L. Cheung, National University of Singapore, Singapore
Rink Hoekstra, University of Groningen, Netherlands

***Correspondence:**
Rogier A. Kievit, Medical Research Council - Cognition and Brain Sciences Unit, 15 Chaucer Rd, Cambridge, CB2 7EF, Cambridgeshire, UK
e-mail: rogier.kievit@mrc-cbu.cam.ac.uk

The direction of an association at the population-level may be reversed within the subgroups comprising that population—a striking observation called Simpson's paradox. When facing this pattern, psychologists often view it as anomalous. Here, we argue that Simpson's paradox is more common than conventionally thought, and typically results in incorrect interpretations—potentially with harmful consequences. We support this claim by reviewing results from cognitive neuroscience, behavior genetics, clinical psychology, personality psychology, educational psychology, intelligence research, and simulation studies. We show that Simpson's paradox is most likely to occur when inferences are drawn across different levels of explanation (e.g., from populations to subgroups, or subgroups to individuals). We propose a set of statistical markers indicative of the paradox, and offer psychometric solutions for dealing with the paradox when encountered—including a toolbox in R for detecting Simpson's paradox. We show that explicit modeling of situations in which the paradox might occur not only prevents incorrect interpretations of data, but also results in a deeper understanding of what data tell us about the world.

Keywords: paradox, measurement, reductionism, Simpson's paradox, statistical inference, ecological fallacy

INTRODUCTION

Two researchers, Mr. A and Ms. B, are applying for the same tenured position. Both researchers submitted a number of manuscripts to academic journals in 2010 and 2011: 60% of Mr. A's papers were accepted, vs. 40% of Ms. B's papers. Mr. A cites his superior acceptance rate as evidence of his academic qualifications. However, Ms. B notes that her acceptance rates were higher in *both* 2010 (25 vs. 0%) *and* 2011 (100 vs. 75%)[1]. Based on these records, who should be hired?[2]

In Simpson (1951) showed that a statistical relationship observed in a population—i.e., a collection of subgroups or individuals—could be reversed within all of the subgroups that make up that population[3]. This apparent paradox has significant implications for the medical and social sciences: A treatment that appears effective at the population-level may, in fact, have adverse consequences within each of the population's subgroups. For instance, a higher dosage of medicine may be associated with higher recovery rates at the population-level; however, *within* subgroups (e.g., for both males and females), a higher dosage may actually result in *lower* recovery rates. **Figure 1** illustrates this situation: Even though a negative relationship exists between "Treatment Dosage" and "Recovery" in both males and females, when these groups are combined a positive trend appears (black, dashed). Thus, if analyzed globally, these data would suggest that a higher dosage treatment is preferable, while the exact opposite is true (the continuous case is often referred to as *Robinson's paradox*, 1950)[4].

Simpson's paradox (hereafter SP) has been formally analyzed by mathematicians and statisticians (e.g., Blyth, 1972; Dawid, 1979; Pearl, 1999, 2000; Schield, 1999; Tu et al., 2008; Greenland, 2010; Hernán et al., 2011), its relevance for human inferences studied by psychologists (e.g., Schaller, 1992; Spellman, 1996a,b; Fiedler, 2000, 2008; Curley and Browne, 2001) and conceptually explored by philosophers (e.g., Cartwright, 1979; Otte, 1985; Bandyoapdhyay et al., 2011). However, few works have discussed the *practical* aspects of SP for empirical science: How might researchers prevent the paradox, recognize it, and deal with it upon detection? These issues are the focus of the present paper.

1

	2010	2011	overall
Mr. A	0 of 20	60 of 80	60%
Ms. B	20 of 80	20 of 20	40%

[2]The years in this example are substitutes for the true relevant variable, namely journal quality (together with diverging base rates of submission). This variable is substituted here to emphasize the puzzling nature of the paradox. See page 3 for further explanation of this (hypothetical) example.

[3]The same observation was made, albeit less explicitly, by Pearson et al. (1899), Yule (1903) and Cohen and Nagel (1934); see also Aldrich (1995).

[4]Julious and Mullee (1994) showed such a pattern in a data set bearing on treatment of kidney stones: Treatment A seemed more effective than treatment B in the dataset as a whole, but when split into small and large kidney stones (which, combined, formed the entire data set), treatment B was more effective for both.

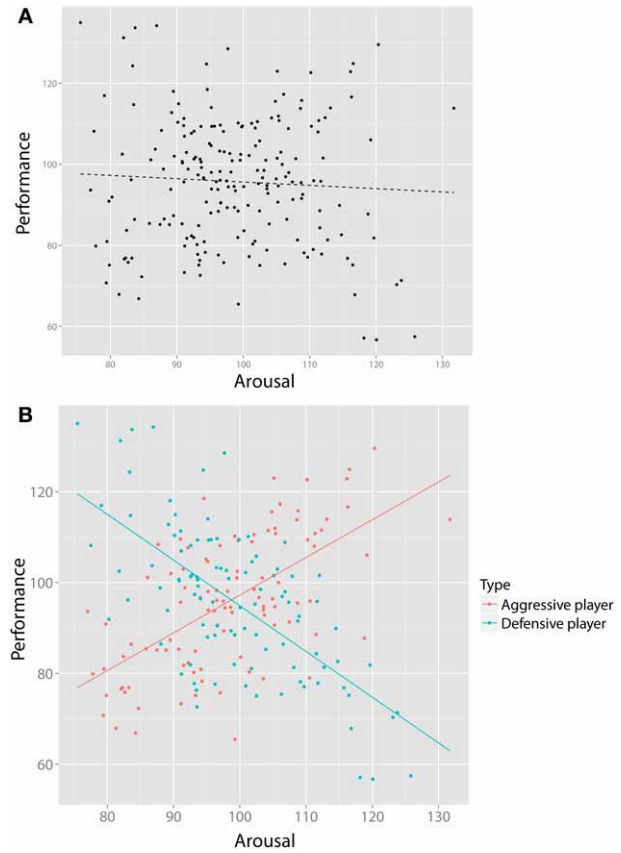


FIGURE 3 | Visualization alone does not always suffice. (A) shows the bivariate relationship between arousal and performance of tennis players, suggesting no relationship. However, after collecting new data on playing styles (e.g., how many winning shots, how many errors) we perform a cluster analysis yielding two types of players ("aggressive" and "defensive"). By including this new, bivariate variable, two clear and opposite relationships **(B)** emerge that would have gone unnoticed otherwise.