Correlation and Causality

Dr. Paul Larsen

February 25, 2020

Why causality matters

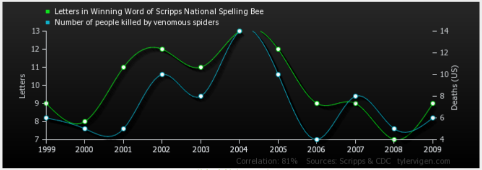
Because correlation is a proxy.

spurious correlations

Letters in Winning Word of Scripps National Spelling Bee

correlates with

Number of people killed by venomous spiders



Upload this image to imgur

	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009
Letters in Winning Word of Scripps National Spelling Bee *Letters (Scripps)*	9	8	11	12	11	13	12	9	9	7	9
Number of people killed by venomous spiders *Deaths (US) (CDC)*	6	5	5	10	8	14	10	4	8	5	6

Correlation: 0.8057

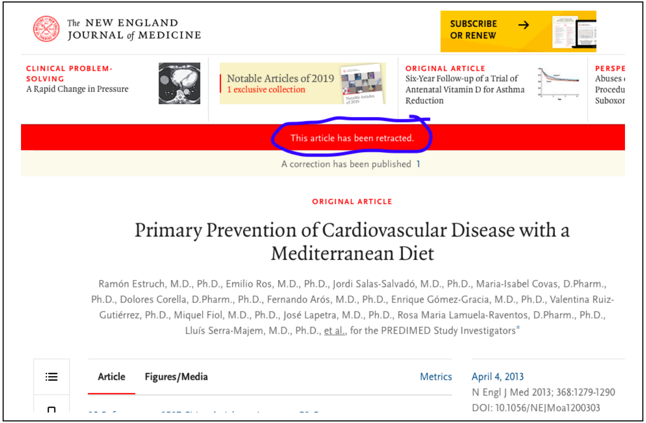
Permalink - Mark as interesting (6,161) - Not interesting (2,981)

View all correlations - Discover a new correlation

[Vig]

Why causality matters

Because A / B testing is not always possible.


The NEW ENGLAND JOURNAL of MEDICINE

SUBSCRIBE OR RENEW

CLINICAL PROBLEM-SOLVING
A Rapid Change in Pressure

Notable Articles of 2019
1 exclusive collection

ORIGINAL ARTICLE
Six-Year Follow-up of a Trial of Antenatal Vitamin D for Asthma Reduction

This article has been retracted.

A correction has been published 1

ORIGINAL ARTICLE

Primary Prevention of Cardiovascular Disease with a Mediterranean Diet

Ramón Estruch, M.D., Ph.D., Emilio Ros, M.D., Ph.D., Jordi Salas-Salvadó, M.D., Ph.D., Maria-Isabel Covas, D.Pharm., Ph.D., Dolores Corella, D.Pharm., Ph.D., Fernando Arós, M.D., Ph.D., Enrique Gómez-Gracia, M.D., Ph.D., Valentina Ruiz-Gutiérrez, Ph.D., Miquel Fiol, M.D., Ph.D., José Lapetra, M.D., Ph.D., Rosa Maria Lamuela-Raventos, D.Pharm., Ph.D., Lluís Serra-Majem, M.D., Ph.D., et al., for the PREDIMED Study Investigators*

Article Figures/Media Metrics

April 4, 2013
N Engl J Med 2013; 368:1279-1290
DOI: 10.1056/NEJMoa1200303

[ERSS$^+$13]

Simpson's paradox: cautionary tales

Simpson's paradox: a phenomenon in probability and statistics in which a trend appears disappears or reverses depending on grouping of data. [Wik], [PGJ16]

Example: University of California, Berkeley 1973 admission figures

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

[FPP98]

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

[BHO75]

A brief, biased history of causality

- Aristotle, 384 - 322 BC
- Isaac Newton, 1643 - 1727 AD
- David Hume, 1711 - 1776 AD
- Francis Galton, 1822 - 1900 AD, Karl Pearson, 1857 - 1936 AD
- Judea Pearl, b. 1936 AD

Counterfactuals and causality

Ideal: Intervention + Multiverse $\rightarrow$ Causality

Examples:

- Medical treatment (e.g. kidney stone treatment)
- Social outomes (e.g. university admissions)
- Business outcomes (e.g. click-through rate, hit rate)

In-practice:

- Correlation: approximate multiverse by comparing intervention at t to result at $t-1$
- Random population: approximate multiverse by splitting sample well
- A / B testing: random populations A / B + intervention in one

Counterfactual example: hit rate for insurance

Variables:

- product_type: Client line of business
- days: Number of days to generate quote
- rating: Binary indication of client risk
- hit: Binary, 1 for success (binding the quote), 0 for failure

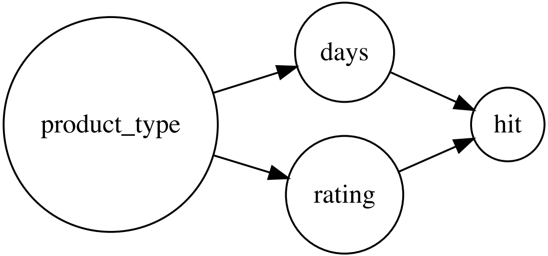
Fake data:

product_type	days	rating	hit
property	3	1	0
financial	2	1	0
financial	1	1	0
financial	0	0	1
financial	0	1	0

Counterfactual example: hit rate for insurance

Variables:

- product_type: Client line of business
- days: Number of days to generate quote
- rating: Binary indication of client risk
- hit: Binary, 1 for success (binding the quote), 0 for failure



Non-counterfactual approach: condition and query

Goal: estimate effect of days on hit.

Calculate

- $P(\mathrm{hit} = 1|\mathrm{days} = 0) - P(\mathrm{hit} = 1|\mathrm{days} = 1)$,
- $P(\mathrm{hit} = 1|\mathrm{days} = 1) - P(\mathrm{hit} = 1|\mathrm{days} = 2)$,
- $\ldots$

From exercise Jupyter notebook:

	hit
days	
0	0.539135
1	0.440035
2	0.326531
3	0.168289

The Structural Causal Model

The definitions in following slides are from [Pea07], [PGJ16].

Definition

A *structural causal model* M consists of two sets of variables U, V and a set of functions F, where

- U are considered *exogenous*, or background variables,
- V are the *causal* variables, i.e. that can be manipulated, and
- F are the functions that represent the process of assigning values to elements of V based on other values in U, V, e.g. $v_i = f(u, v)$.

We denote by G the graph induced on U, V by the functions F, and call it the *causal graph* of (U, V, F).

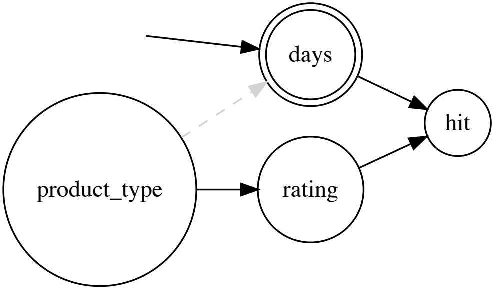
Hit rate example: $U = \{\text{product\_type}, \text{rating}\}$, $V = \{\text{days}, \text{hit}\}$, $F \leftrightarrow$ sample from conditional probabilty tables in directed graphical model.

Formalizing interventions: the intuition of “do”

For business application, quantity of interest is not $P(\text{hit} = 1|\text{days} = d)$, but intervention

$$P(\text{hit} = 1|\text{do}(\text{days} = d))$$

.

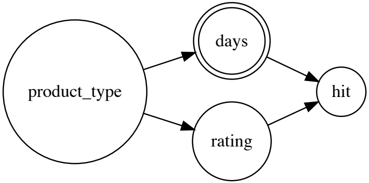


Formalizing interventions: the intuition of "do"

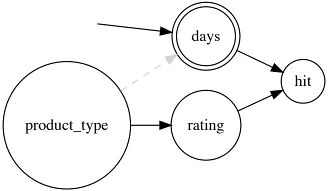
For business application, quantity of interest is effect of intervention / counterfactual

Not $P(\mathrm{hit} = 1 | \mathrm{days} = d)$ but $P(\mathrm{hit} = 1 | \mathrm{do}(\mathrm{days} = d))$

$G =$

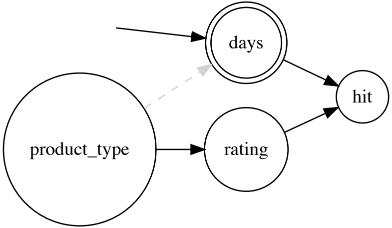


$G' = G_{\underline{\mathrm{days}}} =$



Formalizing interventions: the intuition of "do"

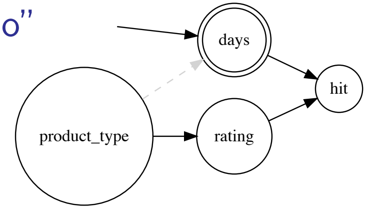
First, find quantities unchanged between G and $G' = G_{\underline{\text{days}}}$



$$P_{G'}(\text{product\_type} = p, \text{rating} = r) = P_G(\text{product\_type} = p, \text{rating} = r) \quad (1)$$

$$P_{G'}(\text{hit} = 1 | \text{product\_type} = p, \text{rating} = r) = P_G(\text{hit} = 1 | \text{product\_type} = p, \text{rating} = r) \quad (2)$$

Formalizing interventions: the intuition of "do"



$$
\begin{aligned}
P(\text{hit} = 1 | \text{do}(\text{days}) = d) \\
&= P_{G'}(\text{hit} = 1 | \text{days} = d), \text{ by definition} \\
&= \sum_{p,r} P_{G'}(\text{hit} = 1 | \text{days} = d, \text{product\_type} = p, \text{rating} = r) \\
&\qquad P_{G'}(\text{product\_type} = p, \text{rating} = r | \text{days} = d), \text{ by total probability} \\
&= \sum_{p,r} P_{G'}(\text{hit} = 1 | \text{days} = d, \text{product\_type} = p, \text{rating} = r) \\
&\qquad P_{G'}(\text{product\_type} = p, \text{rating} = r), \text{ by substitution} \\
&= \sum_{p,r} P_{G}(\text{hit} = 1 | \text{days} = d, \text{product\_type} = p, \text{rating} = r) \\
&\qquad P_{G}(\text{product\_type} = p, \text{rating} = r), \text{ our } \textit{adjustment} \text{ formula}
\end{aligned}
$$

References: [PGJ16], [Pro]

Causal hit rate

Typical quantity of interest: *average treatment effect* or *ATE*

$P(\text{hit} = 1|\text{days} = d)$

	hit
days	
0	0.539135
1	0.440035
2	0.326531
3	0.168289

Example ATE:
$P(\text{hit} = 1|\text{days} = 2)$
$-P(\text{hit} = 1|\text{days} = 3) \approx 16\%$

$P(\text{hit} = 1|\text{do}(\text{days} = d))$

	prob
days	
0	0.549247
1	0.410495
2	0.292335
3	0.215497

Example causal ATE:
$P(\text{hit} = 1|\text{do}(\text{days}) = 2)$
$-P(\text{hit} = 1|\text{do}(\text{days}) = 3) \approx 8\%$

Judea Pearl's Rules of Causality

Let X, Y, Z and W be arbitrary disjoint sets of nodes in a DAG G. Let $G_{\underline{X}}$ be the graph obtained by removing all arrows pointing into (nodes of) X. Denote by $G_{\overline{X}}$ the graph obtained by removing all arrows pointing out of X. If, e.g. we remove arrows pointing out of X and into Z, we the resulting graph is denoted by $G_{\underline{X}\overline{Z}}$

Rule 1: Insertion / deletion of observations

$$P(y|\mathrm{do}(x), z, w) = P(y|\mathrm{do}(x), w) \text{ if } (Y \perp\!\!\!\perp Z|X, W)_{G_{\overline{X}}}$$

Rule 2: Action / observation exchange

$$P(y|\mathrm{do}(x), \mathrm{do}(z), w) = P(y|\mathrm{do}(x), z, w) \text{ if } (Y \perp\!\!\!\perp Z|X, W)_{G_{\overline{X}\underline{Z}}}$$

Rule 3: Insertion / deletion of actions

$$P(y|\mathrm{do}(x), \mathrm{do}(z), w) = P(y|\mathrm{do}(x), w) \text{ if } (Y \perp\!\!\!\perp Z|X, W)_{G_{\overline{X}\overline{Z(W)}}},$$

where $Z(W)$ is the set of Z-nodes that are not ancestors of any W-node in $G_{\underline{X}}$.

Special cases of the causal rules

By judicious setting of sets of nodes to be empty, we obtain some useful corollaries of the causal rules.

Rule 1': Insertion / deletion of observations, with $W = \emptyset$

$$P(y|\mathrm{do}(x), z) = P(y|\mathrm{do}(x)) \text{ if } (Y \perp\!\!\!\perp Z|X)_{G_{\overline{X}}}$$

Rule 2': Action / observation exchange, with $X = \emptyset$

$$P(y|\mathrm{do}(z), w) = P(y|z, w) \text{ if } (Y \perp\!\!\!\perp Z|W)_{G_{\underline{Z}}}$$

Rule 3': Insertion / deletion of actions, with $X, W = \emptyset$

$$P(y|\mathrm{do}(z)) = P(y) \text{ if } (Y \perp\!\!\!\perp Z)_{G_{\overline{Z}}}$$

Special cases of the causal rules

By judicious setting of sets of nodes to be empty, we obtain some useful corollaries of the causal rules.

Rule 1': Insertion / deletion of observations, with $W = \emptyset$

$$P(y|\mathrm{do}(x), z) = P(y|\mathrm{do}(x)) \text{ if } (Y \perp\!\!\!\perp Z|X)_{G_{\overline{X}}}$$

Rule 2': Action / observation exchange, with $X = \emptyset$

$$P(y|\mathrm{do}(z), w) = P(y|z, w) \text{ if } (Y \perp\!\!\!\perp Z|W)_{G_{\underline{Z}}}$$

Rule 3': Insertion / deletion of actions, with $X, W = \emptyset$

$$P(y|\mathrm{do}(z)) = P(y) \text{ if } (Y \perp\!\!\!\perp Z)_{G_{\overline{Z}}}$$

$\Longrightarrow$ d-separation + causal rules = *adjustment formulas*: do queries as normal queries.

References I

[BHO75] P. J. Bickel, E. A. Hammel, and J. W. O'Connell, *Sex Bias in Graduate Admissions: Data from Berkeley*, Science **187** (1975), no. 4175, 398–404.

[ERSS+13] Ramón Estruch, Emilio Ros, Jordi Salas-Salvadó, Maria-Isabel Covas, Dolores Corella, Fernando Arós, Enrique Gómez-Gracia, Valentina Ruiz-Gutiérrez, Miquel Fiol, José Lapetra, et al., *Primary prevention of cardiovascular disease with a mediterranean diet*, New England Journal of Medicine **368** (2013), no. 14, 1279–1290.

[FPP98] D. Freedman, R. Pisani, and R. Purves, *Statistics*, W.W. Norton, 1998.

[Pea07] Judea Pearl, *The mathematics of causal inference in statistics*, To appear in 2007 JSM Proceedings **337** (2007).

[PGJ16] Judea Pearl, Madelyn Glymour, and Nicholas P Jewell, *Causal inference in statistics: A primer*, John Wiley & Sons, 2016.

References II

[Pro] Christopher Prohm, *Causality and function approximation*, https://cprohm.de/article/causality-and-function-approximations.html.

[Vig] Typer Vigen, *Spurious Correlations, Spiders and Spelling-Bees*, http://tylervigen.com/view_correlation?id=2941.

[Wik] Wikipedia, *Simpson's paradox*, https://en.wikipedia.org/wiki/Simpson's_paradox.