Journal of Modern Applied Statistical Methods

Volume 17 | Issue 2 Article 4

3-6-2019

Should We Give Up on Causality?

Tom Knapp
The Ohio State University, tomknapp5@gmail.com

Journal of Modern Applied Statistical Methods
November 2018, Vol. 17, No. 2, eP2720
doi: 10.22237/jmasm/1551907445

בס"ד

ISSN 1538 − 9472

INVITED ARTICLE

Should We Give Up on Causality?

Tom Knapp
The Ohio State University
Columbus, OH

Keywords: Causality, randomization

Introduction

Researcher A randomly assigned forty members of a convenience sample of middle school students to one of five different amounts of remedial reading instructional minutes (eight students for each amount), determined the number of books each student subsequently chose to read, and carried out a test of the significance of the difference among the five mean numbers of books read. Researcher B had access to the school records for a random sample of forty middle school students, determined the number of minutes of remedial reading instruction each student received, the number of books that each student read, and calculated the correlation (Pearson product-moment) between number of minutes and number of books. Researcher A's study has a stronger basis for causality (internal validity). Researcher B's study has a stronger basis for generalizability (external validity). Which of the two studies contributes more to the advancement of knowledge?

Do you need to see the data before you answer the question? The raw data are the same for both studies:

ID	Minutes	Books	ID	Minutes	Books
1	75	5	6	75	15
2	75	10	7	75	15
3	75	10	8	75	20
4	75	10	9	125	10
5	75	15	10	125	15

doi: 10.22237/jmasm/1551907445 | Accepted: January 24, 2018; Published: March 6, 2019.
Correspondence: Tom Knapp, tomknapp5@gmail.com

ID	Minutes	Books	ID	Minutes	Books
11	125	15	26	225	25
12	125	15	27	225	25
13	125	20	28	225	25
14	125	20	29	225	30
15	125	20	30	225	30
16	125	25	31	225	30
17	175	15	32	225	35
18	175	20	33	275	25
19	175	20	34	275	30
20	175	20	35	275	30
21	175	25	36	275	30
22	175	25	37	275	35
23	175	25	38	275	35
24	175	30	39	275	35
25	225	20	40	275	40

Here are the results for the two analyses (using Excel and Minitab):

SUMMARY

Group	Count	Sum	Mean	Variance
75 mins	8	100	12.5	21.43
125 mins	8	140	17.5	21.43
175 mins	8	180	22.5	21.43
225 mins	8	220	27.5	21.43
275 mins	8	260	32.5	21.43

ANOVA

Source of Variation	SS	df	MS	F
Between Groups	2000	4	500	23.33
Within Groups	750	35	21.43	
Total	2750	39		

Correlation between Minutes and Books = 0.853

The regression equation is:
Books = 5.00 + 0.10 Minutes

Predictor	Coef	Standard error	t-ratio
Constant	5.00	1.88	2.67
Minutes	0.10	0.0099	10.07

s = 4.44 R-sq = 72.7% R-sq(adj) = 72.0%

Analysis of Variance Table

SOURCE	DF	SS	MS
Regression	1	2000	2000
Error	38	750	19.7
Total	39	2750	

The results are virtually identical. (Given that both approaches are subsumed under the general linear model, that is not surprising.) There is only that tricky difference in the *df*s associated with the fact that hours is discrete in the ANOVA (its magnitude never entered the analysis) and continuous in the correlation and regression analyses.

But What About the Assumptions?

Here is the overall frequency distribution for Books:

Books	Count
5	1
10	4
15	7
20	8
25	8
30	7
35	4
40	1

It appears normally distributed. Here is the frequency distribution of number of books read for each of the five groups: (This is relevant for homogeneity of variance in the ANOVA and for homoscedasticity in the regression.)

Books	Count	Mins = 75	*n* = 8
5	1		
10	3		
15	3		
20	1		

Books	Count	Mins = 125	*n* = 8
10	1		
15	3		
20	3		
25	1		

Books	Count	Mins = 175	*n* = 8
15	1		
20	3		
25	3		
30	1		

Books	Count	Mins = 225	*n* = 8
20	1		
25	3		
30	3		
35	1		

Books	Count	Mins = 275	*n* = 8
25	1		
30	3		
35	3		
40	1		

Those distributions are as normal as they can be for eight observations per group. (They're actually the binomial coefficients for $n = 3$.)

So What?

The "So what?" is that the conclusion is essentially the same for the two studies; i.e., there is a strong linear association between minutes of remedial reading instruction and number of books read. The regression equation for Researcher B's study can be used to predict books from minutes quite well for the population from which their sample was randomly drawn. They are likely to be only off by 5-10 books in number of books read, since the standard error of estimate, s, = 4.44. Why do we need the causal interpretation provided by Researcher A's study? Isn't the greater generalizability of Researcher B's study more important than whether or not the effect of minutes on books is causal for the non-random sample? These data are

admittedly artificial (for illustrative purposes). Real data are never that clean, but they could be.

What is Typically Stated Regarding Causation, Correlation, and Prediction?

The sources cited most often for distinctions among causation (using the terms "causality" and "causation" interchangeably), correlation, and prediction are classics written by philosophers such as Mill (1884) and Popper (1959); textbook authors such as Pearl (2000); and journal articles such as Hill (1965) and Holland (1986a) (with comments Rubin, 1986; Cox, 1986; Glymour, 1986; Granger, 1986; and rejoinder Holland, 1986b). However, consider Frakt (2009) and White (2010):

Frakt (2009)

In an unusual twist, Frakt (2009) argued causation may exist without correlation. (The usual minimum three criteria for a claim that *X* causes *Y* are strong correlation, temporal precedence, and non-spuriousness.) An example was given in which the true relationship between *X* and *Y* is mediated by a third variable *W*, where the correlation between *X* and *Y* is equal to zero.

White (2010)

White (2010) decried the endless repetition of "correlation is not causation."

He argued most knowledge is correlational knowledge; causal knowledge is only necessary when control is needed. Causation is a slippery concept, and correlation and causation go hand-in-hand more often than imagined.

In the spirit of this distinction between correlational knowledge and causal knowledge, can it be agreed the focus of research efforts should be on two non-overlapping strategies: true experiments (randomized controlled trials) carried out on non-random samples, with replications wherever possible; and non-experimental correlational studies carried out on random samples, also with replications?

What about the effect of smoking (firsthand, secondhand, thirdhand...) on lung cancer? It may be necessary to give up on causality even there. There are problems regarding the difficulty of establishing a causal connection between the two, even for firsthand smoking. See, for example, Spirtes, Glymour, and Scheines (2000, pp. 239-240).

References

Cox, D. R. (1986). Statistics and causal inference: Comment. *Journal of the American Statistical Association, 81*(396), 963-964. doi: 10.2307/2289066

Frakt, A. (2009, December 16). Causation without correlation is possible [web log post]. Retrieved from https://theincidentaleconomist.com/wordpress/causation-without-correlation-is-possible/

Glymour, C. (1986). Statistics and causal inference: Comment: Statistics and metaphysics. *Journal of the American Statistical Association, 81*(396), 964-966. doi: 10.2307/2289067

Granger, C. (1986). Statistics and causal inference: Comment. *Journal of the American Statistical Association, 81*(396), 967-968. doi: 10.2307/2289068

Hill, A. B. (1965). The environment and disease: Association or causation? *Proceedings of the Royal Society of Medicine, 58*(5), 295-300.

Holland, P. W. (1986a). Statistics and causal inference. *Journal of the American Statistical Association, 81*(396), 945-960. doi: 10.2307/2289064

Holland, P. W. (1986b). Statistics and causal inference: Rejoinder. *Journal of the American Statistical Association, 81*(396), 968-970. doi: 10.2307/2289064

Mill, J. S. (1884). *A system of logic, ratiocinative and inductive: Being a connected view of the principles of evidence, and methods of scientific investigation*. London, UK: Longmans, Green, and Co.

Pearl, J. (2000). *Causality: Models, reasoning, and inference*. New York, NY: Cambridge University Press.

Popper, K. (1959). *The logic of scientific discovery*. London, UK: Routledge.

Rubin, D. B. (1986). Statistics and causal inference: Comment: Which ifs have causal answers. *Journal of the American Statistical Association, 81*(396), 961-962. doi: 10.2307/2289065

Spirtes, P., Glymour, C., & Scheines, R. (2001). Causation, prediction, and search (2nd ed.). Cambridge, MA: The MIT Press. doi: 10.7551/mitpress/1754.001.0001

White, J. M. (2010, October 1). Three-quarter truths: Correlation is not causation [web log post]. Retrieved from http://www.johnmyleswhite.com/notebook/2010/10/01/three-quarter-truths-correlation-is-not-causation/