Should we give up on causality?

Thomas R. Knapp

Introduction

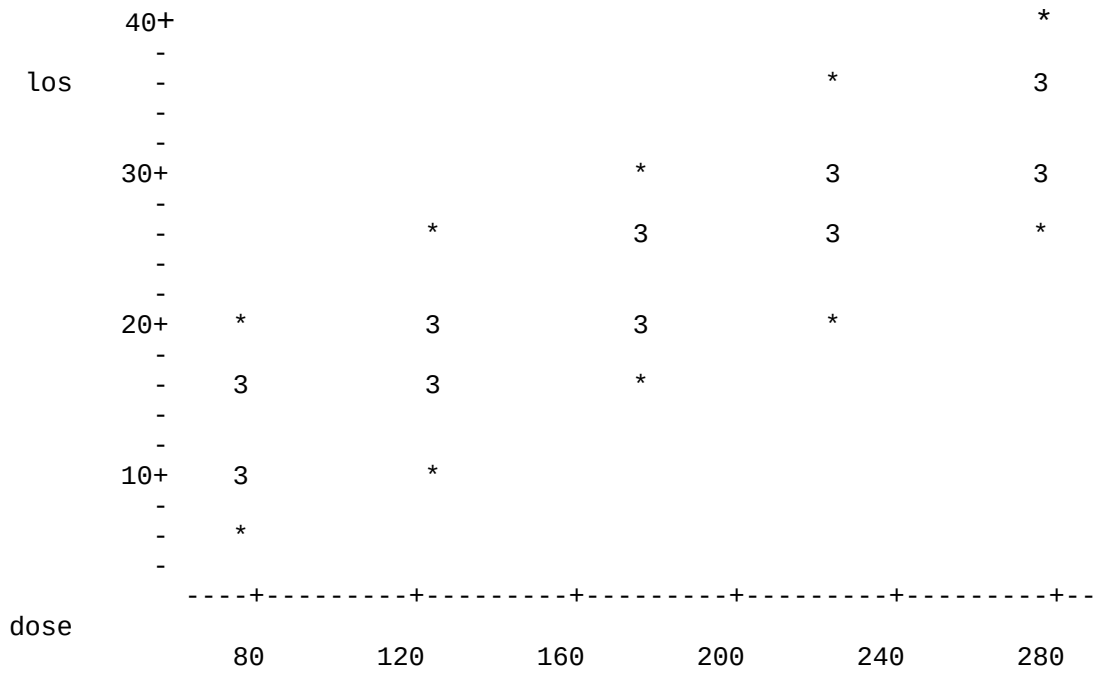
Researcher A randomly assigns forty members of a convenience sample of hospitalized patients to one of five different daily doses of aspirin (eight patients per dose), determines the length of hospital stay for each person, and carries out a test of the significance of the difference among the five mean stays. Researcher B has access to hospital records for a random sample of forty patients, determines the daily dose of aspirin given to, and the length of hospital stay for, each person, and calculates the correlation (Pearson product-moment) between dose of aspirin and length of stay. 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?

Oh; do you need to see the data before you answer the question? The raw data are the same for both studies. Here they are:

ID	Dose(in mg)	LOS(in days)	ID	Dose(in mg)	LOS(in days)
1	75	5	21	175	25
2	75	10	22	175	25
3	75	10	23	175	25
4	75	10	24	175	30
5	75	15	25	225	20
6	75	15	26	225	25
7	75	15	27	225	25
8	75	20	28	225	25
9	125	10	29	225	30
10	125	15	30	225	30
11	125	15	31	225	30
12	125	15	32	225	35
13	125	20	33	275	25
14	125	20	34	275	30
15	125	20	35	275	30
16	125	25	36	275	30
17	175	15	37	275	35
18	175	20	38	275	35
19	175	20	39	275	35
20	175	20	40	275	40

And here are the results for the two analyses (courtesy of Minitab):

```
ANALYSIS OF VARIANCE ON los
 SOURCE     DF        SS        MS        F
 treat       4    2000.0     500.0    23.33
 ERROR      35     750.0      21.4
 TOTAL      39    2750.0
                                    INDIVIDUAL 95 PCT CI'S FOR MEAN
                                    BASED ON POOLED STDEV
 LEVEL       N      MEAN     STDEV  ---------+---------+---------+-----
--
     1       8    12.500     4.629  (----*---)
     2       8    17.500     4.629         (---*---)
     3       8    22.500     4.629               (---*---)
     4       8    27.500     4.629                     (---*----)
     5       8    32.500     4.629                           (----*---)
                                    ---------+---------+---------+-----
--
 POOLED STDEV =    4.629                 16.0      24.0      32.0
```



Correlation of dose and los = 0.853

The regression equation is
los = 5.00 + 0.100 dose

Predictor	Coef	Stdev	t-ratio
Constant	5.000	1.875	2.67
dose	0.100000	0.009934	10.07

s = 4.443 R-sq = 72.7% R-sq(adj) = 72.0%

Analysis of Variance

SOURCE	DF	SS	MS
Regression	1	2000.0	2000.0
Error	38	750.0	19.7
Total	39	2750.0	

The results are almost identical. (For those of you familiar with "the general linear model" that is not surprising.) There is only that tricky difference in the df's associated with the fact that dose is discrete in the ANOVA (its magnitude never even enters the analysis) and continuous in the correlation and regression.

But what about the assumptions?

Here is the over-all frequency distribution for LOS:

```
    Midpoint   Count
        5       1  *
       10       4  ****
       15       7  *******
       20       8  ********
       25       8  ********
       30       7  *******
       35       4  ****
       40       1  *
```

Looks pretty normal to me.

And here is the LOS frequency distribution for each of the five treatment groups: (This is relevant for homogeneity of variance in the ANOVA and for homoscedasticity in the regression.)

```
 Histogram of los   treat = 1   N = 8
 Midpoint   Count
        5       1  *
       10       3  ***
       15       3  ***
       20       1  *
 Histogram of los   treat = 2   N = 8
 Midpoint   Count
       10       1  *
       15       3  ***
       20       3  ***
       25       1  *
 Histogram of los   treat = 3   N = 8
 Midpoint   Count
       15       1  *
       20       3  ***
       25       3  ***
       30       1  *
 Histogram of los   treat = 4   N = 8
 Midpoint   Count
       20       1  *
       25       3  ***
       30       3  ***
       35       1  *
 Histogram of los   treat = 5   N = 8
 Midpoint   Count
       25       1  *
       30       3  ***
       35       3  ***
       40       1  *
```

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

So what?

The "So what?" is that the statistical conclusion is essentially the same for the two studies; i.e., there is a strong linear association between dose and stay. The regression equation for Researcher B's study can be used to predict stay from dose quite well for the population from which his (her) sample was randomly drawn. (You're only likely to be off by 5-10 days in length of stay.) Why do we need the causal interpretation for Researcher A's study? Isn't the greater generalizability of Researcher B's study more important than whether or not the "effect" of dose on stay is causal?

You're probably thinking "Yeah; big deal, for this one example of artificial data." Of course the data are artificial (for illustrative purposes). Real data are never that clean, but they could be.

Read on.

What do other people have to say about causation, correlation, and prediction?

The sources cited most often for distinctions among causation (I use the terms "causality" and "causation" interchangeably), correlation, and prediction are usually classics written by philosophers such as Mill (1884) and Popper (1959); textbook authors such as Pearl (2000); and journal articles such as Bradford Hill (1965) and Holland (1986). I would like to cite a few other lesser known people who have had something to say for or against the position I have just taken. I happily exclude those who say only that "correlation is not causation" and let it go at that.

Schield (1995):

My friend Milo Schield is very big on emphasizing the matter of causation in the teaching of statistics. Although he included in his conference presentation the mantra "correlation is not causality", he carefully points out that students might mistakenly think that correlation can never be causal. He goes on to argue for the need to make other important distinctions among causality, explanation, determination, prediction, and other terms that are often confused with one another. Nice piece.

Frakt (2009):

In an unusual twist, Austin Frakt argues that you can have causation without correlation. (The usual minimum three criteria for a claim that X causes Y are strong correlation, temporal precedence, and non-spuriousness.) He gives an

example for 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):

John Myles White decries the endless repetiton of "correlation is not causation". He argues that most of our knowledge is correlational knowledge; causal knowledge is only necessary when we want to control things; causation is a slippery concept; and correlation and causation go hand-in-hand more often than some people think. His take-home message is that it's much better to know X and Y are related than it is to know nothing at all.

Anonymous (2012):

Anonymous starts out his (her) two-part article with this: "The ultimate goal of social science is causal explanation. The actual goal of most academic research is to discover significant relationships between variables." Ouch! But true? He (she) contends that we can detect a statistically significant effect of X on Y but still not know why and when Y occurs.

That looks like three (Schield, Frakt, and Anonymous) against two (White and me), so I lose? Perhaps. How about a compromise? In the spirit of White's distinction between correlational knowledge and causal knowledge, can we agree that we should concentrate our research efforts on two non-overlapping strategies: true experiments (randomized clinical trials) carried out on admittedly handy non-random samples, with replications wherever possible; and non-experimental correlational studies carried out on random samples, also with replications?

A closing note

What about the effect of smoking (firsthand, secondhand, thirdhand...whatever) on lung cancer? Would you believe that we might have to give up on causality there? There are problems regarding the difficulty of establishing a causal connection between the two even for firsthand smoking. You can look it up (in Spirtes, Glymour, & Scheines, 2000, pp.239-240). You might also want to read the commentary by Lyketsos and Chisolm (2009), the letter by Luchins (2009) regarding that commentary, and the reply by Lyketsos and Chisolm (2009) concerning why it is sometimes not reported that smoking was responsible for the death of a smoker who had lung cancer (whereas stress as a cause for suicide almost always is).

References

Anonymous (2012). Explanation and the quest for 'significant' relationships. Parts 1 and 2. Downloaded from the Rules of Reason website on the internet.

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

Frakt, A. (2009). Causation without correlation is possible. Downloaded from The Incidental Economist website on the internet.

Holland, P.W. (1986). Statistics and causal inference. Journal of the American Statistical Association, 81 (396), 945-970. [Includes comments by D.B. Rubin, D.R. Cox, C.Glymour, and C.Granger, and a rejoinder by Holland.]

Luchins, D.J. (2009). Meaningful explanations vs. scientific causality. JAMA, 302 (21), 2320.

Lyketsos, C.G., & Chisolm, M.S. (2009). The trap of meaning: A public health tragedy. JAMA, 302 (4), 432-433.

Lyketsos, C.G., & Chisolm, M.S. (2009). In reply. JAMA, 302 (21), 2320-2321.

Mill, J. S. (1884). A system of logic, ratiocinative and Inductive. London: Longmans, Green, and Co.

Pearl, J. (2000). Causality. New York: Cambridge University Press.

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

Schield, M. (1995). Correlation, determination, and causality in introductory statistics. Conference presentation, Annual Meeting of the American Statistical Association.

Spirtes, P., Glymour, C., & Scheines, R. (2000). Causation, prediction, and search. (2nd. ed.) Cambridge, MA: The MIT Press.

White, J.M. (2010). Three-quarter truths: correlation is not causation. Downloaded from his website on the internet.