

Revision list for Pearl's
THE FOUNDATIONS OF CAUSAL INFERENCE

insert – p. 90:

in graphical terms or plain causal language. The mediation problem of Section 6 illustrates how such symbiosis clarifies the definition and identification of direct and indirect effects, a task deemed insurmountable, “deceptive” and “ill-defined” by advocates of the structureless potential-outcome approach (Rubin, 2004, 2005).

remove paragraph – pp. 90-91 starting:

“In contrast, when the mediation problem is approached...” including printed footnote 29.

new footnote – p. 99:

The need to control for mediator-outcome confounders (e.g., W_2 in Figure 6(b)) was evidently overlooked in the classical paper of Baron and Kenny (1986), and has subsequently been ignored by most social science researchers.

revised paragraph – p. 100:

Graphical identification conditions for multi-action expressions of the type $E(Y|do(x), do(z_1), do(z_2), \dots, do(z_k))$ in the presence of unmeasured confounders were derived by Pearl and Robins (1995) (see Pearl, 2000, ch. 4) using sequential application of the back-door conditions discussed in Section 3.2.

new footnote – p. 101

Pearl (2001) used the acronym *NDE* to denote the natural direct effect. We will delete the letter “*N*” from the acronyms of both the direct and indirect effect, and use *DE* and *IE*, respectively.

best break for equation 48 – p. 104:

$$DE_{x,x'}(Y) = \sum_z \sum_{w_1} P(w_1)[E(Y|x', z, w_1) - E(Y|x, z, w_1)] \sum_{w_2} P(z|x, w_2)P(w_2). \quad (48)$$

revised – pp. 110–11

linear version of Figure 6(a) (equation 45), which reads

$$\begin{aligned}x &= u_X \\z &= b_0 + b_x x + u_Z \\y &= c_0 + c_x x + c_z z + u_Y\end{aligned}\tag{54}$$

with u_X, u_Y , and u_Z uncorrelated, zero-mean error terms. Computing the conditional expectation in (53) gives

$$E(Y|x, z) = E(c_0 + c_x x + c_z z + u_Y) = c_0 + c_x x + c_z z$$

and yields

$$\begin{aligned}IE_{x,x'}(Y) &= \sum_z (c_x x + c_z z) [P(z|x') - P(z|x)] \\&= c_z [E(Z|x') - E(Z|x)]\end{aligned}\tag{55}$$

$$= (x' - x)(c_z b_x)\tag{56}$$

$$= (x' - x)(b - c_x)\tag{57}$$

where b is the total effect coefficient,

$$b = (E(Y|x') - E(Y|x))/(x' - x) = c_x + c_z b_x.$$

We thus obtained the standard expressions for indirect effects in linear systems, which can be estimated either as a difference in two regression coefficients (equation 57) or a product of two regression coefficients (equation 56), with Y regressed on both X and Z . (see MacKinnon, Lockwood, et al., 2007). These two strategies do not generalize to nonlinear systems as shown in Pearl (2010a); direct application of (53) is necessary.

To understand the difficulty, consider adding an interaction term $c_{xz}xz$ to the model in equation (54), yielding

$$y = c_0 + c_x x + c_z z + c_{xz}xz + u_Y$$

Now assume that, through elaborate regression analysis, we obtain accurate estimates of all parameters in the model. It is still not clear what combinations of parameters measure the direct and indirect effects of X on Y , or, more specifically, how to assess the fraction of the total effect that is *explained* by mediation and the fraction that is *owed* to mediation. In linear

analysis, the former fraction is captured by the product $c_z b_x / b$ (equation 56), the latter by the difference $(b - c_x) / b$ (equation 57) and the two quantities coincide. In the presence of interaction, however, each fraction demands a separate analysis, as dictated by the Mediation Formula.

To witness, substituting the nonlinear equation in (49), (52) and (53) and assuming $x = 0$ and $x' = 1$, yields the following decomposition:

$$\begin{aligned} DE &= c_x + b_0 c_{xz} \\ IE &= b_x c_z \\ TE &= c_x + b_0 c_{xz} + b_x (c_z + c_{xz}) \\ &= DE + IE + b_x c_{xz} \end{aligned}$$

We therefore conclude that the fraction of output change for which mediation would be *sufficient* is

$$IE/TE = b_x c_z / (c_x + b_0 c_{xz} + b_x (c_z + c_{xz}))$$

while the fraction for which mediation would be *necessary* is

$$1 - DE/TE = b_x (c_z + c_{xz}) / (c_x + b_0 c_{xz} + b_x (c_z + c_{xz}))$$

We note that, due to interaction, a direct effect can be sustained even when the parameter c_x vanishes and, moreover, a total effect can be sustained even when both the direct and indirect effects vanish. This illustrates that estimating parameters in isolation tells us little about the effect of mediation and, more generally, mediation and moderation are intertwined and cannot be assessed separately.

If the policy evaluated aims to prevent the outcome Y by ways of weakening the mediating pathways, the target of analysis should be the difference $TE - DE$, which measures the highest prevention effect of any such policy. If, on the other hand, the policy aims to prevent the outcome by weakening the direct pathway, the target of analysis should shift to IE , for $TE - IE$ measures the highest preventive impact of this type of policies.

The main power of the Mediation Formula shines in studies involving categorical variables, especially when we have no parametric model of the data generating process. To illustrate, consider the case where [...continue]

new footnote – p. 105:

Some authors (e.g., VanderWeele 2009), define the natural indirect effect as the difference $TE - DE$. This renders the additive formula a tautology of definition, rather than a theorem, conditioned upon the anti-symmetry $IE_{x,x'}(Y) = -IE_{x',x}(Y)$. Violation of (52) will be demonstrated in the next section.

corrected Figure 7 – p. 112 (number of samples → Number of Samples)

Number of Samples	X	Z	Y	$E(Y x, z) = \mathbf{g}_{xz}$	$E(Z x) = \mathbf{h}_x$
n_1	0	0	0	$\frac{n_2}{n_1+n_2} = g_{00}$	$\frac{n_3+n_4}{n_1+n_2+n_3+n_4} = h_0$
n_2	0	0	1		
n_3	0	1	0	$\frac{n_4}{n_3+n_4} = g_{01}$	
n_4	0	1	1		
n_5	1	0	0	$\frac{n_6}{n_5+n_6} = g_{10}$	$\frac{n_7+n_8}{n_5+n_6+n_7+n_8} = h_1$
n_6	1	0	1		
n_7	1	1	0	$\frac{n_8}{n_7+n_8} = g_{11}$	
n_8	1	1	1		

Figure 7: Computing the Mediation Formula for the model in Figure 6(a), with X, Y, Z binary.

References

- BARON, R. and KENNY, D. (1986). The moderator-mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology* **51** 1173–1182.
- MACKINNON, D., LOCKWOOD, C., BROWN, C., WANG, W. and HOFFMAN, J. (2007). The intermediate endpoint effect in logistic and probit regression. *Clinical Trials* **4** 499–513.
- PEARL, J. (2000). *Causality: Models, Reasoning, and Inference*. Cambridge University Press, New York. 2nd edition, 2009.
- PEARL, J. and ROBINS, J. (1995). Probabilistic evaluation of sequential plans from causal models with hidden variables. In *Uncertainty in Artificial Intelligence 11* (P. Besnard and S. Hanks, eds.). Morgan Kaufmann, San Francisco, 444–453.
- RUBIN, D. (2004). Direct and indirect causal effects via potential outcomes. *Scandinavian Journal of Statistics* **31** 161–170.
- RUBIN, D. (2005). Causal inference using potential outcomes: Design, modeling, decisions. *Journal of the American Statistical Association* **100** 322–331.