Introduction to Classical Mediation Methods

Ashley I Naimi

September 2024

Contents

1 Mediation: Background, Context, and Eras

Mediation analysis describes an analytic scenario in which we are interested in understanding what role a potential intermediary variable plays in explaining the relationship between an exposure and an outcome of interest. The simplest diagram we can use to illustrate the concept of mediation is displayed in Figure **??**.

Figure 1: Illustration of overly simplified mediation scenario with a single expsoure (X), a single mediator (M), and a single outcome (Y).

Mediation questions are everywhere in the empirical sciences. For example, Mendola and colleagues evaluated the role that gestational age at birth plays in mediating the relationship between preeclampsia and perinatal fetal/infant mortality (Mendola et al., 2015). VanderWeele et al looked at the role that smoking plays in mediating the relationship between the haplotype locus 15q25.1 and lung cancer development (VanderWeele et al., 2012). Chatterjee et al looked at whether serum potassium concentrations mediated the relation between race and incident diabetes (Chatterjee et al., 2011). And Ananth and VanderWeele looked at the role that preterm birth played in mediating the relationship between placental abruption and perinatal fetal/infant mortality (Ananth and VanderWeele, 2011).

In all cases, there is a known or established relationship between the exposure and the outcome, and the goal of the analysis is to evaluate whether some intermediary explains all, some, or none of this relationship.

1.1 Preclassical Mediation

Mediation analysis has a long history in science. Indeed Sewall Wright indended mediation analysis to be the "… primary purpose of the method of path coefficients" [Wright (1934); p177].

Some of the earliest modern examples of mediation analysis begin to appear in the scientific record in the mid-1950s. For example, Hyman (1955) was one of the first examples, referring to mediation analysis as *elaboration*.[1]

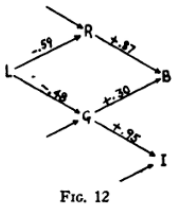

FIG. 12

[1] It seems, however, that the author of the chapter on "elaboration" was not Hyman himself, but Patricia Kendall, who was not permitted to be on faculty at Columbia due to the nepotism rules of the time. See https://is.gd/p134mN

1.2 Classical Mediation

In 1975, Alwin and Hauser revived the method of path analysis, and described a "general method for decomposing total effects into their constituent direct and indirect components". This was perhaps the first modern usage of terms more commonly used in the modern literature, including effect decomposition, total effects, direct effects, and indirect effects. This was later developed and applied to specific problems by a number of authors, notably Judd and Kenny (1981) and Baron and Kenny (1986).

The publication by Baron and Kenny (1986) became the cornerstone of the development and use of methods for mediation analysis in a wide range of settings for nearly three decades following it's publication. Indeed, by 2014, this paper had been cited nearly 50,000 times according to Google Scholar.[2]

[2] At this point ten years later, Google Scholar notes over 130,000 citations. However, I assume that a growing proportion of these since 2014 have been notably about the limitations of "Baron and Kenny" approach to mediation, as we will soon see.

1.3 Modern Mediation: Pure/Total Direct/Indirect Effects

It was only in the early 1990s when researchers began to realize that there was some unrecognized nuance in how we defined things like total, direct, and indirect effects. Furthermore, this nuance had important implications for how we might conduct a mediation analysis.

Jamie Robins and Sander Greenland were the first to use potential response types to separate a total effect into two components they termed pure direct/indirect and total direct/indirect effects (Robins and Greenland, 1992). To understand these direct and indirect effects, it's important to first understand counterfactual or potential outcome notation. A potential outcome is an outcome that would have been observed had an individual's exposure (or mediator) variable been set to a specific value, possibly counter to fact.

We can write potential outcomes in a number of ways. Using subscripts, for example, we might write:

$$Y_{x=0}$$

which we would read as the outcome Y that would have been observed if the exposure x was set to zero. Sometimes, we omit some elements of the counterfactual notation. For example, we can simplify the above to be Y_0.

With mediation quantities, we can use counterfactuals such as:

$$Y_{xm}$$

which is the outcome that would be observed if the exposure was set to some value x AND the mediator was set to some value m.

However, things can get more complicated, in which case we might need to use **nested** counterfactuals. For example,

$$Y_{0M_0}$$

This value is the outcome that we would observe if the exposure was set to zero, and the mediator was set to the value that would be observed if the exposure was set to zero. This more complicated quantity is a nested counterfactual, because the counterfactual mediator (M_0) is nested into the counterfactual outcome Y_{0M_0}.

With these nested counterfactuals, we can now rely on a review article by Hafeman and Schwartz (2009) to define the total and pure direct/indirect effects:

Table 2 Natural indirect and direct effects

Natural effect	Potential outcomes
Pure direct effect (PDE)	$\mathrm{P}(Y_{1M_0}=1)-\mathrm{P}(Y_{0M_0}=1)$
Pure indirect effect (PIE)	$\mathrm{P}(Y_{0M_1}=1)-\mathrm{P}(Y_{0M_0}=1)$
Total direct effect (TDE)	$\mathrm{P}(Y_{1M_1}=1)-\mathrm{P}(Y_{0M_1}=1)$
Total indirect effect (TIE)	$\mathrm{P}(Y_{1M_1}=1)-\mathrm{P}(Y_{1M_0}=1)$

Mediational effects can be defined in terms of potential outcomes.

Figure 2: Table 2 from Hafeman and Schwartz 2009.

There are a couple of important takeaways from this Table that are informative. The first is to note the key difference between a pure direct/indirect and a total direct/indirect effect. Let's focus first on the difference between the pure direct and total direct effects.

The pure direct effect is defined as a comparison of: Y_{1M_0} versus Y_{0M_0}. This is a contrast of the outcome that would be observed if an individual was **exposed**, but if there mediator value was set to what it would be if they were **unexposed** (i.e., Y_{1M_0}), relative to the outcome that would be observed if an

individual was **unexposed**, and if their mediator value was set to what would be observed if they were **unexposed** (i.e., Y_{0M_0}).

In contrast, the total direct effect is defined in a similar way for the exposure. But in both cases, the mediator value is what it would have been had they been **exposed**, instead of unexposed as in the pure direct case.

A similar situation arises if we were to focus on pure indirect versus total indirect effects (see Table).

The one key takeaway in the distinction between pure direct/indirect and total direct/indirect is that they are not necessarily equal, and this fact arises from the value to which we set the constant portion of the contrast. For example, the distinction between pure direct and total direct effects can be summed up in the following question: Is the effect of the exposure different if we set the mediator to M_0 versus M_1?

This takeaway is important because it was the first time that the concept of "moderation" was formalized in the literature. Moderation is a tricky concept to pin down, and as a result the analytic methods we should use to evaluate moderation were, for a long time, *ad hoc*. This distinction between pure direct/indirect and total direct/indirect help set the stage for later work that made the concept of moderation more precise.

A second curious but arguably less important takeaway comes from the fact that nested counterfactuals imply something strange about the world. In particular, they require that we conceptualize an outcome under two incompatible states. For example, the outcome if the exposure X were set to 1, under the mediator value that would be observed if the exposure X were set to zero.

Since there is no way to observe such an outcome ever (even in a perfectly designed randomized trials), these concepts of pure direct/indirect and total direct/indirect effects led to some consternation about whether they were scientifically viable metrics.

1.4 Modern Mediation: Natural and Controlled Effects

Years after Robins and Greenland (1992), Pearl (2001) grouped pure direct/indirect and total direct/indirect effects into a single class of effects called natural effects. He also proposed a new effect which was named the **controlled direct effect**, and which is defined as:

$$E(Y_{1m} - Y_{0m})$$

This controlled direct effect is basically the effect of the exposure if we were to hold the mediator fixed at a specific value for everyone in the population. Note that this is different from natural effects, because in this case we hold the mediator fixed to what it would have been had everyone been exposed/unexposed, which may not be constant for all individuals in the population.

Controlled direct effects were easier to define, easier to estimate, and easier to ground scientifically, since we could, in fact, design a randomized trial that allows us to estimate them (Pearl, 2001). However, one drawback of controlled direct effects is that there is no easy way to generate a corresponding controlled "indirect" effect (VanderWeele, 2011). Since this is often of scientific interest, researchers have often treaded the border between natural and controlled effects.

2 Classical Mediation Approaches

With some important context and background behind us, let's now explore the classical mediation methods that became very popular, but that are characterized by some important problems we need to understand. These problems are particularly relevant for health disparities work, as we will see.

To contextualize our discussion, let's assume we're interested in the racial disparity in greater than median weight change, and whether this disparity is explained by different rates of quitting smoking across self-reported racial status [e.g., "whites" (i.e., race = 0) versus among "black or other" (i.e., race = 1), per the NHEFS codebook].

Let's start by loading the data:

```
pacman::p_load(tidyverse,
               here,
               broom,
               boot)

nhefs <- read_csv(here("data","analytic_data.csv"))
```

```
nhefs <- nhefs %>%
  mutate(wt_delta = as.numeric(wt82_71 > median(wt82_71)))

names(nhefs)
```

```
##  [1] "seqn"     "qsmk"     "sex"      "age"      "race"     "income"
##  [7] "wt82_71"  "death"    "map"      "wt_delta"
```

```
summary(lm(qsmk ~ race, data = nhefs))$coefficients
```

```
##               Estimate Std. Error   t value     Pr(>|t|)
## (Intercept)  0.2581649 0.01196899 21.569471 6.675870e-90
## race        -0.1002701 0.03335984 -3.005713 2.694359e-03
```

```
summary(lm(wt_delta ~ race, data = nhefs))$coefficients
```

```
##               Estimate Std. Error   t value      Pr(>|t|)
## (Intercept)  0.5062208 0.01394493 36.301427 1.114188e-206
## race        -0.0483261 0.03886715 -1.243366  2.139306e-01
```

The mediation question we are going to try to answer is: if everybody in the population quit smoking (i.e., `qsmk` was set to zero), would there be a change in the racial disparity in greater than median weight change?

We'll start by using classical methods, and then we'll discuss problems with these approaches in the context of our NHEFS data and question.

2.1 The Difference Method

We'll start with what is known as the difference method, which is fairly simple. We can compute "direct" and "indirect" effects with this method in two stages. The first stage regresses the outcome (in our case, greater than median weight change) against the exposure (in our case, race). Different types of models have been used to do this in the past. We'll keep things simple and use linear regression:

$$E(Y \mid X) = \alpha_0 + \alpha_1 X$$

In this first stage model, α_1 might be interpreted as the magnitude of the disparity between race and greater than median weight change.

Next, we'll deploy the second stage, which adds the potential mediator to the model:

$$E(Y \mid X, M) = \beta_0 + \beta_1 X + \beta_2 M$$

In this second stage model, β_1 represents the association between race (X) and greater than median weight change (Y) that is not "due to" quitting smoking. In other words, if everyone were to quit smoking, the risk difference between race and greater than median weight change would be β_1. Furthermore, according to the logic underlying these classical methods, to compute the magnitude of the disparity that occurs through quitting smoking, we can simply take the difference between α_1 and β_1. Thus, with this difference method, the "direct" effect is β_1, and the "indirect" effect is $\alpha_1 - \beta_1$.

In R, we could do this with our data as follows:

```
alpha1 <- summary(lm(wt_delta ~ race, data = nhefs))$coefficients["race",]

alpha1
```

```
##    Estimate  Std. Error     t value    Pr(>|t|)
## -0.04832610  0.03886715 -1.24336623  0.21393063
```

```
beta1 <- summary(lm(wt_delta ~ race + qsmk, data = nhefs))$coefficients["race",]

beta1
```

```
##    Estimate  Std. Error     t value    Pr(>|t|)
## -0.03606788  0.03878402 -0.92996742  0.35254033
```

The direct effect estimate is thus -0.036, and the indirect effect estimate is -0.012.

This means that, according to this difference method, if everybody quit smoking, a greater than median weight change among those labeled "black or other" would be 3.6% lower than among those labeled "white".

Similarly, the "indirect" portion of the racial disparity due to quitting smoking is -1.2 cases of greater than median weight change per 100 individuals in the sample.

This method is trivially simple to implement. It is fairly straightforward to obtain standard errors for the direct effect, and one can easily compute standard errors for the indirect effect (or use the bootstrap). However, this method has several problems, which we'll discuss.

2.2 The Product Method

Before discussing problems with the difference method, let's review another common classical technique known as the product method. This product method is the approach articulated by Baron and Kenny (1986), and is equally easy to use. It is again a two stage approach, with the first stage equal to the `beta1` model in the difference method.

Stage 1:

$$E(Y \mid X, M) = \beta_0 + \beta_1 X + \beta_2 M$$

However, the second stage of the product method approach is to fit a regression model for the mediator, in our case, quitting smoking:

$$E(M \mid X) = \gamma_0 + \gamma_1 X$$

One can then (again) use the β_1 parameter as an estimate of the "direct" effect. The "indirect" effect can be computed as the product of γ_1 and β_2. In R, this might look like:

```
beta1 <- summary(lm(wt_delta ~ race + qsmk, data = nhefs))$coefficients["race",]


beta1
```

```
##    Estimate  Std. Error     t value    Pr(>|t|)
## -0.03606788  0.03878402 -0.92996742  0.35254033
```

```
beta2 <- summary(lm(wt_delta ~ race + qsmk, data = nhefs))$coefficients["qsmk",]


beta2
```

```
##     Estimate   Std. Error      t value     Pr(>|t|)
## 1.222520e-01 3.018931e-02 4.049514e+00 5.399485e-05
```

```
gamma1  <- summary(lm(qsmk ~ race, data = nhefs))$coefficients["race",]

gamma1
```

```
##     Estimate   Std. Error      t value     Pr(>|t|)
## -0.100270115  0.033359843 -3.005713060  0.002694359
```

Which yields a direct effect estimate of (again) -0.036, and an indirect effect estimate of -0.012.

Again, it's straightforward to get standard errors for these estimates. Let's now talk about some of the problems with these approaches.

3 Problems with Classical Approaches

These methods were commonly used in sociology, epidemiology, psychology, and the health sciences. These methods are still being used, actually, but more and more researchers are relying on more generally applicable methods.

Why? There are at least four problems with the above approaches.

3.1 Problem 1: Mediator-Outcome Confounding

Control for confounders of the mediator outcome relation is required, even in settings where the exposure is randomized. This need to control for mediator-outcome confounders is often overlooked. Interestingly, it was raised in an early article on the product method by Judd and Kenny (1981), but Baron and Kenny (1986) neglected to mention this issue. Without adjusting for mediator-outcome confounders, we get a biased estimate of the parameter for the relation between the mediator and the outcome. This biases both the direct and indirect effect estimates that arise from these approaches.

For example, several studies (**??**Hernandez-Diaz et al., 2006) have examined the effect of smoking on infant mortality adjusting for gestational age as a mediator. These analyses have led to the identification of a "paradox". After adjusting for a presumed mediator of the relation between smoking status

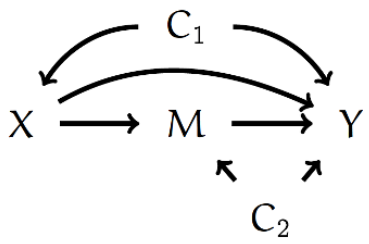

Figure 3: When doing mediation analysis, both exposure-outcome and mediator-outcome (and sometimes exposure-mediator) confounding.

and infant mortality, fetal/gestational weight, they find smoking is protective for infant mortality among babies weighing <2,000g (OR = 0.79, 95% CI: 0.76, 0.82).

Hernandez-Diaz et al. (2006) argued convincingly that this is the result of unmeasured mediator-outcome confounding.

3.2 Problem 2: Mediator-Outcome Confounding Affected by the Exposure

Another problem is that neither of these methods can handle the scenario where there is an association between the exposure, and confounders of the mediator outcome association. This scenario is illustrated in Figure 4:

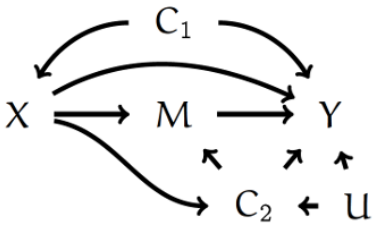

Figure 4: Illustration of mediator-outcome confounders affected by the exposure.

This is a particularly important problem in health disparities research. Race, for example, is an "upstream" or "fundamental cause" variable (Link and Phelan, 1995, Phelan et al. (2010)). Consequently, we'd expect a variable like race to be associated with a whole host of downstream variables, including many that would confound the association between an mediator and outcome of interest.

3.3 Problem 3: Exposure-Mediator Interactions

A third problem is that these approaches (particularly the difference method) presuppose no exposure-mediator interaction.

For the difference method, it's not clear what to do with the exposure-mediator interaction term if present.

For the product method, Valeri and Vanderweele (2013) have generalized the product method to accommodate such interactions, but the approach still cannot resolve problem 2.

3.4 Problem 4: Non-Collapsibility

In the above, we used linear models, but these techniques are often applied irrespective of model form. It is common to see these methods applied to log-linear, logistic, and even time-to-event (e.g., Cox proportional hazards) models.

4 When they Work

So, overall, what are the assumptions or conditions that need to hold for these two methods to work? Here is a list:

- No exposure-mediator interaction
- No mediator-outcome confounders affected by (or associated with) the exposure
- Exposure-outcome and mediator-outcome confounders accounted for in both models
- Model is collapsible

If any one of these conditions does not hold (e.g., is violated), then we should not really trust the "direct" and "indirect" effect estimates obtained from the difference and product methods above.

References

Cande V Ananth and Tyler J VanderWeele. Placental abruption and perinatal mortality with preterm delivery as a mediator: disentangling direct and indirect effects. *Am J Epidemiol*, 174(1):99–108, 2011.

R M Baron and D A Kenny. The moderator-mediator variable distinction in social psychological research: conceptual, strategic, and statistical considerations. *J Pers Soc Psychol*, 51(6):1173–1182, 1986.

Ranee Chatterjee, Hsin-Chieh Yeh, Tariq Shafi, Cheryl Anderson, James S Pankow, Edgar R Miller, David Levine, Elizabeth Selvin, and Frederick L Brancati. Serum potassium and the racial disparity in diabetes risk: the atherosclerosis risk in communities (aric) study. *Am J Clin Nutr*, 93(5): 1087–1091, 2011.

Danella M Hafeman and Sharon Schwartz. Opening the black box: a motivation for the assessment of mediation. *Int J Epidemiol*, 38(3):838–845, Jun 2009.

Sonia Hernandez-Diaz, Enrique F Schisterman, and Miguel A Hernan. The birth weight "paradox" uncovered? *Am J Epidemiol*, 164(11):1115–1120, Dec 2006. DOI: 10.1093/aje/kwj275.

H.H. Hyman. *Survey Design and Analysis: Principles, Cases, and Procedures*. Free Press, 1955. URL https://books.google.com/books?id=8ygYAAAAIAAJ.

Charles M. Judd and David A. Kenny. Process analysis: Estimating mediation in treatment evaluations. *Eval Rev*, 5(5):602–619, 1981.

B G Link and J Phelan. Social conditions as fundamental causes of disease. *J Health Soc Behav*, 35((Extra Issue)):80–94, 1995.

Pauline Mendola, Sunni L Mumford, Tuija I Mannisto, Alexander Holston, Uma M Reddy, and S Katherine Laughon. Controlled direct effects of preeclampsia on neonatal health after accounting for mediation by preterm birth. *Epidemiology*, 26(1):17–26, 2015.

J. Pearl. Direct and Indirect Effects. In John Breese and Daphne Koller, editors, *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence*, pages 411–20. Morgan Kaufmann, San Francisco, CA, 2001.

Jo C Phelan, Bruce G Link, and Parisa Tehranifar. Social conditions as fundamental causes of health inequalities: theory, evidence, and policy implications. *J Health Soc Behav*, 51 Suppl(1):S28–40, 2010.

James M. Robins and Sander Greenland. Identifiability and exchangeability for direct and indirect effects. *Epidemiol*, 3(2):143–155, 1992.

Linda Valeri and Tyler J Vanderweele. Mediation analysis allowing for exposure-mediator interactions and causal interpretation: theoretical assumptions and implementation with SAS and SPSS macros. *Psychol Methods*, 18(2):137–150, Jun 2013. DOI: 10.1037/a0031034.

Tyler J VanderWeele. Controlled direct and mediated effects: definition, identification and bounds. *Scandinavian Journal of Statistics*, 38(3):551–563, 2011.

Tyler J VanderWeele, Kofi Asomaning, Eric J Tchetgen Tchetgen, Younghun Han, Margaret R Spitz, Sanjay Shete, Xifeng Wu, Valerie Gaborieau, Ying Wang, John McLaughlin, Rayjean J Hung, Paul Brennan, Christopher I Amos, David C Christiani, and Xihong Lin. Genetic variants on 15q25.1, smoking, and lung cancer: an assessment of mediation and interaction. *Am J Epidemiol*, 175(10):1013–1020, May 2012. ISSN 1476-6256 (Electronic); 0002-9262 (Print); 0002-9262 (Linking). DOI: 10.1093/aje/kwr467.

Sewall Wright. The method of path coefficients. *The Annals of Mathematical Statistics*, 5(3):161–215, 1934.