Mediation analysis: stratification and standardization

Boris Sobolev, Lisa Kuramoto *The University of British Columbia*

The paper presents explicit formulas for estimating effects in mediation analysis using stratification and standardization.

Reasoning

The causal analysis is about the comparability of treatment groups. The control group and the treatment group should be as similar as possible so that we can be sure that any effects are due to the treatment alone. We compare treatment responses in carefully defined strata to make sure that we are comparing like with like. The only difference is whether or not the units in the strata received the treatment.

For stratification, we consider pre-treatment factors that influence both treatment choice and outcome. Causal mediation analysis extends this approach beyond pre-treatment factors, which can create co-variation between treatment and outcome.

Post-treatment factors can also influence outcomes. When factors change in response to treatment and then, in turn, affect the outcome, we call them mediators. Mediators are intermediary outcomes. They lie on the causal pathway between the treatment and the outcome of primary interest.

Here is the key assumption – the total treatment effect describes changes in outcome in response to changes in treatment, directly and through mediators. Therefore, total, direct, and indirect effects are the subject of causal mediation analysis.

Take direct effect as an example. This is the expected outcome after the application of a treatment, when the mediating factors are considered at the value they would have taken in the absence of treatment. Here we contrast two counterfactual situations: all units without treatment and all units with treatment, but the units in both groups have mediator values naturally occurring in the absence of treatment.

In the case of indirect effect, our goal is to find the fraction of outcomes that would be sustained solely by the presence of mediator absent any other effect the treatment may have on the outcome.

Confounding-free mediation

Treatment can change the marginal mean outcome by changing the outcome probability in the groups determined by mediator values, the composition of mediators in the population, or both. Therefore, the difference in marginal means can be due to differences in group-specific risks and differences in composition.

Table 1: *Frequencies of binary T, M and Y*

	T	M	Y	$E\{Y \mid T{=}t, M{=}m\}$	$P(M \mid T{=}t)$
n_1	0	0	0	$g_{00}{=}\dfrac{n_2}{n_1+n_2}$	$h_0{=}\dfrac{n_3+n_4}{\sum_{i=1}^{4} n_i}$
n_2	0	0	1		
n_3	0	1	0	$g_{01}{=}\dfrac{n_4}{n_3+n_4}$	
n_4	0	1	1		
n_5	1	0	0	$g_{10}{=}\dfrac{n_6}{n_5+n_6}$	$h_1{=}\dfrac{n_7+n_8}{\sum_{i=5}^{8} n_i}$
n_6	1	0	1		
n_7	1	1	0	$g_{11}{=}\dfrac{n_8}{n_7+n_8}$	
n_8	1	1	1		

In the case of un-confounded mediation, we find the total effect of treatment in two steps. As defined in Table 1, let $g_{tm} = E\{Y \mid t, m\}$ and $h_t = P(M{=}\,m \mid t)$ for binary treatment T, mediator M and outcome Y. We average mean outcomes in the groups with and without treatment by the proportion of mediators in each group and take the difference between the marginal means

$$\text{TE} = g_{11}h_1 + g_{10}(1-h_1) - \big(g_{01}h_0 + g_{00}(1-h_0)\big).$$

For the direct effect, we follow a different set of steps. We take the difference in mean outcome between the treated and untreated groups and then average it over the proportion of mediators in the untreated group

$$\text{DE} = (1-h_0)(g_{10}-g_{00}) + h_0(g_{11}-g_{01}).$$

Why exactly in the untreated group? It standardizes the number of outcomes in the treated group by the composition of mediator levels in the untreated group. This eliminates the composition effect.

Standardization

Basic concepts

Let us review the basic concepts of standardization of the mean outcomes of the study population by the composition of the standard population. In our case, the treated units are the study population and the untreated units are the standard population. The mediator values determine strata in both groups.

Direct standardization allows us to find the marginal mean for the treated population if its mediator composition were the same as in the untreated population. Let $E\{Y_{1M_0}\}$ denote the expectation of the intervention distribution of outcome Y if all units had treatment 1 and their mediators M took values that would naturally occur under treatment 0.[1] We calculate it by averaging the treated mean outcomes over the mediator composition in the untreated units:

$$E\{Y_{1M_0}\} = \sum_m E\{Y \mid T{=}1, M{=}m\}P(M{=}m \mid T{=}0).$$

Indirect standardization determines the marginal mean for the treated population if the mean outcomes in its strata were the same as in the untreated population. Let $E\{Y_{0M_1}\}$ denote the expectation of the intervention probability of outcome Y if all units had treatment 0 and their mediators took values that would naturally occur with treatment 1. In this approach, we average the untreated mean outcomes over the mediator composition in the treated units:

$$E\{Y_{0M_1}\} = \sum_m E\{Y \mid T{=}0, M{=}m\}P(M{=}m \mid T{=}1).$$

Both directly and indirectly standardized means are then compared to the marginal means in the populations studied. However, these comparisons may target different populations.[2]

Applied to mediation analysis

By comparing the standardized mean $E\{Y_{1M_0}\}$ with the marginal mean for the untreated group $\mathbf{E}^0$, we measure the effect of changes in mean outcomes in mediator strata

$$\begin{aligned}\delta_0 &\stackrel{\text{def}}{=} E\{Y_{1M_0}\} - \mathbf{E}^0 \\ &= \sum_m P(M{=}\,m \mid T{=}\,0)\Big(E\{Y \mid T{=}\,1, M{=}\,m\} \\ &\qquad - E\{Y \mid T{=}\,0, M{=}\,m\}\Big).\end{aligned}$$

By comparing it with the marginal mean for the treated group, we measure the effect of a reverse change in mediator composition, i.e., from the treated to untreated,

$$\begin{aligned}\zeta_1 &\stackrel{\text{def}}{=} E\{Y_{1M_0}\} - \mathbf{E}^1 \\ &= \sum_m E\{Y \mid T{=}\,1, M{=}\,m\}\Big(P(M{=}\,m \mid T{=}\,0) \\ &\qquad - P(M{=}\,m \mid T{=}\,1)\Big).\end{aligned}$$

By comparing the standardized mean $E\{Y_{0M_1}\}$ with the marginal mean for the treated group, we

measure the effect of changes in mean outcomes in mediator strata

$$\begin{aligned}\delta_1 &\stackrel{\text{def}}{=} E\{Y_{0M_1}\} - \mathbf{E}^1 \\ &= \sum_m P(M=m \mid T=1)\Big(E\{Y \mid T=0, M=m\} \\ &\quad - E\{Y \mid T=1, M=m\}\Big).\end{aligned}$$

By comparing it with the marginal mean in the untreated group, we measure the effect of changes in mediator composition:

$$\begin{aligned}\zeta_0 &\stackrel{\text{def}}{=} E\{Y_{0M_1}\} - \mathbf{E}^0 \\ &= \sum_m E\{Y \mid T=0, M=m\}\Big(P(M=m \mid T=1) \\ &\quad - P(M=m \mid T=0)\Big).\end{aligned}$$

The differences

$$\begin{aligned}\delta_x &= E\{Y_{1-xM_x}\} - \mathbf{E}^x,\ x=0,1 \\ \zeta_x &= E\{Y_{xM_{1-x}}\} - \mathbf{E}^x,\ x=0,1\end{aligned}$$

are called the direct and indirect effect, correspondingly. Their index refers to the target population of the comparison.

Counterfactual interpretation

We interpret direct effects as counterfactuals: changes in the marginal means that would occur without the mediator responding to changes in treatment. For example, the natural direct effect

$$\delta_0 = E\{Y_{1M_0}\} - \mathbf{E}^0$$

is interpreted as the difference in marginal means between groups of treated and untreated units if both groups had taken the mediator values that would naturally occur without treatment.[1]

Similarly, we interpret indirect effects as changes in the marginal mean that would occur if only the mediator responded to treatment. For example, the natural indirect effect

$$\zeta_0 = E\{Y_{0M_1}\} - \mathbf{E}^0$$

is interpreted as the difference in marginal means between two groups of untreated units if one group had taken the mediator values that would naturally occur with treatment.

Total effect partitioning

We use direct effects δ_x and indirect effects ζ_x to partition the total effect of treatment. The general form of the partitioning is given by

$$\text{TE} = (1-2t)(\delta_t - \zeta_{1-t});\ t=0,1.$$

For example, the total effect of changing treatment from $T=0$ to $T=1$ is equal to the effect of change in mediator composition in untreated units minus the effect of change in mean outcomes in treated units.

$$\text{TE} = \zeta_0 - \delta_1.$$

In other words, the total treatment effect is the indirect effect in the untreated reduced by the direct effect in the treated. Alternatively, the total effect of change in treatment is equal to the effect of change in mean outcomes in untreated units minus the effect of change in mediator composition in treated units.

$$\text{TE} = \delta_0 - \zeta_1.$$

In other words, the total treatment effect is the direct effect in the untreated reduced by the indirect effect in the treated. The second term is difference in marginal means between two groups of treated units if one group had not taken mediator values that would naturally occur in the treated units. Remarkably, this implies that the negative difference between the total effect and the natural direct effect identifies the effect of a reverse change in mediator composition:

$$-\big(\text{TE} - \delta_0\big) = E\{Y_{1M_0}\} - \mathbf{E}^1.$$

Mediation analysis through stratification

How do we get the marginal mean of outcome if the treatment had no effect on the mediator, but the outcome would depend on the mediator value? Suppose that the mediator variable encodes the presence and absence of a mediator as 1 and 0. We take the proportion of outcomes in treated units, with and without mediator, in each stratum defined by the adjustment set. We then take the proportion of untreated units, with and without mediator. We then apply the proportions of strata in the study population.

Let's use the following notation to define all three proportions:

n^1_{sgm}: the number of events in group g, stratum s, mediator m
n_{sgm}: the number of patients in group g, stratum s, mediator m
n_{sg}: the number of patients in group g, stratum s

Stratification and standardization

We multiply these three proportions and sum up the products from each stratum to obtain the directly standardized mean

$$E\{Y_{1M_0}\} = \sum_s \frac{n^1_{s10}}{n_{s10}} \frac{n_{s00}}{n_{s0}} \frac{n_s}{N} + \sum_s \frac{n^1_{s11}}{n_{s11}} \frac{n_{s01}}{n_{s0}} \frac{n_s}{N} \quad (1)$$

To see the relationship of eq.1 to direct standardization, let's re-arrange the formula to make it look like the inverse probability weighting. We take the inverse of the untreated proportion outside of the brackets in each stratum

$$E\{Y_{1M_0}\} = N^{-1} \sum_s \left(\frac{n^1_{s10}}{n_{s10}} n_{s00} + \frac{n^1_{s11}}{n_{s11}} n_{s01} \right) \frac{n_s}{n_{s0}}$$

What's left in brackets? It's the expected number of events in the treated group, which we standardized on the proportion of mediated in the untreated group. Let's denote it as $\tilde{n}^1_{s1}$, then

$$E\{Y_{1M_0}\} = N^{-1} \sum_s \tilde{n}^1_{s1} \times \frac{n_s}{n_{s0}} \quad (2)$$

We use the term *standardized* in an epidemiological sense.[3] For direct standardization, we calculate the strata-specific rate in one population and apply it to a population with standard strata composition. Here, we use the composition of mediator levels in the untreated, as a standard population. We then stratify it by the adjustment covariate. And then we apply mean outcomes from the corresponding strata of the treated group.

As a result, we get the expected number of outcomes in the treated group, standardized by the composition of mediator levels in the untreated group. Weighted by the inverse of the prevalence of untreated in each stratum and averaged over the entire study population, it gives us the counterfactual risk of outcome among the treated units in the presence of mediators typical in the untreated. Sounds like classical epidemiology.

Strata-specific direct effects

To find formulas for computing the direct effect, we subtract the marginal risk in the untreated population from its standardized counterpart for the treated

$$\begin{aligned} \text{NDE} &= E\{Y_{1M_0}\} - \mathbf{P}^0 \\ &= \sum_s \left(\frac{n^1_{s10}}{n_{s10}} \frac{n_{s00}}{n_{s0}} + \frac{n^1_{s11}}{n_{s11}} \frac{n_{s01}}{n_{s0}} - \frac{n^1_{s0}}{n_{s0}} \right) \frac{n_s}{N} \end{aligned}$$

In the individual stratum, the direct effect is given by

$$\begin{aligned} \text{SRD} &= \frac{n^1_{s10}}{n_{s10}} \frac{n_{s00}}{n_{s0}} + \frac{n^1_{s11}}{n_{s11}} \frac{n_{s01}}{n_{s0}} - \frac{n^1_{s0}}{n_{s0}} \\ &= \frac{\tilde{n}^1_{s1} - n^1_{s0}}{n_{s0}}. \end{aligned}$$

Strata-specific indirect effects

To find formulas for computing the indirect effect, we subtract the marginal risk from its standardized counterpart in the untreated population

$$\begin{aligned} \text{NIE} &= E\{Y_{0M_1}\} - \mathbf{P}^0 \\ &= \sum_s \left(\frac{n^1_{s00}}{n_{s00}} \frac{n_{s10}}{n_{s1}} + \frac{n^1_{s01}}{n_{s01}} \frac{n_{s11}}{n_{s1}} - \frac{n^1_{s0}}{n_{s0}} \right) \frac{n_s}{N} \end{aligned}$$

In this case the indirect effect in the individual stratum will be given by

$$\begin{aligned} \text{SRD} &= \frac{n^1_{s00}}{n_{s00}} \frac{n_{s10}}{n_{s1}} + \frac{n^1_{s01}}{n_{s01}} \frac{n_{s11}}{n_{s1}} - \frac{n^1_{s0}}{n_{s0}} \\ &= \frac{\tilde{n}^1_{s0}}{n_{s1}} - \frac{n^1_{s0}}{n_{s0}} \end{aligned}$$

where $\tilde{n}^1_{s0}$ is the expected number of events in the untreated group standardized on the proportion of mediated in the treated group.

Numerical example

Timing of treatment as mediator

Let's consider an observational study comparing effectiveness of stenting ($T = 1$) versus surgery ($T = 0$) in preventing the need for future revascularization, Y. In this setting, the method of treatment may change the timing of treatment, M. Our research question asks

Table 2: *Outcome in relation to treatment and timing*

S	T	M	$E(Y \mid T, M, S)$	$P(M{=}m \mid T, S)$
1	1	1	0.111	0.850
1	1	2	0.077	0.096
1	1	3	0.128	0.054
1	0	1	0.051	0.663
1	0	2	0.012	0.187
1	0	3	0.023	0.151
2	1	1	0.160	0.629
2	1	2	0.091	0.081
2	1	3	0.143	0.289
2	0	1	0.039	0.348
2	0	2	0.034	0.162
2	0	3	0.030	0.490

to what extent the timing of treatment contributes to the difference in outcome probability. The policy implications of this question are either to shorten the time to surgery or to encourage stenting instead of waiting for surgery.

With stratification

Table 2 shows the aggregate data collected in two strata determined by the treatment priority S, which may confound all 3 relations: $T \to Y$, $T \to M$, and $M \to Y$. Because the priority is not affected by treatment, adjusting for S renders all relations unconfounded in each stratum, and the direct and indirect effects are the average of stratum-specific effects over the strata distribution. The following adjustment formulas were used to compute the effects[4]

$$\begin{aligned}\text{NDE} = \sum_{s,m} & P(s)P(M = m \mid T = 0, S = s) \\ & \times \big(E\{Y \mid T = 1, M = m, S = s\} \\ & - E\{Y \mid T = 0, M = m, S = s\}\big)\end{aligned}$$

$$\begin{aligned}\text{NIE} = \sum_{s,m} & P(s)E\{Y \mid T = 0, M = m, S = s\} \\ & \times \big(P(M = m \mid T = 1, S = s) \\ & - P(M = m \mid T = 0, S = s)\big).\end{aligned}$$

Using the figures from the last column[1] and the stratum distribution, 0.37 and 0.63, we obtain

$$\begin{aligned}\text{TE} &= E\{Y_{1M_1}\} - E\{Y_{0M_0}\} = 10.3\% \\ \text{NDE} &= E\{Y_{1M_0}\} - E\{Y_{0M_0}\} = 9.7\% \\ \text{NIE} &= E\{Y_{0M_1}\} - E\{Y_{0M_0}\} = 3.1\%\end{aligned}$$

We can conclude that stenting has increased the risk of revascularization by 10% and that up to one-third of the difference, NIE/TE = 0.3, is explained by the treatment timing alone without the effect of treatment. At the same time, only small portion of this difference, $1 - \text{NDE/TE} = 0.06$, is owed to the capacity of the treatment to affect the timing of treatment by the natural occurrence.

[1] It's the two-model numbers actually

References

[1] Pearl J. Interpretation and identification of causal mediation. Psychological methods. 2014;19(4):459–481.

[2] Miettinen OS. Components of the crude risk ratio. American Journal of Epidemiology. 1972;96(2):168–172.

[3] Keiding N, Clayton D. Standardization and Control for Confounding in Observational Studies: A Historical Perspective. Statistical Science. 2014;29(4):529 – 558.

[4] Shpitser I, VanderWeele TJ. A complete graphical criterion for the adjustment formula in mediation analysis. The international Journal of Biostatistics. 2011;7(1):1–16.