## ETT: $E[Y_x / X = x']$

- Recruitment
- Additive interventions

## Probability of Necessity

$PN = P(Y_0 = 0 / X = 1, Y = 1)$

- **Attribution**
  - Personal decision problems
  - Possible cases of discrimination
- **Legal Liability**
  - $P(Damage_{Action=0} = 0 \,/\, Damage = 1, Ac = 1)$

## Mediation

$E[Y_{x, M_{x'}}]$

"The expected outcome (Y) had the treatment been $X = x$ and, simultaneously, had the mediator M attained the value ($M_{x'}$) it would have attained had X been $x'$."

- Indirect effect of gender on hiring mediated by qualification.

# Attribution in Legal Setting

**Example 4.5.1 (Attribution in Legal Setting)** *A lawsuit is filed against the manufacturer of drug x, charging that the drug is likely to have caused the death of Mr A, who took it to relieve back pains. The manufacturer claims that experimental data on patients with back pains show conclusively that drug x has only minor effects on death rates. However, the plaintiff argues that the experimental study is of little relevance to this case because it represents average effects on patients in the study, not on patients like Mr A who did not participate in the study. In particular, argues the plaintiff, Mr A is unique in that he used the drug of his own volition, unlike subjects in the experimental study, who took the drug to comply with experimental protocols. To support this argument, the plaintiff furnishes nonexperimental data on patients who, like Mr A, chose drug x to relieve back pains but were not part of any experiment, and who experienced higher death rates than those who didn't take the drug. The court must now decide, based on both the experimental and nonexperimental studies, whether it is "more probable than not" that drug x was in fact the cause of Mr A's death.*

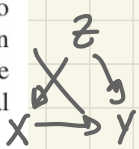
**Theorem 4.5.1** *If Y is monotonic relative to X, that is, $Y_1(u) \geq Y_0(u)$ for all u, then PN is identifiable whenever the causal effect $P(y|do(x))$ is identifiable, and*

$$PN = \frac{P(y) - P(y|do(x'))}{P(x,y)} \tag{4.28}$$

*or, substituting $P(y) = P(y|x)P(x) + P(y|x')(1 - P(x))$, we obtain*

$$PN = \frac{P(y|x) - P(y|x')}{P(y|x)} + \frac{P(y|x') - P(y|do(x'))}{P(x,y)} \tag{4.29}$$

The first term on the r.h.s. of (4.29) is called the *excess risk ratio* (ERR) and is often used in court cases in the absence of experimental data (Greenland 1999). It is also known as the Attributable Risk Fraction among the exposed (Jewell 2004, Chapter 4.7). The second term (the *confounding factor* (CF)) represents a *correction* needed to account for confounding bias, that is, $P(y|do(x')) \neq P(y|x')$. Put in words, confounding occurs when the proportion of population for whom $Y = y$, when $X$ is set to $x'$ for everyone is not the same as the proportion of the population for whom $Y = y$ among those acquiring $X = x'$ by choice. For instance, suppose there is a case brought against a car manufacturer, claiming that its car's faulty design led to a man's death in a car crash. The ERR tells us how much more likely people are to die in crashes when driving one of the manufacturer's cars. If it turns out that people who buy the manufacturer's cars are more likely to drive fast (leading to deadlier crashes) than the general populations, the second term will correct for that bias.




Equation (4.29) thus provides an estimable measure of necessary causation, which can be used for monotonic $Y_x(u)$ whenever the causal effect $P(y|do(x))$ can be estimated, be it from randomized trials or from graph-assisted observational studies (e.g., through the backdoor criterion). More significantly, it has also been shown (Tian and Pearl 2000) that the expression

Assuming that drug $x$ can only cause (but never prevent) death, monotonicity holds, and Theorem 4.5.1 (Eq. (4.29)) yields

$$PN = \frac{P(y|x) - P(y|x')}{P(y|x)} + \frac{P(y|x') - P(y|do(x'))}{P(x,y)}$$
$$= \frac{0.002 - 0.028}{0.002} + \frac{0.028 - 0.014}{0.001} = -13 + 14 = 1 \quad (4.41)$$

We see that while the observational ERR is negative (−13), giving the impression that the drug is actually preventing deaths, the bias-correction term (+14) rectifies this impression and sets the probability of necessity (PN) to unity. Moreover, since the lower bound of Eq. (4.30) becomes 1, we conclude that $PN = 1.00$ even without assuming monotonicity. Thus, the plaintiff was correct; barring sampling errors, the data provide us with 100% assurance that drug $x$ was in fact responsible for the death of Mr A.

To complete this tool kit for attribution, we note that the other two probabilities that came up in the discussion on personal decision-making (Example 4.4.3), PS and PNS, can be bounded by similar expressions; see (Pearl 2000, Chapter 9) and (Tian and Pearl 2000).

In particular, when $Y_x(u)$ is monotonic, we have

$$PNS = P(Y_x = 1, Y_{x'} = 0)$$
$$= P(Y_x = 1) - P(Y_{x'} = 1) \quad (4.42)$$

as asserted in Example 4.4.3, Eq. (4.26).

# A Toolkit for Mediation



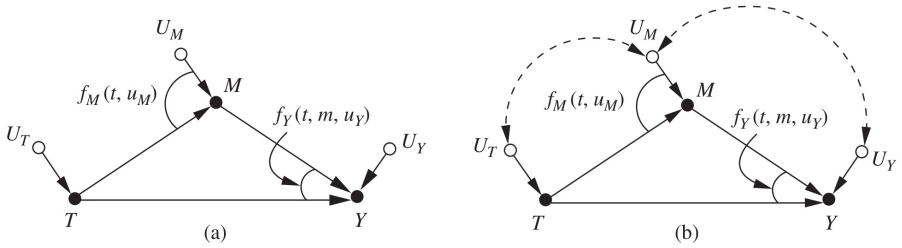



**Figure 4.6** (a) The basic nonparametric mediation model, with no confounding. (b) A confounded mediation model in which dependence exists between $U_M$ and $(U_T, U_Y)$

**Counterfactual definition of direct and indirect effects**

Using the structural model of Eq. (4.43) and the counterfactual notation defined in Section 4.2.1, four types of effects can be defined for the transition from $T = 0$ to $T = 1$. Generalizations to arbitrary reference points, say from $T = t$ to $T = t'$, are straightforward[1]:

**(a) Total effect –**

$$\begin{aligned} TE &= E[Y_1 - Y_0] \\ &= E[Y|do(T = 1)] - E[Y|do(T = 0)] \end{aligned} \tag{4.44}$$

*TE* measures the expected increase in $Y$ as the treatment changes from $T = 0$ to $T = 1$, while the mediator is allowed to track the change in $T$ naturally, as dictated by the function $f_M$.

**(b) Controlled direct effect –**

$$\begin{aligned} CDE(m) &= E[Y_{1,m} - Y_{0,m}] \\ &= E[Y|do(T = 1, M = m)] - E[Y|do(T = 0, M = m)] \end{aligned} \tag{4.45}$$

*CDE* measures the expected increase in $Y$ as the treatment changes from $T = 0$ to $T = 1$, while the mediator is set to a specified level $M = m$ uniformly over the entire population.

**(c) Natural direct effect –**

$$NDE = E[Y_{1,M_0} - Y_{0,M_0}] \tag{4.46}$$

*NDE* measures the expected increase in $Y$ as the treatment changes from $T = 0$ to $T = 1$, while the mediator is set to whatever value it *would have attained* (for each individual) prior to the change, that is, under $T = 0$.

**(d) Natural indirect effect –**

$$NIE = E[Y_{0,M_1} - Y_{0,M_0}] \tag{4.47}$$

*NIE* measures the expected increase in $Y$ when the treatment is held constant, at $T = 0$, and $M$ changes to whatever value it would have attained (for each individual) under $T = 1$. It captures, therefore, the portion of the effect that can be explained by mediation alone, while disabling the capacity of $Y$ to respond to $X$.

---

[1] These definitions apply at the population levels; the unit-level effects are given by the expressions under the expectation. All expectations are taken over the factors $U_M$ and $U_Y$.

We note that, in general, the total effect can be decomposed as

$$TE = NDE - NIE_r \tag{4.48}$$

where $NIE_r$ stands for the NIE under the reverse transition, from $T = 1$ to $T = 0$. This implies that *NIE* is identifiable whenever *NDE* and *TE* are identifiable. In linear systems, where reversal of transitions amounts to negating the signs of their effects, we have the standard additive formula, $TE = NDE + NIE$.

We further note that *TE* and *CDE*(*m*) are *do*-expressions and can, therefore, be estimated from experimental data or in observational studies using the backdoor or front-door adjustments. Not so for the *NDE* and *NIE*; a new set of assumptions is needed for their identification.

**Conditions for identifying natural effects**

The following set of conditions, marked *A*-1 to *A*-4, are sufficient for identifying both direct and indirect natural effects.

We can identify the *NDE* and *NIE* provided that there exists a set $W$ of measured covariates such that

A-1 No member of $W$ is a descendant of $T$.
A-2 $W$ blocks all backdoor paths from $M$ to $Y$ (after removing $T \to M$ and $T \to Y$).
A-3 The $W$-specific effect of $T$ on $M$ is identifiable (possibly using experiments or adjustments).
A-4 The $W$-specific joint effect of $\{T, M\}$ on $Y$ is identifiable (possibly using experiments or adjustments).

**Theorem 4.5.2 (Identification of the *NDE*)** *When conditions A-1 and A-2 hold, the natural direct effect is experimentally identifiable and is given by*

$$NDE = \sum_m \sum_w [E[Y|do(T = 1, M = m), W = w] - E[Y|do(T = 0, M = m), W = w]] \times P(M = m|do(T = 0), W = w)P(W = w) \tag{4.49}$$

*The identifiability of the do-expressions in Eq. (4.49) is guaranteed by conditions A-3 and A-4 and can be determined using the backdoor or front-door criteria.*

**Corollary 4.5.1** *If conditions A-1 and A-2 are satisfied by a set W that also deconfounds the relationships in A-3 and A-4, then the do-expressions in Eq. (4.49) are reducible to conditional expectations, and the natural direct effect becomes*

$$NDE = \sum_m \sum_w [E[Y|T = 1, M = m, W = w] - E[Y|T = 0, M = m, W = w]] P(M = m|T = 0, W = w)P(W = w) \tag{4.50}$$

In the nonconfounding case (Figure 4.6(a)), *NDE* reduces to

$$NDE = \sum_m [E[Y \mid T = 1, M = m] - E[Y \mid T = 0, M = m]]P(M = m \mid T = 0). \tag{4.51}$$

Similarly, using (4.48), the NIE becomes

$$NIE = \sum_m E[Y \mid T = 0, M = m][P(M = m \mid T = 1) - P(M = m \mid T = 0)] \tag{4.52}$$

The last two expressions are known as the *mediation formulas*. We see that while *NDE* is a weighted average of *CDE*, no such interpretation can be given to *NIE*.

The counterfactual definitions of *NDE* and *NIE* (Eqs. (4.46) and (4.47)) permit us to give these effects meaningful interpretations in terms of "response fractions." The ratio $NDE/TE$ measures the fraction of the response that is transmitted directly, with $M$ "frozen." $NIE/TE$ measures the fraction of the response that may be transmitted through $M$, with $Y$ blinded to $X$. Consequently, the difference $(TE - NDE)/TE$ measures the fraction of the response that is necessarily due to $M$.

**Numerical example: Mediation with binary variables**

To anchor these mediation formulas in a concrete example, we return to the encouragement-design example of Section 4.2.3 and assume that $T = 1$ stands for participation in an enhanced training program, $Y = 1$ for passing the exam, and $M = 1$ for a student spending more than 3 hours per week on homework. Assume further that the data described in Tables 4.6 and 4.7 were obtained in a randomized trial with no mediator-to-outcome confounding (Figure 4.6(a)). The data shows that training tends to increase both the time spent on homework and the rate of success on the exam. Moreover, training and time spent on homework together are more likely to produce success than each factor alone.

Our research question asks for the extent to which students' homework contributes to their increased success rates regardless of the training program. The policy implications of such questions lie in evaluating policy options that either curtail or enhance homework efforts, for example, by counting homework effort in the final grade or by providing students with



adequate work environments at home. An extreme explanation of the data, with significant impact on educational policy, might argue that the program does not contribute substantively to students' success, save for encouraging students to spend more time on homework, an encouragement that could be obtained through less expensive means. Opposing this theory, we may have teachers who argue that the program's success is substantive, achieved mainly due to the unique features of the curriculum covered, whereas the increase in homework efforts cannot alone account for the success observed.

Substituting the data into Eqs. (4.51) and (4.52) gives

$$NDE = (0.40 - 0.20)(1 - 0.40) + (0.80 - 0.30)0.40 = 0.32$$
$$NIE = (0.75 - 0.40)(0.30 - 0.20) = 0.035$$
$$TE = 0.80 \times 0.75 + 0.40 \times 0.25 - (0.30 \times 0.40 + 0.20 \times 0.60) = 0.46$$
$$NIE/TE = 0.07, NDE/TE = 0.696, 1 - NDE/TE = 0.304$$

We conclude that the program as a whole has increased the success rate by 46% and that a significant portion, 30.4%, of this increase is due to the capacity of the program to stimulate improved homework effort. At the same time, only 7% of the increase can be explained by stimulated homework alone without the benefit of the program itself.