THE UNIVERSITY of EDINBURGH

Methods for Causal Inference Lecture 15: Counterfactual

Ava Khamseh

School of Informatics
2025-2026

Counterfactuals (revisited)

You wish to go to a bar with your friends to have a drink on Friday night. You would prefer if the bar is not too busy so that the drink can be served quickly.

You choose to go to a bar in Cowgate, and it turns out to be busy. You might be saying “We should have just gone to Teviot.”

Colloquially: “If we had gone to Teviot, it would have been quieter and we would have been served more quickly.”

Counterfactuals: An ‘if’ statement in which the ‘if’ position is untrue or unrealised. (‘if’ here is a hypothetical condition).

We wish to compare two outcomes (serving time above), under the same conditions differing only in one aspect (going to Teviot above).

Counterfactuals (revisited)

Knowing the outcome of the action we have already taken is important, because our estimate of serving time at Teviot may be different after observing the time at Cowgate, which may offer valuable evidence of our assessment (e.g. break/ weekend, more students are going to the bar).

If we use the do-expression to estimate the serving time we will have a problem:

E[serving time | do(Teviot), serving time = 40 mins]

Clash: Actual serving time (at Cowgate, known) and the hypothetical one under do(Teviot)

Counterfactuals (potential outcomes notation)

Use different subscripts to label the two outcomes (as done before):

$Y_{T=1}$: Serving time at Teviot
$Y_{T=0}$: Serving time at Cowgate

So we wish to estimate: $\mathbb{E}[Y_{T=1}|T=0, Y=Y_0=40\text{mins}]$

Notice the apparent contradiction in this notation, two different worlds.
We seek to estimate the serving time in a world where we chose Teviot given that in the actual world, Cowgate, the time it took to serve the drink was 40mins.

Compare with do-operator: $\mathbb{E}[Y|do(T=t)]$
This expression seeks to estimate the serving time in a world that we chose Teviot, without reference to the other world (Cowgate) whatsoever.

Counterfactuals: Keep identification & estimation separate

Estimation: Measure the serving time at Teviot at a different time.
Here will be making approximations under certain assumptions. We have to make sure these assumptions are appropriate.

BUT, **defining** the counterfactual should not require approximation. Definitions should accurately capture what we wish to estimate precisely.
(How we then estimate it is a different problem).

Defining and estimating counterfactual allows us to address complex problems:

- efficacy of a job training programme by identifying how may enrolled would have gotten jobs had they not enrolled
- Predict the effect of an additive intervention (adding 5 mg/l of insulin to a group of patient with varying insulin levels), from experimental studies
- Obtain the likelihood that an individual cancer patient would have had a different outcome, had they chosen a different treatment

Counterfactuals: SCM

In a fully specified causal model M: Know all functions F, and the value of all exogenous variables U. Then every assignment of U=u correspond to a single member (unit) in a population or a situation in nature.

People: e.g., salary, area, education, hobbies, etc.
i.e., so many of the defining properties that cannot all be included in the model, such that, when taken together, they can uniquely distinguish each individuals

So U=u here represents a unique individual.

Example: Simple causal model $T = aU$

$Y = bT + U$

Counterfactuals: SCM

In a fully specified causal model M: Know all functions F, and the value of all exogenous variables U. Then every assignment of U=u correspond to a single member (unit) in a population or a situation in nature.

People: e.g., salary, area, education, hobbies, etc.
i.e., so many of the defining properties that cannot all be included in the model, such that, when taken together, they can uniquely distinguish each individuals

So U=u here represents a unique individual.

Example: Simple causal model $T = aU$

$Y = bT + U$

What Y would be had T been t in situation U = u: $Y_t(u) = bt + u$

Counterfactuals: SCM

In a fully specified causal model M: Know all functions F, and the value of all exogenous variables U. Then every assignment of U=u correspond to a single member (unit) in a population or a situation in nature.

People: e.g., salary, area, education, hobbies, etc.
i.e., so many of the defining properties that cannot all be included in the model, such that, when taken together, they can uniquely distinguish each individuals

So U=u here represents a unique individual.

Notice also that the do(t)-operator capture the behaviour of a population under intervention, while $Y_t(u)$ describes the behaviour of a specific individual U=u under such an intervention.

Counterfactuals: The Fundamental Law (seen before)

If we observe $T(u) = 1, Y(u) = 0$ then $Y_{T=1}(u) = 0$

(Setting T to a value it already has should not produce a change in the world.)

Consistency rule:

$$\text{if} \;\; T = t \;\; \text{then} \;\; Y_t = Y$$

If T is binary, the consistency rule can be written as:

$$Y = TY_1 + (1 - T)Y_0$$

Counterfactual interpretation of Backdoor (Theorem)

If a set Z of variables satisfies the backdoor condition relative to (T,Y), then, for all t, the counterfactual Y_t is conditionally independent of T given Z:

$$P(Y_t|T,Z) = P(Y_t|Z)$$

Counterfactual interpretation of Backdoor (Theorem)

If a set Z of variables satisfies the backdoor condition relative to (T,Y), then, for all t, the counterfactual Y_t is conditionally independent of T given Z:

$$P(Y_t|T,Z) = P(Y_t|Z)$$

This implies **we can estimate counterfactuals from observational studies**, in particular $P(Y_t = y)$ is identifiable via the adjustment formula:

$$\begin{aligned} P(Y_t = y) &= \sum_z p(Y_t = y|Z = z)P(z) && \text{Marginalise} \\ &= \sum_z p(Y_t = y|T = t, Z = z)P(z) && \text{Above theorem} \\ &= \sum_z p(Y = y|T = t, Z = z)P(z) && \text{Consistency} \end{aligned}$$

Counterfactuals: Example

4.4.1 Recruitment to a Program

Example 4.4.1 *A government is funding a job training program aimed at getting jobless people back into the workforce. A pilot randomized experiment shows that the program is effective; a higher percentage of people were hired among those who finished the program than among those who did not go through the program. As a result, the program is approved, and a recruitment effort is launched to encourage enrollment among the unemployed, by offering the job training program to any unemployed person who elects to enroll.*

Lo and behold, enrollment is successful, and the hiring rate among the program's graduates turns out even higher than in the randomized pilot study. The program developers are happy with the results and decide to request additional funding.

Pearl's Primer book Chapter 4, page 107

Counterfactuals: Example

Critics: This programme is a waste of tax payer's money and should be terminated.
Reasoning: The one's who self-enrol are more intelligent, resourceful, socially connected, so they were more likely to find a job anyway, regardless of the training programme. (i.e., reality is different from the randomised experiment perform prior to the full launch).

Hence, we need to estimate the differential benefit of the programme to know if it's useful.

What is the extent to which hiring rate has increased among the enrolled, compared to what would have been had they not been trained.

Pearl's Primer book Chapter 4, page 107

Counterfactuals: Example

Training: T = 1
Hiring: Y = 1

Effect of treatment on the treated (ETT) $\mathbb{E}[Y_1 - Y_0|T = 1]$

The difference in the causal effect of training, conditional on those who actually chose the training programme on their own initiative.

Recall out Cowgate/Teviot example, with Y_0 being the issue here. Not serious consequences there, in this new example we are talking about important socio-economical/political implications.

Shall we stop the programme or redesign or keep going?

Pearl's Primer book Chapter 4, page 107

Counterfactuals: Example

To compute $\mathbb{E}[Y_1 - Y_0|T = 1]$ we need,

$$p(Y_t = y|T = t') = \sum_z p(Y_t = y|z, T = t')p(z|t') \qquad \text{Marginalise}$$

$$= \sum_z p(Y_t = y|z, T = t)p(z|t') \qquad \text{Prev theorem}$$

$$= \sum_z p(Y = y|z, T = t)p(z|t') \qquad \text{Consistency}$$

Pearl's Primer book Chapter 4, page 107

Counterfactuals: Example

To compute $\mathbb{E}[Y_1 - Y_0|T = 1]$ we need,

$$p(Y_t = y|T = t') = \sum_z p(Y_t = y|z, T = t')p(z|t') \qquad \text{Marginalise}$$

$$= \sum_z p(Y_t = y|z, T = t)p(z|t') \qquad \text{Prev theorem}$$

$$= \sum_z p(Y = y|z, T = t)p(z|t') \qquad \text{Consistency}$$

Contrast with do-operator (different weights):

$$p(Y = y|do(T = t)) = \sum_z p(Y = y|T = t, Z = z)P(z)$$

Pearl's Primer book Chapter 4, page 107

Counterfactuals: Example

To compute $\mathbb{E}[Y_1 - Y_0|T = 1]$ we need,

$$p(Y_t = y|T = t') = \sum_z p(Y_t = y|z, T = t')p(z|t') \qquad \text{Marginalise}$$

$$= \sum_z p(Y_t = y|z, T = t)p(z|t') \qquad \text{Prev theorem}$$

$$= \sum_z p(Y = y|z, T = t)p(z|t') \qquad \text{Consistency}$$

Continuing:

$$\mathbb{E}[Y_1 - Y_0|T = 1] = \mathbb{E}[Y|T = 1] - \sum_z \mathbb{E}[Y|T = 0, Z = z]p(Z = z|T = 1)$$

Pearl's Primer book Chapter 4, page 107

Additive interventions: Idea

The operator $\mathrm{do}(T = t)$ sets the variable T to some amount *by disabling all pre-existing causes of T* (i.e., removing any incoming arrows to T)

This might not be what we want! (Want to keep effect of incoming arrows to T)

Additive interventions: Idea

The operator $\mathrm{do}(T = t)$ sets the variable T to some amount *by disabling all pre-existing causes of T* (i.e., removing any incoming arrows to T)

This might not be what we want! (Want to keep effect of incoming arrows to T)

Example

A doctor might prescribe an *extra* 5mg/L of insulin to diabetes patients
These patients already have varying levels of insulin in their blood, e.g., affected by what they ate that morning.

Specifying the total level of insulin (i.e., applying do-operator) is not desired.
Instead the doctor simply wishes to *add* some insulin
From experience, they expect/know the causal effect on health to be positive

Goal: Quantify the effect of *additive interventions* from exp/observational data

Additive interventions: Identifiability

Goal: Understand if additive interventions can be identified from observational studies, or from experimental studies in which T is uniformly set to some T = t?

Approach: Write out the required quantity using counterfactuals

1. Suppose we add a quantity *q* to the treatment variable T currently at T = t' for an individual. The resulting outcome is the counterfactual $Y_{t'+q}$
2. We average this counterfactual over all units at level T = t'

Additive interventions: Identifiability

Goal: Understand if additive interventions can be identified from observational studies, or from experimental studies in which T is uniformly set to some T = t?

Approach: Write out the required quantity using counterfactuals

1. Suppose we add a quantity *q* to the treatment variable T currently at T = t' for an individual. The resulting outcome is the counterfactual $Y_{t'+q}$
2. We average this counterfactual over all units at level T = t'

Use what we just derived for ATT, with $\mathbb{E}[Y_t \mid T = t']$ where $t = t' + q$

Multiply both sides by y and sum over y, and by p(T=t') sum over t':

$$\sum_{t',y} y\ p(Y_{t'+q} = y|T = t')p(t') = \sum_{t',y,z} y\ p(Y = y|Z = z, T = t' + q)p(z|t')p(T = t')$$

$$\Rightarrow \underbrace{\sum_{t'} \mathbb{E}[Y_{t'+q}|T = t']p(T = t')}_{\equiv \mathbb{E}[Y|\mathrm{add}(q)]} = \sum_{t',z} \mathbb{E}[Y|Z = z, T = t' + q]p(Z = z|T = t')p(T = t')$$

Additive interventions: Identifiability

Goal: Understand if additive interventions can be identified from observational studies, or from experimental studies in which T is uniformly set to some T = t?

Approach: Write out the required quantity using counterfactuals

1. Suppose we add a quantity *q* to the treatment variable T currently at T = t' for an individual. The resulting outcome is the counterfactual $Y_{t'+q}$
2. We average this counterfactual over all units at level T = t'

Use what we just derived for ATT, with $\mathbb{E}[Y_t \mid T = t']$ where $t, = t' + q$

Therefore, with Z a set of variables satisfying the backdoor for (T,Y):

$$\mathbb{E}[Y|\mathrm{add}(q)] - \mathbb{E}[Y] = \sum_{t',z} \mathbb{E}[Y|Z = z, T = t' + q]p(Z = z|T = t')p(T = t') - \mathbb{E}[Y]$$

Additive interventions: Why are they necessary?

Shows the use of counterfactuals to estimate effect of practical interventions, which cannot always be (easily) described as *do*-expressions.

So why are counterfactuals needed, and what would a simple *do*-operator do instead? This intervention resembles a typical *"scientific"* intervention, asking the causal effect of increasing insulin from T = t to another level T = t+q:

$$\mathbb{E}[Y|\mathrm{do}(T = t + q)] - \mathbb{E}[Y|\mathrm{do}(T = t)]$$

Additive interventions: Why are they necessary?

Shows the use of counterfactuals to estimate effect of practical interventions, which cannot always be (easily) described as *do*-expressions.

So why are counterfactuals needed, and what would a simple *do*-operator do instead? This intervention resembles a typical *"scientific"* intervention, asking the causal effect of increasing insulin from T = t to another level T = t+q:

$$\mathbb{E}[Y|\mathrm{do}(T=t+q)] - \mathbb{E}[Y|\mathrm{do}(T=t)]$$

Scientific: Population-*independent* (we work with a random sample), thus revealing biologically meaningful effect of insulin on health outcome

Policy question (our intervention): Effect of adding q to each individual's insulin, regardless of their current level, so population-*dependent*

Note: Crucial distinction between scientific query and policy question

Pearl's Primer book Chapter 4, page 110

Additive interventions: Why are they necessary?

Another guess for the policy question $\mathbb{E}[Y|\mathrm{add}(q)]$ would be to try and take the population-specificity into account, would be the average causal effect

$$\sum_t \Big(\mathbb{E}[Y|\mathrm{do}(T=t+q)] - \mathbb{E}[Y|\mathrm{do}(T=t)]\Big) P(T=t)$$

However, this represents an experiment in which subjects are chosen at random from a population, a fraction $P(T=t)$ are given an additional dose q, and the rest are left alone … that is not the policy question!

Additive interventions: Why are they necessary?

Another guess for the policy question $\mathbb{E}[Y|\mathrm{add}(q)]$ would be to try and take the population-specificity into account, would be the average causal effect

$$\sum_t \Big(\mathbb{E}[Y|\mathrm{do}(T=t+q)] - \mathbb{E}[Y|\mathrm{do}(T=t)]\Big) P(T=t)$$

However, this represents an experiment in which subjects are chosen at random from a population, a fraction $P(T=t)$ are given an additional dose q, and the rest are left alone … that is not the policy question!

In the policy question, $P(T=t)$ is the proportion taking T = t by free choice. Those that attain T = t by free choice might respond to $\mathrm{add}(q)$ differently than those who are to receive T = t by experimental design. Similar to ATE vs ATT:

$$\mathbb{E}[Y|\mathrm{add}(q)] = \sum_t \mathbb{E}[Y_{t+q}|T=t]P(T=t) \neq \sum_t \mathbb{E}[Y_{t+q}]P(T=t)$$

Equality holds when there are no confounders: $Y_t \perp\!\!\!\perp T$

Personal decision making

Counterfactuals allow us to extract subtle information on individual risks (i.e., probabilities) from population data, under appropriate assumptions

Example
Cancer patients face a tough decision between two treatments: (i) surgery, or (ii) surgery + chemotherapy. We consider the situation of three patients:

Personal decision making

Counterfactuals allow us to extract subtle information on individual risks (i.e., probabilities) from population data, under appropriate assumptions

Example
Cancer patients face a tough decision between two treatments: (i) surgery, or (ii) surgery + chemotherapy. We consider the situation of three patients:

A. Patient A takes option (ii). Ten years later, she is alive and the tumour has not come back. **Q:** Was (ii) *necessary* for this outcome, or was (i) enough?

Personal decision making

Counterfactuals allow us to extract subtle information on individual risks (i.e., probabilities) from population data, under appropriate assumptions

Example
Cancer patients face a tough decision between two treatments: (i) surgery, or (ii) surgery + chemotherapy. We consider the situation of three patients:

A. Patient A takes option (ii). Ten years later, she is alive and the tumour has not come back. **Q:** Was (ii) *necessary* for this outcome, or was (i) enough?
B. Patient B takes option (i). One year later, unfortunately, the tumour has come back. **Q:** Would (ii) have been *sufficient* to prevent recurrence?

Personal decision making

Counterfactuals allow us to extract subtle information on individual risks (i.e., probabilities) from population data, under appropriate assumptions

Example

Cancer patients face a tough decision between two treatments: (i) surgery, or (ii) surgery + chemotherapy. We consider the situation of three patients:

A. Patient A takes option (ii). Ten years later, she is alive and the tumour has not come back. **Q:** Was (ii) *necessary* for this outcome, or was (i) enough?
B. Patient B takes option (i). One year later, unfortunately, the tumour has come back. **Q:** Would (ii) have been *sufficient* to prevent recurrence?
C. Patient C is faced with the same choice: If (ii) is *necessary* to remove the tumour, she wants to go for (ii). However, if neither (i) nor (ii) is *sufficient* to eliminate the tumour, she would rather avoid the chemotherapy. **Q:** Is surgery + chemo both necessary *and* sufficient for eliminating the tumour?

Personal decision making: Patient A

Cancer patients face a tough decision between two treatments: (i) surgery, or (ii) surgery + chemotherapy. We consider the situation of three patients:

A. Patient A takes option (ii). Ten years later, she is alive and the tumour has not come back. **Q:** Was (ii) *necessary* for this outcome, or was (i) enough?

Write Y = 1 for no tumour, and T = 1 for option (ii). Patient A wants to know the *probability of necessity* in attributing Y = 1 to T = 1:

$$\mathrm{PN} = P(Y_0 = 0 \mid T = 1, Y = 1)$$

In words: the probability that the tumour would have come back (Y = 0) had Patient A not chosen chemotherapy, given that Patient A in fact chose option (ii) (i.e. T = 1) and the tumour did not come back (Y = 1).

—> Was T = 1 *necessary* for the positive outcome Y = 1?

Personal decision making: Patient B

Cancer patients face a tough decision between two treatments: (i) surgery, or (ii) surgery + chemotherapy. We consider the situation of three patients:

B. Patient B takes option (i). One year later, unfortunately, the tumour has come back. **Q:** Would (ii) have been *sufficient* to prevent recurrence?

Write Y = 1 for no tumour, and T = 1 for option (ii). Patient B wants to know the *probability of sufficiency* if the action **not** taken, T = 1, would have sufficed:

$$\mathrm{PS} = P(Y_1 = 1 \mid T = 0, Y = 0)$$

In words: the probability that the tumour would not have come back had Patient B gone through option (ii) ($Y_1 =$)1 , given that Patient B in fact did not choose option (ii) (i.e. T = 0) and the tumour came back (Y = 0).

—> Would T = 1 have been *sufficient* to guarantee a positive outcome Y = 1?

Personal decision making: Why bother?

An intermediate philosophical question: What is gained by assessing these retrospective counterfactual parameters? After all, decisions have been taken.

If shown successful or not, it informs future decision-making strategies (on any kind of problem) if the patients know they made the better call or not.

“It is through counterfactual reinforcement that we learn to improve our own decision-making processes and achieve higher performance”

Personal decision making: Patient C

Cancer patients face a tough decision between two treatments: (i) surgery, or (ii) surgery + chemotherapy. We consider the situation of three patients:

C. Patient C is faced with the same choice: If (ii) is *necessary* to remove the tumour, she wants to for (ii). However, if neither (i) nor (ii) is *sufficient* to remove the tumour, she would rather avoid the chemotherapy. **Q:** Is surgery + chemo both necessary *and* sufficient for eliminating the tumour?

Patient C wants to know the probability that additional chemotherapy (i.e. option (ii)) is both *necessary* and *sufficient* for eliminating her tumour:

$$\text{PNS} = P(Y_1 = 1, Y_0 = 0)$$

Variables mean eliminating the tumour with chemo (Y_1) and without (Y_0)

—> Seems fundamentally inestimable from experimental studies!

Personal decision making: Patient C

Patient C wants to know the probability that additional chemotherapy (i.e. option (ii)) is both *necessary and sufficient* for eliminating her tumour:

$$\text{PNS} = P(Y_1 = 1, Y_0 = 0)$$

Variables mean eliminating the tumour with chemo (Y_1) and without (Y_0)

However, mathematical framework allows us to investigate algebraically if and under what assumptions we can estimate PNS from what type of data.

Under *monotonicity* (i.e., chemo cannot cause the tumour to recur if it was about to go away), experimental data are sufficient to conclude (Pearl 2000, Chapter 9 and Tian and Pearl 2000, not proven in this course):

$$\text{PNS} = P(Y = 1|\text{do}(T = 1)) - P(Y = 1|\text{do}(T = 0))$$

Thus, remarkably, experimental data from Fisher et al. (2002) allow us to estimate this quantity at

$$\text{PNS} = 0.39 - 0.14 = 0.25$$

Pearl's Primer book
Chapter 4, page 113 & 120

Back to Mediation

Mediation and Path-disabling Interventions

Example 4.4.5 *A policy maker wishes to assess the extent to which gender disparity in hiring can be reduced by making hiring decisions gender-blind, rather than eliminating gender inequality in education or job training. The former concerns the "direct effect" of gender on hiring, whereas the latter concerns the "indirect effect," or the effect* mediated *via job qualification.*

Aim: Which of the two causal effects is greater (i) the direct effect (gender on hiring), or (ii) the indirect effect (education on job qualification on hiring)?
—> Could inform policy where to invest resources to address disparity

Pearl's Primer book Chapter 4, page 114

Mediation and Path-disabling Interventions

Example 4.4.5 *A policy maker wishes to assess the extent to which gender disparity in hiring can be reduced by making hiring decisions gender-blind, rather than eliminating gender inequality in education or job training. The former concerns the "direct effect" of gender on hiring, whereas the latter concerns the "indirect effect," or the effect* mediated *via job qualification.*

Aim: Which of the two causal effects is greater (i) the direct effect (gender on hiring), or (ii) the indirect effect (education on job qualification on hiring)?
—> Could inform policy where to invest resources to address disparity

This concerns enabling/disabling processes (e.g., educational reforms) rather than lowering/raising values of specific variables. Thus, the do-operator and the controlled direct effect (CDE) seen earlier do not suffice …

… as before, we phrase the problem mathematically via counterfactuals!

Pearl's Primer book Chapter 4, page 114

Mediation and Path-disabling Interventions

How do we phrase this in a counterfactual manner?

For example, we want to know how the gender disparity changes *after* successfully implementing gender-blind hiring procedures.

In words: We estimate gender disparity under the counterfactual condition that all female applicants be treated as males

Pearl's Primer book Chapter 4, page 115

Mediation and Path-disabling Interventions

How do we phrase this in a counterfactual manner?

For example, we want to know how the gender disparity changes *after* successfully implementing gender-blind hiring procedures.

In words: We estimate gender disparity under the counterfactual condition that all female applicants be treated as males

Hiring status (Y) of a female applicant with qualification Q = q, given that the employer treats her as though she is a male (X=1) is captured by the counterfactual $Y_{X=1,Q=q}$

Since Q varies over the population, we average this quantity according to the distribution of the qualification of female applicants, $p(Q = q|X = 0)$

Pearl's Primer book Chapter 4, page 115

Mediation and Path-disabling Interventions

Since Q varies over the population, we average this quantity according to the distribution of the qualification of female applicants, $p(Q = q|X = 0)$

The result is $\sum_q \mathbb{E}\left[Y_{X=1,Q=q}\right] p(Q = q|X = 0)$

Male applicants have similar chances, but averaging over $p(Q = q|X = 1)$

Mediation and Path-disabling Interventions

Since Q varies over the population, we average this quantity according to the distribution of the qualification of female applicants, $p(Q = q|X = 0)$

The result is $\sum_q \mathbb{E}\left[Y_{X=1,Q=q}\right] p(Q = q|X = 0)$

Male applicants have similar chances, but averaging over $p(Q = q|X = 1)$

Subtracting the two quantities yields the *Natural Indirect Effect (NIE)* of gender on hiring, mediated by the level of qualification Q:

$$\text{NIE} = \sum_q \mathbb{E}\left[Y_{X=1,Q=q}\right] \left(p(Q = q|X = 0) - p(Q = q|X = 1)\right)$$

Allow Q to vary naturally between applicants, as opposed to the CDE. Here we disable the capacity of Y to respond to X but leave its response to Q unaltered.

Pearl's Primer book Chapter 4, page 115

Mediation and Path-disabling Interventions

It remains to identify the *Natural Indirect Effect (NIE)* of gender on hiring, mediated by the level of qualification Q, in order to allow estimation:

$$\mathrm{NIE} = \sum_q \mathbb{E}\left[Y_{X=1,Q=q}\right]\left(p(Q=q|X=0) - p(Q=q|X=1)\right)$$

The following result is known as Pearl's ***Mediation formula***

Pearl's Primer book Chapter 4, page 115

Mediation and Path-disabling Interventions

It remains to identify the *Natural Indirect Effect (NIE)* of gender on hiring, mediated by the level of qualification Q, in order to allow estimation:

$$\text{NIE} = \sum_q \mathbb{E}\left[Y_{X=1,Q=q}\right]\left(p(Q=q|X=0) - p(Q=q|X=1)\right)$$

The following result is known as Pearl's ***Mediation formula***

Theorem (Pearl, 2001)

In the absence of confounding, the NIE can be identified as follows

$$\text{NIE} = \sum_q \mathbb{E}\left[Y|X=1,Q=q\right]\left(p(Q=q|X=0) - p(Q=q|X=1)\right)$$

In words: It measures the extent to which the effect of X on Y is *explained* by its effect on the mediator Q. In the NIE we "freeze" the direct effect of X on Y, yet allow the mediator Q of each unit to react to X in a natural "unfrozen" way.

Pearl's Primer book Chapter 4, page 115