# An introduction to g-formula

## Causal fanclub

V.Perduca - MAP5 (Université Paris Cité) and CESP

26 june 2026

# Contents

1. Counterfactual framework
2. g-formula
3. Ignorability and backdoor criterion
4. Estimation

## 1. Counterfactual framework

# Counterfactual variables, Rubin (1980)

We consider i.i.d. data $(Z_i, X_i, Y_i, Y_i(0), Y_i(1))$ for subjects $i = 1, \dots, n$ where:

- $Z_i$: **binary** indicator of treatment
- $X_i$: baseline covariate(s)
- $Y_i$: observed outcome
- $Y_i(1)$: counterfactual (or potential) outcome under the assignment $Z_i = 1$
- $Y_i(0)$: counterfactual outcome under the assignment $Z_i = 0$

Under the Stable Unit Value Assumption (SUTVA) each unit has only two counterfactual outcomes $Y_i(0), Y_i(1)$ which are connected to $Y_i$ by the *consistency* identity

$$(Y_i(z) = Y_i) | Z_i = z$$

that is

$$Y_i = Z_i Y_i(1) + (1 - Z_i) Y_i(0).$$

# Ideal setting for causal inference

- To establish whether/to what extent exposure $Z$ is a cause of $Y$, the ideal (unrealistic) experiment would provide

| $i$ | $Y_i(0)$ | $Y_i(1)$ |
|---|---|---|
| 1 | 0 | 1 |
| 2 | 0 | 0 |
| $\vdots$ | $\vdots$ | $\vdots$ |
| $n$ | 1 | 1 |

This would allow to evaluate causal effects

- at the individual level:

$$Y_i(1) - Y_i(0)$$

- at the population level (*average causal effect*):

$$\tau = E[Y_i(1) - Y_i(0)] = E[Y_i(1)] - E[Y_i(0)].$$

# Fundamental problem of observational data

Ideally, data would be ...

| $i$ | $Y_i(0)$ | $Y_i(1)$ |
|---|---|---|
| 1 | 0 | 1 |
| 2 | 0 | 0 |
| $\vdots$ | $\vdots$ | $\vdots$ |
| $n$ | 1 | 1 |

... while observational data are

| $i$ | $Z_i$ | $Y_i$ | $Y_i(0)$ | $Y_i(1)$ |
|---|---|---|---|---|
| 1 | 0 | 0 | 0 | NA |
| 2 | 1 | 0 | NA | 0 |
| $\vdots$ | $\vdots$ | $\vdots$ | | |
| $n$ | 0 | 1 | 1 | NA |

*How to identify, e.g., $E[Y_i(1) - Y_i(0)]$ from an observed experiment generating $Z_i, Y_i$ and other covariates $X_i$?*

Without further assumptions we cannot!

# Assumptions

1. SUTVA
2. Unconfoundedness (conditional ignorability):

$$Z \perp Y(z) | X \text{ for } z \in \{0, 1\}$$

3. Positivity:

$$0 < e(X) < 1 \quad \text{where} \quad e(X) = P(Z = 1 | X)$$

# 2. g-formula

# g-formula: standardization and adjustment variables

Let $P$ the distribution of $(Z, X, Y, Y(0), Y(1))$. For simplicity we suppose $P$ discrete. Under ignorability and positivity, for each $z \in \{0, 1\}$:

$$P(Y(z) = y) = \sum_x P(Y = y | Z = z, X = x) P(X = x)$$

Proof:

$$\begin{aligned}
P(Y(z) = y) &= \sum_x P(Y(z) = y, X = x) \text{ (total probability)} \\
&= \sum_x P(Y(z) = y | X = x) P(X = x) \text{ (conditional prob.)} \\
&= \sum_x P(Y(z) = y | Z = z, X = x) P(X = x) \text{ (ignorability)} \\
&= \sum_x P(Y = y | Z = z, X = x) P(X = x) \text{ (consistency)}
\end{aligned}$$

This formula is also referred to as *standardization* and $X$ is called an *adjustment* variable.

## g-formula: expectation of the counterfactual outcomes

Under ignorability and positivity, for each $z \in \{0,1\}$:

$$E[Y(z)] = E[\mu_z(X)]$$

where $\mu_z(X) = E[Y|Z = z, X]$.

Proof:

$$\begin{aligned}
\mathbb{E}[Y(z)] &= \sum_y yP(Y(z) = y) \text{ (def. of expectation)} \\
&= \sum_y y \sum_x P(Y = y|Z = z, X = x)P(X = x) \\
&= \sum_x \left( \sum_y yP(Y = y|Z = z, X = x) \right) P(X = x) \\
&= \sum_x E[Y|Z = z, X = x] \\
&= E[E[Y|Z = z, X = x]].
\end{aligned}$$

# Example: comparing kidney stone removal modus operandi

- Summary statistics from observational data:

| | $d < 2$cm | | $d \geq 2$cm | | all $d$s | |
|---|---|---|---|---|---|---|
| | success | failure | success | failure | success | failure |
| $m_1$ | 81 | 6 | 192 | 71 | 273 | 77 |
| $m_2$ | 234 | 36 | 55 | 25 | 289 | 61 |

- Modus operandi vs Success, unconditional inference:

$$P_n(\text{success}|m_1) = \frac{273}{350} \approx 78\% < P_n(\text{success}|m_2) = \frac{289}{350} = 83\%$$

- It look likes chances of success are higher with $m_2$...

# Example (cont'nd)

- Modus operandi vs Success conditionally on Stone size:

$$P_n(\text{success}|d < 2, m_1) = \frac{81}{87} \approx 93\% > P_n(\text{success}|d < 2, m_2) = \frac{234}{270} \approx 87\%$$
$$P_n(\text{success}|d \geq 2, m_1) = \frac{192}{263} \approx 73\% > P_n(\text{success}|d \geq 2, m_2) = \frac{55}{80} \approx 69\%$$

- For each stone size, chances of success are higher with $m_1$. Note the **association reversal**! This is an instance of the so called Simpson's paradox.

# Example (cont'nd)

- $P(\text{success}(m_1))$ is more interesting than $P(\text{success}|m_1)$
- It turns out that the operation success given $m_1$, $\text{success}(m_1)$, is independent from the observed modus operandi $m$ given the stone size $d$ (and similarly for $m_2$)
- $\Rightarrow$ by standardization we have

$$\begin{aligned}\mathbb{P}_n(\text{success}(m_1)) &= P_n(\text{success}|d<2, m_1) \times P_n(d<2) + \\ &+ P_n(\text{success}|d \geq 2, m_1) \times P_n(d \geq 2) \\ &= \frac{81}{87} \times \frac{357}{700} + \frac{192}{263} \times \frac{343}{700} \approx 83\%.\end{aligned}$$

- Similarly, we can compute

$$\mathbb{P}_n(\text{success}(m_2)) \approx 78\%.$$

# g-formula: identification of the ATE

- Under ignorability and positivity the ATE is identifiable:

$$\tau = E[Y(1)] - E[Y(0)] = E[\mu_1(X)] - E[\mu_0(X)]$$

  where $\mu_1(X) = E[Y|Z=1, X]$ and $\mu_0(X) = E[Y|Z=0, X]$ (Q-models)

- This formula justifies several estimation strategies (more on this later...).

- Problem: Ignorability **cannot** be tested on data!

# 3. Ignorability and backdoor criterion

# Causal DAGs

- $\rightarrow$ how to find $X$ such that $Z \perp Y(z)|X$, that is a set of **adjustment variables**?
- One can adopt the DAG approach, Pearl (2016), and apply graphical criteria like the *backdoor criterion*
- Instead of reasoning on the abstract relation $Y(z) \perp Z|X$ for each $z$, we
  - represent our assumptions on the data generating mechanism in the form of a DAG
  - use graphical rules to look for variable $X$ verifying ignorability

# Quick reminder: d-separation and do operator

- A DAG is a graph where
  - vertices represent variables
  - arrows represent dependence relationships
- the DAG topology defines a distribution of $(X, Z, Y)$
- *d-separation*: $A \perp B | C$ if and only if all paths connecting $A$ and $B$ are blocked by $C$
- causal semantics is brought by the do$(\cdot)$ operator:

$$Y | do(Z = z) \text{ is equivalent to } Y(z)$$

- the do$(\cdot)$ operator allows to define the post-interventional distribution, that is the distribution $P$ of $(X, Z, Y, Y(0), Y(1))$.

# Backdoor-criterion

We say that $X$ satisfies the back-door criterion w.r.t. $(Z, Y)$ if

1. $X$ does not contain any descendant of $Z$
2. $X$ blocks all **back-door paths** between $Z$ and $Y$, that is all paths terminating with an arrow pointing to $Z$.

**Theorem**

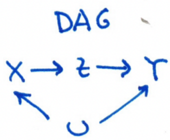
Consider a DAG and let $P$ be the distribution of $(X, Z, Y, Y(0), Y(1))$ associated to the DAG. If $X$ satisfies the back-door criterion, then $Y(z) \perp Z|X$ for all $z$ in the distribution $P$.

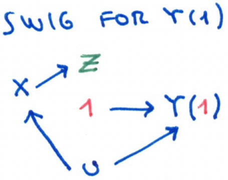
To prove the theorem it is useful to build a new DAG with both $Z, z$ and $Y(z)$ because then we can use d-separation to check if $Y(z) \perp Z|X$.

# Single world intervention graphs (SWIGs)

The *single-world intervention graph (SWIG)* associated to a DAG under the intervention that sets $Z$ to $z$ is built from the DAG as follows:

1. split the treatment node $Z$ into two nodes: $Z$ and $z$
2. $z$ inherits all the arrows that were out of $Z$ in the original DAG
3. $Z$ inherits all the arrows that were into $Z$
4. replace $Y$ with $Y(z)$



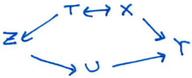



# Simplified proof of the backdoor-criterion

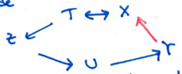
We assume

- $T$ is the parent of $Z$, $U$ is the child of $Z$
- $X$ is an individual variable
- there are no more variables

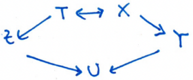
We have the following "important" possible configurations:

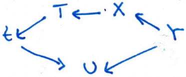



We exclude

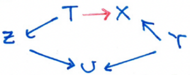



because X would be a descendant of Z



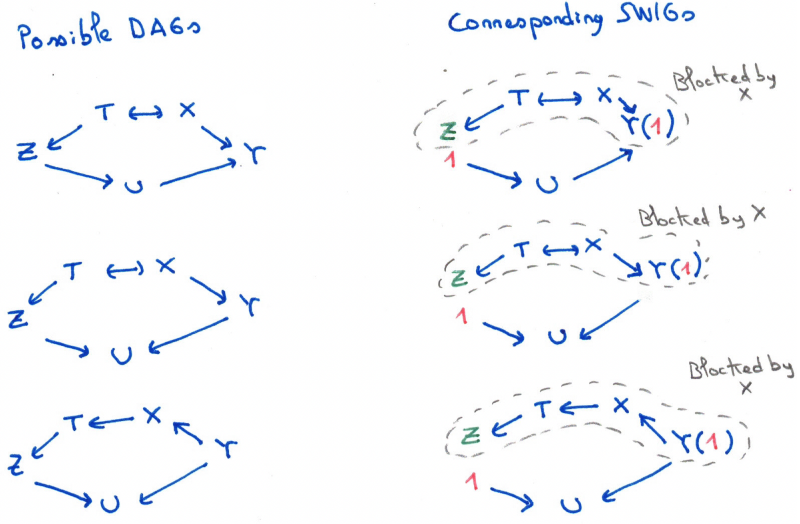




We exclude




because X would open the backdoor

# Simplified proof of the backdoor-criterion (cont'nd)




In all possible SWIGS all paths from $Z$ to $Y(z)$ are *blocked* by $X$, henceforth $Z \perp Y(z)|X$.

# 4. Estimation

# g-formula estimator

- The g-formula
$$\tau = E\left[\mu_1(X) - \mu_0(X)\right]$$
justifies the plug-in estimator:
$$\hat{\tau} = \frac{1}{n}\sum_{i=1}^{n}\left[\hat{\mu}_1(X_i) - \hat{\mu}_0(X_i)\right]$$
where $\hat{\mu}_1(X)$ is an estimate of $E[Y|Z=1,X]$ and $\hat{\mu}_0(X)$ is an estimate of $E[Y|Z=0,X]$
- This estimator is known as plug-in g-formula or outcome regression
- Q-models can be estimated with linear models, random forests, ...
- The estimator is consistent if the Q-models are **well specified** and **consistently estimated**

# A special case: outcome regression is linear

- Suppose $P$ entails $\mu_Z(X) = E[Y|Z, X] = \beta_0 + \beta_1 X + \beta_2 Z$ and ignorability holds
- Then

$$\begin{aligned}\tau &= E[\mu_1(X)] - E[\mu_0(X)] \\ &= E[\beta_0 + \beta_1 X + \beta_2 \times 1] - E[\beta_0 + \beta_1 X + \beta_2 \times 0] \\ &= \beta_2\end{aligned}$$

- $\Rightarrow$ the ATE is equal to the coefficient of the treatment $Z$ in the outcome regression
- This justifies the interpretation of $\beta_2$ estimates, e.g. ordinary least squares (OLS) estimates, in causal terms **if in the true model** $E[Y|Z, X] = \beta_0 + \beta_1 X + \beta_2 Z$
- In general, if $\mu_Z(X)$ is not linear, we cannot interpret estimates of $\beta_2$ causally!

# Example: ATE $\neq$ coefficient of $X$

$$P : \begin{cases} X \sim \mathcal{N}(0,2) \\ Z|X = x \sim \mathcal{B}(p_x) \text{ with } p_x = \begin{cases} 0.8 \text{ if } x \leq 0 \\ 0.3 \text{ if } x > 0 \end{cases} \\ Y(0) \sim X + \mathcal{N}(0,2) \\ Y(1) \sim X + \exp(Z) + \mathcal{N}(0,2) \end{cases}$$

It is easy to see that

$$\begin{aligned} \tau &= E[Y(1) - Y(0)] \\ &= E[\exp(Z)] \\ &= \exp(0 + 2^2/2) = \exp(2) \simeq 7.4 \end{aligned}$$

and $E[Y|Z,X]$ is not linear in $X$ and $Z$:

$$\begin{aligned} E[Y|Z,X] &= E\left[(1-Z)Y(0) + ZY(1)\right] \\ &= X + Z\exp(Z) \end{aligned}$$

# Example (cont'nd)

```r
N <- 1E6
df <- data.frame(x = rnorm(N, 0, 2)) |>
  mutate(z = rbinom(N, 1, 0.3*(x>0) + 0.8*(x<=0)),
         y0 = x + rnorm(N,0,2),
         y1 = x + exp(x) + rnorm(N,0,2),
         y = ifelse(z==1, y1, y0))

mean(df$y1) - mean(df$y0) # True ATE = exp(2)
```

```
## [1] 7.416355
```

```r
lm(y ~ x + z, data=df)$coeff['z']
```

```
##        z
## 9.126802
```

```r
# OLS estimate of b2 in the linear approximation
# Y = b0 + b1Z + b2X
```

## g-formula for binary outcome, causal relative risk

- When $Y$ is binary we are often interested in the ATE in the risk ratio scale

$$\tau_{RR} = \frac{E(Y(1))}{E(Y(0))} = \frac{P(Y(1) = 1)}{P(Y(0) = 1)}$$

- Under ignorability and positivity:

$$\tau_{RR} = \frac{E[P(Y = 1|Z = 1, X)]}{E[P(Y = 1|Z = 0, X)]}$$

- Plug-in estimator:

$$\hat{\tau}_{RR} = \frac{\sum_{i=1}^{n} \hat{\mu}_1(X_i)}{\sum_{i=1}^{n} \hat{\mu}_0(X_i)}$$

where $\hat{\mu}_z(X)$ is a an estimate of $P(Y = 1|Z = z, X)$.

# g-formula for binary outcome, causal odds ratio

- We can also define the causal OR as

$$\tau_{OR} = \frac{P(Y(1) = 1)/P(Y(1) = 0)}{P(Y(0) = 1)/P(Y(0) = 0)}$$

- Under ignorability and positivity:

$$\tau_{OR} = \frac{E[P(Y = 1|Z = 1, X)]/E[P(Y = 0|Z = 1, X)]}{E[P(Y = 1|Z = 0, X)]/E[P(Y = 0|Z = 0, X)]}$$

- Plug-in estimator:

$$\hat{\tau}_{OR} = \frac{\left[\sum_{i=1}^{n} \hat{\mu}_1(X_i)\right] / \left[1 - \sum_{i=1}^{n} \hat{\mu}_1(X_i)\right]}{\left[\sum_{i=1}^{n} \hat{\mu}_0(X_i)\right] / \left[1 - \sum_{i=1}^{n} \hat{\mu}_0(X_i)\right]}$$

# Causal ORs and logistic regression

- One can show that if $P$ fulfills
  - ignorability and positivity
  - $\text{logit}(P(Y|X,Z)) = \beta_0 + \beta_1 X + \beta_2 Z$
  - no effect modification:

$$\tau_{\text{OR}} = \frac{P(Y(1)=1|X)/P(Y(1)=0|X)}{P(Y(0)=1|X)/P(Y(0)=0|X)}$$

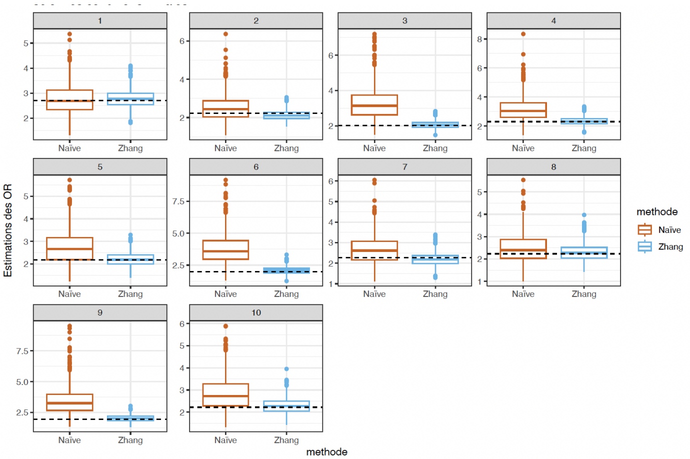
  for each $X$, then

$$\tau_{\text{OR}} = \exp(\beta_2)$$

- However, in general, the coefficient of $Z$ in the logistic regression model of $Y$ given $Z$ and adjustment variables $X$ cannot be interpreted as the log causal OR!

# Simulation study (projet tutoré W.Bakenga, S.Vidil M2 IMB 2022)

- Several scenarios with interactions
- Naive = coefficient $Z$ logistic regression, Zhang = g-formula plug-in estimator



# Conclusions and perspectives

- g-formula application requires
  - identifying variables $X$ for ignorability $\rightarrow$ DAGs, backdoor criterion...
  - estimating the outcome regressions with consistent estimators
  - plug-in estimators
- Many other things to discuss:
  - inference?
  - sensitivity to violation of ignorability?
  - IPTW: alternative identification result, requires estimating consistently the propensity score
  - APTW combines the two approaches and is doubly robust
  - historically, g-formula referred to standardization with time-varying exposure, Robins (1986)

# Sensitivity analysis, Chapter 18.3 Ding (2023)

- Ignorability implies

$$\frac{E[Y(1)|Z=1,X]}{E[Y(1)|Z=0,X]} = 1$$

and

$$\frac{E[Y(0)|Z=0,X]}{E[Y(0)|Z=1,X]} = 1$$

- In general we have

$$\frac{E[Y(1)|Z=1,X]}{E[Y(1)|Z=0,X]} = \epsilon_1(X),$$
$$\frac{E[Y(0)|Z=1,X]}{E[Y(0)|Z=0,X]} = \epsilon_0(X).$$

# Sensitivity analysis (cont'nd)

- It is not difficult to show that

$$\tau = E\left[ZY + (1-Z)\frac{\mu_1(X)}{\epsilon_1(X)}\right] - E[Z\mu_0(X)\epsilon_0(X) + (1-Z)Y]$$

- If we assume $\epsilon_1(X) = \epsilon_1$ and $\epsilon_0(X) = \epsilon_0$, the left hand side is identifiable!
- The idea is to use this formula to calculate $\tau$ for values of $(\epsilon_1, \epsilon_2)$ different from $(1, 1)$ (ignorability)
- If $\tau$ changes abruptly as we go away from $(1, 1)$ our estimate is not robust to violation of ignorability.

# Example: contour plot fonction $\tau = \tau(\epsilon_1, \epsilon_2)$