Generalization of risk ratio using observational data

Julie Josse (Inria) & Erwan Scornet (Sorbonne Université)

Ahmed BOUGHDIRI

May 30, 2024

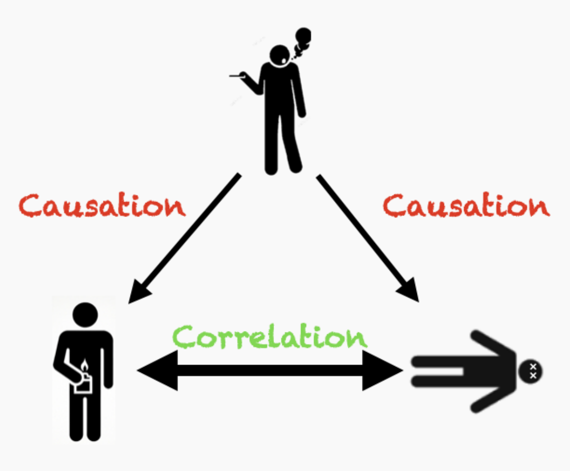
Causal Inference

$\Rightarrow$ **Effect of a policy/intervention/treatment T on an outcome Y**

- What is the effect of smoking on COVID-19 mortality rate ?
- What is the effect of online classes on student performance ?
- How does 4 days work week affect the economy ?

Causal Inference

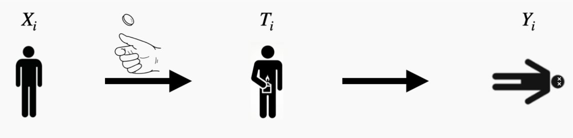
Effect of a policy/intervention/treatment T on an outcome Y



"People who have a lighter tend to have a smaller life expectancy"

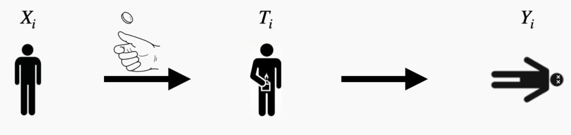
Causal Inference

"People who have a lighter tend to have a smaller life expectancy"



Causal Inference

"People who have a lighter tend to have a smaller life expectancy"



$$\Rightarrow N \textbf{ i.i.d. } (\underbrace{X_i}_{\textbf{covariates}}, \underbrace{T_i}_{\textbf{treatment}}, \underbrace{Y_i}_{\textit{outcome}}) \in \mathbb{R}^d \times \{0,1\} \times \mathbb{R}$$

$$Y = TY^{(1)} + (1 - T)Y^{(0)}.$$

Covariates			Treatment	Outcome	Potential outcomes	
X_1	X_2	X_3	T	Y	$Y^{(0)}$	$Y^{(1)}$
1.1	20	F	1	67	**?**	67
6	45	F	0	83	83	**?**
0	15	M	1	57	**?**	57
	...		...	...	...	...
12	52	M	0	100	100	**?**

Ways to measure the causal effect

$$\mathbb{E}\left[Y^{(1)}\right] \qquad \mathbb{E}\left[Y^{(0)}\right]$$

Expected outcome if treated (1) or control (0)

Ways to measure the causal effect

$$\mathbb{E}\left[Y^{(1)}\right] \qquad \mathbb{E}\left[Y^{(0)}\right]$$

Expected outcome if treated (1) or control (0)

Count the dead

$$\tau_{RR} = \frac{\mathbb{E}\left[Y^{(1)}\right]}{\mathbb{E}\left[Y^{(0)}\right]}$$

Count the living

$$\tau_{SR} = \frac{1 - \mathbb{E}\left[Y^{(1)}\right]}{1 - \mathbb{E}\left[Y^{(0)}\right]}$$

$$\tau_{RD} = \mathbb{E}\left[Y^{(1)}\right] - \mathbb{E}\left[Y^{(0)}\right]$$

Risk Difference

$$\tau_{NNT} = \tau_{RD}^{-1}$$

Number Needed to Treat

Odds Ratio

$$\tau_{OR} = \frac{\mathbb{E}[Y^{(1)}]}{1 - \mathbb{E}[Y^{(1)}]} \left(\frac{\mathbb{E}[Y^{(0)}]}{1 - \mathbb{E}[Y^{(0)}]} \right)^{-1}$$

Ways to measure the causal effect

$$\mathbb{E}\left[Y^{(1)}\right] \qquad \mathbb{E}\left[Y^{(0)}\right]$$

Expected outcome if treated (1) or control (0)

Count the dead

$$\tau_{RR} = \frac{\mathbb{E}\left[Y^{(1)}\right]}{\mathbb{E}\left[Y^{(0)}\right]}$$

Count the living

$$\tau_{SR} = \frac{1 - \mathbb{E}\left[Y^{(1)}\right]}{1 - \mathbb{E}\left[Y^{(0)}\right]}$$

$$\tau_{RD} = \mathbb{E}\left[Y^{(1)}\right] - \mathbb{E}\left[Y^{(0)}\right]$$

Risk Difference

$$\tau_{NNT} = \tau_{RD}^{-1}$$

Number Needed to Treat

Odds Ratio

$$\tau_{OR} = \frac{\mathbb{E}[Y^{(1)}]}{1 - \mathbb{E}[Y^{(1)}]} \left(\frac{\mathbb{E}[Y^{(0)}]}{1 - \mathbb{E}[Y^{(0)}]} \right)^{-1}$$

RESEARCH METHODS & REPORTING

CONSORT 2010 Explanation and Elaboration: updated guidelines for reporting parallel group randomised trials

David Moher,[1] Sally Hopewell,[2] Kenneth F Schulz,[3] Victor Montori,[4] Peter C Gøtzsche,[5] P J Devereaux,[6] Diana Elbourne,[7] Matthias Egger,[8] Douglas G Altman[2]

"[...] both the relative effect (risk ratio (relative risk) or odds ratio) and the absolute effect (risk difference) should be reported (with confidence intervals), as neither the relative measure nor the absolute measure alone gives a complete picture of the effect and its implications."

Risk Ratio

Randomized Controlled Trial (RCT)

(1) $T_i \perp\!\!\!\perp \{Y_i(0), Y_i(1), X_i\} \implies e = \mathbb{P}[T_i = 1]$

Risk Ratio

Randomized Controlled Trial (RCT)

(1) $T_i \perp\!\!\!\perp \{Y_i(0), Y_i(1), X_i\} \implies e = \mathbb{P}[T_i = 1]$

$$\begin{aligned}\tau_{\mathrm{RR}} &= \mathbb{E}[Y(1)]/\mathbb{E}[Y(0)] \\ &\overset{\textbf{(1)}}{=} \mathbb{E}[Y|T=1]/\mathbb{E}[Y|T=0]\end{aligned}$$

Risk Ratio

Randomized Controlled Trial (RCT)

(1) $T_i \perp\!\!\!\perp \{Y_i(0), Y_i(1), X_i\} \implies e = \mathbb{P}[T_i = 1]$

$$\tau_{\mathrm{RR}} = \mathbb{E}[Y(1)]/\mathbb{E}[Y(0)]$$
$$\overset{\textbf{(1)}}{=} \mathbb{E}[Y|T=1]/\mathbb{E}[Y|T=0]$$

Simple mean approach:

$$\hat{\tau}_{\mathrm{RR,N,n}} = \frac{\frac{1}{n_1}\sum_{i=1}^{n} T_i Y_i}{\frac{1}{n-n_1}\sum_{i=1}^{n}(1-T_i)Y_i} \quad \text{with} \quad n_1 = \sum_{i=1}^{n} T_i$$

$$\sqrt{n}\left(\hat{\tau}_{\mathrm{RR,N,n}} - \tau_{\mathrm{RR}}\right) \xrightarrow{d} \mathcal{N}\left(0, V_{\mathrm{RR,N}}\right)$$

where

$$V_{\mathrm{RR,N}} = \tau_{\mathrm{RR}}^2 \left(\frac{\mathrm{Var}\left(Y^{(1)}\right)}{e\mathbb{E}[Y^{(1)}]^2} + \frac{\mathrm{Var}\left(Y^{(0)}\right)}{(1-e)\mathbb{E}[Y^{(0)}]^2}\right)$$

Risk Ratio

Randomized Controlled Trial (RCT)

- **Gold standard** (allocation )
- $T_i \perp\!\!\!\perp \{Y_i(0), Y_i(1), X_i\}$ (1)
- Small sample size: expensive, long, ethical limitations
- Not representative sample:restrictive inclusion criteria

Observational Data

- Low quality Data: contains **confounding bias**
- $\{Y_i(0), Y_i(1)\} \perp\!\!\!\perp T_i \,|\, X_i$ (2)
- Large sample size (registries, biobanks, EHR, claims)
- Representative sample

Risk Ratio

Randomized Controlled Trial (RCT)

- **Gold standard** (allocation )
- $T_i \perp\!\!\!\perp \{Y_i(0), Y_i(1), X_i\}$ (1)
- Small sample size: expensive, long, ethical limitations
- Not representative sample:restrictive inclusion criteria

Observational Data

- Low quality Data: contains **confounding bias**
- $\{Y_i(0), Y_i(1)\} \perp\!\!\!\perp T_i \mid X_i$ (2)
- Large sample size (registries, biobanks, EHR, claims)
- Representative sample

$$
\begin{aligned}
\tau_{\text{RR}} &= \mathbb{E}[Y(1)]/\mathbb{E}[Y(0)] \\
&\overset{(2)}{=} \mathbb{E}[\mathbb{E}[Y|X, T=1]]/\mathbb{E}[\mathbb{E}[Y|X, T=0]]
\end{aligned}
$$

Risk Ratio

Observational Data

$$\begin{aligned} T_i \perp\!\!\!\perp \{Y_i(0), Y_i(1), X_i\} \quad &\implies \quad \{Y_i(0), Y_i(1)\} \perp\!\!\!\perp T_i \mid X_i \\ e = \mathbb{P}\left[T = 1\right] \quad &\implies \quad \underbrace{e(X) = \mathbb{P}\left[T = 1 | X\right]}_{\text{propensity score}} \end{aligned}$$

Risk Ratio

Observational Data

$$
\begin{aligned}
T_i \perp\!\!\!\perp \{Y_i(0), Y_i(1), X_i\} \quad &\implies \quad \{Y_i(0), Y_i(1)\} \perp\!\!\!\perp T_i \mid X_i \\
e = \mathbb{P}\left[T = 1\right] \quad &\implies \quad \underbrace{e(X) = \mathbb{P}\left[T = 1 | X\right]}_{\text{propensity score}}
\end{aligned}
$$

- **Inverse weighted approach**

Risk Ratio

Observational Data

$$
\begin{aligned}
T_i \perp\!\!\!\perp \{Y_i(0), Y_i(1), X_i\} \quad &\Longrightarrow \quad \{Y_i(0), Y_i(1)\} \perp\!\!\!\perp T_i \mid X_i \\
e = \mathbb{P}\left[T = 1\right] \quad &\Longrightarrow \quad \underbrace{e(X) = \mathbb{P}\left[T = 1 | X\right]}_{\text{propensity score}}
\end{aligned}
$$

- **Inverse weighted approach**
- **Regression based approach**

Risk Ratio

Observational Data

$$\begin{aligned} T_i \perp\!\!\!\perp \{Y_i(0), Y_i(1), X_i\} \quad &\implies \quad \{Y_i(0), Y_i(1)\} \perp\!\!\!\perp T_i \mid X_i \\ e = \mathbb{P}[T = 1] \quad &\implies \quad \underbrace{e(X) = \mathbb{P}[T = 1|X]}_{\text{propensity score}} \end{aligned}$$

- **Inverse weighted approach**
- **Regression based approach**
- **Double robust based approach**

Risk Ratio

Observational Data

Inverse weighted approach: We estimate $\hat{e}(X) = \mathbb{P}\left[T = 1 | X\right]$

Risk Ratio

Observational Data

Inverse weighted approach: We estimate $\hat{e}(X) = \mathbb{P}\left[T = 1 | X\right]$

$$\hat{\tau}_{\text{RR,IPW}} = \frac{\sum_{i=1}^{n} \frac{T_i Y_i}{\hat{e}(X_i)}}{\sum_{i=1}^{n} \frac{(1-T_i)Y_i}{1-\hat{e}(X_i)}}$$

Risk Ratio

Observational Data

Inverse weighted approach: We estimate $\hat{e}(X) = \mathbb{P}\left[T = 1 | X\right]$

$$\hat{\tau}_{\mathrm{RR,IPW}} = \frac{\sum_{i=1}^{n} \frac{T_i Y_i}{\hat{e}(X_i)}}{\sum_{i=1}^{n} \frac{(1-T_i)Y_i}{1-\hat{e}(X_i)}} \implies \tau^{\star}_{\mathrm{RR,IPW}} = \frac{\sum_{i=1}^{n} \frac{T_i Y_i}{e(X_i)}}{\sum_{i=1}^{n} \frac{(1-T_i)Y_i}{1-e(X_i)}}$$

Risk Ratio

Observational Data

Inverse weighted approach: We estimate $\hat{e}(X) = \mathbb{P}[T = 1|X]$

$$\hat{\tau}_{\mathrm{RR,IPW}} = \frac{\sum_{i=1}^{n} \frac{T_i Y_i}{\hat{e}(X_i)}}{\sum_{i=1}^{n} \frac{(1-T_i)Y_i}{1-\hat{e}(X_i)}} \implies \tau^{\star}_{\mathrm{RR,IPW}} = \frac{\sum_{i=1}^{n} \frac{T_i Y_i}{e(X_i)}}{\sum_{i=1}^{n} \frac{(1-T_i)Y_i}{1-e(X_i)}}$$

$$\sqrt{n}\left(\tau^{\star}_{\mathrm{RR,IPW}} - \tau_{\mathrm{RR}}\right) \xrightarrow{d} \mathcal{N}\left(0, V_{\mathrm{RR,IPW}}\right)$$

where

$$V_{\mathrm{RR,IPW}} = \tau_{\mathrm{RR}}^2 \left(\frac{\mathbb{E}\left[\frac{(Y^{(1)})^2}{e(X)}\right]}{\mathbb{E}\left[Y^{(1)}\right]^2} + \frac{\mathbb{E}\left[\frac{(Y^{(0)})^2}{1-e(X)}\right]}{\mathbb{E}\left[Y^{(0)}\right]^2} \right).$$

Risk Ratio

Observational Data

Inverse weighted approach: We estimate $\hat{e}(X) = \mathbb{P}\left[T = 1|X\right]$

$$\hat{\tau}_{\mathrm{RR,IPW}} = \frac{\sum_{i=1}^{n} \frac{T_i Y_i}{\hat{e}(X_i)}}{\sum_{i=1}^{n} \frac{(1-T_i)Y_i}{1-\hat{e}(X_i)}} \implies \tau^{\star}_{\mathrm{RR,IPW}} = \frac{\sum_{i=1}^{n} \frac{T_i Y_i}{e(X_i)}}{\sum_{i=1}^{n} \frac{(1-T_i)Y_i}{1-e(X_i)}}$$

$$\sqrt{n}\left(\tau^{\star}_{\mathrm{RR,IPW}} - \tau_{\mathrm{RR}}\right) \xrightarrow{d} \mathcal{N}\left(0, V_{\mathrm{RR,IPW}}\right)$$

where

$$V_{\mathrm{RR,IPW}} = \tau_{\mathrm{RR}}^2 \left(\frac{\mathbb{E}\left[\frac{(Y^{(1)})^2}{e(X)}\right]}{\mathbb{E}\left[Y^{(1)}\right]^2} + \frac{\mathbb{E}\left[\frac{(Y^{(0)})^2}{1-e(X)}\right]}{\mathbb{E}\left[Y^{(0)}\right]^2} \right).$$

Furthermore, if for all i, $0 < m \leq Y_i \leq M$ we have:

$$\left|Bias(\tau^{\star}_{\mathrm{RR,IPW}})\right| \leq \frac{2M^3(1-\eta)^3}{nm^3\eta^3} \quad \left|\mathrm{Var}(\tau^{\star}_{\mathrm{RR,IPW}})\right| \leq \frac{4M^4(1-\eta)^6}{nm^6\eta^4}$$

Risk Ratio

Observational Data

Regression based approach: We estimate $\hat{\mu}_t(X) = \mathbb{E}[Y|T = t, X]$

Risk Ratio

Observational Data

Regression based approach: We estimate $\hat{\mu}_t(X) = \mathbb{E}\left[Y | T = t, X\right]$

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

Risk Ratio

Observational Data

Regression based approach: We estimate $\hat{\mu}_t(X) = \mathbb{E}\left[Y \middle| T = t, X\right]$

$$\hat{\tau}_{\mathrm{RR,G,\ n}} = \frac{\sum_{i=1}^{n} \hat{\mu}_1(X_i)}{\sum_{i=1}^{n} \hat{\mu}_0(X_i)} \implies \tau^{\star}_{\mathrm{RR,G,\ n}} = \frac{\sum_{i=1}^{n} \mu_1^{\star}(X_i)}{\sum_{i=1}^{n} \mu_0^{\star}(X_i)}$$

Risk Ratio

Observational Data

Regression based approach: We estimate $\hat{\mu}_t(X) = \mathbb{E}\left[Y | T = t, X\right]$

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

$$\sqrt{n}\left(\tau^{\star}_{\mathrm{RR,G,\ n}} - \tau_{\mathrm{RR}}\right) \xrightarrow{d} \mathcal{N}\left(0, V_{\mathrm{RR,G}}\right)$$

where

$$V_{\mathrm{RR,G}} = \tau_{\mathrm{RR}}^2 \operatorname{Var}\left(\frac{\mu^{\star}_1(X)}{\mathbb{E}\left[Y^{(1)}\right]} - \frac{\mu^{\star}_0(X)}{\mathbb{E}\left[Y^{(0)}\right]}\right)$$

Risk Ratio

Observational Data

Regression based approach: We estimate $\hat{\mu}_t(X) = \mathbb{E}\left[Y | T = t, X\right]$

$$\hat{\tau}_{\mathrm{RR,G,\ n}} = \frac{\sum_{i=1}^{n} \hat{\mu}_1(X_i)}{\sum_{i=1}^{n} \hat{\mu}_0(X_i)} \implies \tau^{\star}_{\mathrm{RR,G,\ n}} = \frac{\sum_{i=1}^{n} \mu_1^{\star}(X_i)}{\sum_{i=1}^{n} \mu_0^{\star}(X_i)}$$

$$\sqrt{n}\left(\tau^{\star}_{\mathrm{RR,G,\ n}} - \tau_{\mathrm{RR}}\right) \xrightarrow{d} \mathcal{N}\left(0, V_{\mathrm{RR,G}}\right)$$

where

$$V_{\mathrm{RR,G}} = \tau_{\mathrm{RR}}^2 \operatorname{Var}\left(\frac{\mu_1^{\star}(X)}{\mathbb{E}\left[Y^{(1)}\right]} - \frac{\mu_0^{\star}(X)}{\mathbb{E}\left[Y^{(0)}\right]}\right)$$

Assume for all i, $M_0 \geq \mu_0^{\star}(x_i) \geq m_0 > 0$ and $M_1 \geq \mu_1^{\star}(x_i)$. Then, we have:

$$|Bias(\hat{\tau}_{\mathrm{RR,\ G,\ n}})| \leq \frac{2M_1 M_0^2}{n m_0^3} \quad |\operatorname{Var}(\hat{\tau}_{\mathrm{RR,\ G\ ,\ n}})| \leq \frac{2M_0^2 M_1(M_1 + M_0)}{n m_0^6}$$

Risk Ratio

Observational Data

Double robust based approach: We estimate $\hat{\mu}_t(X) = \mathbb{E}\left[Y | T = t, X\right]$ and $\hat{e}(X) = \mathbb{P}\left[T = 1 | X\right]$:

Risk Ratio

Observational Data

Double robust based approach: We estimate $\hat{\mu}_t(X) = \mathbb{E}\left[Y \middle| T = t, X\right]$ and $\hat{e}(X) = \mathbb{P}\left[T = 1 | X\right]$:

$$\hat{\tau}_{\mathrm{RR,AIPW}} = \frac{\sum_{i=1}^n \hat{\mu}_1(X_i) + \frac{T_i(Y_i - \hat{\mu}_1(X_i))}{\hat{e}(X_i)}}{\sum_{i=1}^n \hat{\mu}_0(X_i) + \frac{(1-T_i)(Y_i - \hat{\mu}_0(X_i))}{1-\hat{e}(X_i)}}.$$

Observational Data

Double robust based approach: We estimate $\hat{\mu}_t(X) = \mathbb{E}\left[Y | T = t, X\right]$ and $\hat{e}(X) = \mathbb{P}\left[T = 1 | X\right]$:

$$\hat{\tau}_{\mathrm{RR,AIPW}} = \frac{\sum_{i=1}^{n} \hat{\mu}_1(X_i) + \frac{T_i(Y_i - \hat{\mu}_1(X_i))}{\hat{e}(X_i)}}{\sum_{i=1}^{n} \hat{\mu}_0(X_i) + \frac{(1-T_i)(Y_i - \hat{\mu}_0(X_i))}{1-\hat{e}(X_i)}}.$$

If $\mathbb{E}\left[\left(\hat{\mu}_{(t)}(X_i) - \mu_{(t)}(X_i)\right)^2\right]\mathbb{E}\left[(\hat{e}(X_i) - e(X_i))^2\right] = o\left(\frac{1}{n}\right)$ then:

Risk Ratio

Observational Data

Double robust based approach: We estimate $\hat{\mu}_t(X) = \mathbb{E}\left[Y | T = t, X\right]$ and $\hat{e}(X) = \mathbb{P}\left[T = 1 | X\right]$:

$$\hat{\tau}_{\mathrm{RR,AIPW}} = \frac{\sum_{i=1}^{n} \hat{\mu}_1(X_i) + \frac{T_i(Y_i - \hat{\mu}_1(X_i))}{\hat{e}(X_i)}}{\sum_{i=1}^{n} \hat{\mu}_0(X_i) + \frac{(1-T_i)(Y_i - \hat{\mu}_0(X_i))}{1-\hat{e}(X_i)}}.$$

If $\mathbb{E}\left[\left(\hat{\mu}_{(t)}(X_i) - \mu_{(t)}(X_i)\right)^2\right] \mathbb{E}\left[(\hat{e}(X_i) - e(X_i))^2\right] = o\left(\frac{1}{n}\right)$ then:

$$\sqrt{n}\left(\hat{\tau}_{\mathrm{RR,AIPW}} - \tau_{\mathrm{RR}}\right) \xrightarrow{d} \mathcal{N}\left(0, V_{\mathrm{RR,AIPW}}\right)$$

where

$$\frac{V_{\mathrm{RR,AIPW}}}{\tau_{\mathrm{RR}}^2} = \mathrm{Var}\left(\frac{\mu_1(X)}{\mathbb{E}\left[Y^{(1)}\right]} - \frac{\mu_0(X)}{\mathbb{E}\left[Y^{(0)}\right]}\right) + \mathbb{E}\left[\frac{\mathrm{Var}\left(Y^{(1)} | X\right)}{e(X)\mathbb{E}\left[Y^{(1)}\right]^2}\right] + \mathbb{E}\left[\frac{\mathrm{Var}\left(Y^{(0)} | X\right)}{(1-e(X))\mathbb{E}\left[Y^{(0)}\right]^2}\right].$$

Risk Ratio

Observational Data

Double robust based approach: We estimate $\hat{\mu}_t(X) = \mathbb{E}\left[Y|T = t, X\right]$ and $\hat{e}(X) = \mathbb{P}\left[T = 1|X\right]$:

Risk Ratio

Observational Data

Double robust based approach: We estimate $\hat{\mu}_t(X) = \mathbb{E}\left[Y | T = t, X\right]$ and $\hat{e}(X) = \mathbb{P}\left[T = 1 | X\right]$:

$$\hat{\tau}_{\mathrm{RR,OS}} = \frac{\sum_{i=1}^{n} \hat{\mu}_1(X_i)}{\sum_{i=1}^{n} \hat{\mu}_0(X_i)} \left(1 - \frac{\sum_{i=1}^{n} \hat{\mu}_1(X_i) + \frac{T_i(Y_i - \hat{\mu}_1(X_i))}{\hat{e}(X_i)}}{\sum_{i=1}^{n} \hat{\mu}_0(X_i)} \right) + \frac{\sum_{i=1}^{n} \hat{\mu}_0(X_i) + \frac{(1-T_i)(Y_i - \hat{\mu}_0(X_i))}{\hat{e}(X_i)}}{\sum_{i=1}^{n} \hat{\mu}_0(X_i)}$$

Risk Ratio

Observational Data

Double robust based approach: We estimate $\hat{\mu}_t(X) = \mathbb{E}\left[Y|T=t,X\right]$ and $\hat{e}(X) = \mathbb{P}\left[T=1|X\right]$:

$$\hat{\tau}_{\mathrm{RR,OS}}=\frac{\sum_{i=1}^n \hat{\mu}_1(X_i)}{\sum_{i=1}^n \hat{\mu}_0(X_i)}\left(1-\frac{\sum_{i=1}^n \hat{\mu}_1(X_i)+\frac{T_i(Y_i-\hat{\mu}_1(X_i))}{\hat{e}(X_i)}}{\sum_{i=1}^n \hat{\mu}_0(X_i)}\right)+\frac{\sum_{i=1}^n \hat{\mu}_0(X_i)+\frac{(1-T_i)(Y_i-\hat{\mu}_0(X_i))}{\hat{e}(X_i)}}{\sum_{i=1}^n \hat{\mu}_0(X_i)}$$

If $\mathbb{E}\left[\left(\hat{\mu}_{(t)}(X_i)-\mu_{(t)}(X_i)\right)^2\right]\mathbb{E}\left[(\hat{e}(X_i)-e(X_i))^2\right]=o\left(\frac{1}{n}\right)$ and [...] then:

Risk Ratio

Observational Data

Double robust based approach: We estimate $\hat{\mu}_t(X) = \mathbb{E}\left[Y \mid T = t, X\right]$ and $\hat{e}(X) = \mathbb{P}\left[T = 1 \mid X\right]$:

$$\hat{\tau}_{\mathrm{RR,OS}} = \frac{\sum_{i=1}^n \hat{\mu}_1(X_i)}{\sum_{i=1}^n \hat{\mu}_0(X_i)} \left(1 - \frac{\sum_{i=1}^n \hat{\mu}_1(X_i) + \frac{T_i(Y_i - \hat{\mu}_1(X_i))}{\hat{e}(X_i)}}{\sum_{i=1}^n \hat{\mu}_0(X_i)} \right) + \frac{\sum_{i=1}^n \hat{\mu}_0(X_i) + \frac{(1-T_i)(Y_i - \hat{\mu}_0(X_i))}{\hat{e}(X_i)}}{\sum_{i=1}^n \hat{\mu}_0(X_i)}$$

If $\mathbb{E}\left[\left(\hat{\mu}_{(t)}(X_i) - \mu_{(t)}(X_i)\right)^2\right] \mathbb{E}\left[(\hat{e}(X_i) - e(X_i))^2\right] = o\left(\frac{1}{n}\right)$ and [...] then:

$$\sqrt{n}\left(\hat{\tau}_{\mathrm{RR,OS}} - \tau_{\mathrm{RR}}\right) \xrightarrow{d} \mathcal{N}\left(0, V_{\mathrm{RR,AIPW}}\right)$$

where

$$\frac{V_{\mathrm{RR,AIPW}}}{\tau_{\mathrm{RR}}^2} = \mathrm{Var}\left(\frac{\mu_1(X)}{\mathbb{E}\left[Y^{(1)}\right]} - \frac{\mu_0(X)}{\mathbb{E}\left[Y^{(0)}\right]} \right) + \mathbb{E}\left[\frac{\mathrm{Var}\left(Y^{(1)} \mid X\right)}{e(X)\mathbb{E}\left[Y^{(1)}\right]^2} \right] + \mathbb{E}\left[\frac{\mathrm{Var}\left(Y^{(0)} \mid X\right)}{(1-e(X))\mathbb{E}\left[Y^{(0)}\right]^2} \right].$$

Risk Ratio

Observational Data

Efficiency Theory:

Suppose we want to estimate a function f in a value a:

$$a \approx \hat{a} \implies f(a) \approx f(\hat{a}) \qquad f(a) = f(\hat{a}) + \int_{\hat{a}}^{a} f'(t)\,dt$$

Risk Ratio

Observational Data

Efficiency Theory:

Suppose we want to estimate a function f in a value a:

$$a \approx \hat{a} \implies f(a) \approx f(\hat{a}) \qquad f(a) = f(\hat{a}) + \int_{\hat{a}}^{a} f'(t)\, dt$$

$$f(a) \implies \Psi(\mathcal{P}_{X,T,Y}) = \frac{\mathbb{E}\left[\mathbb{E}\left[Y|\, T = 1, X\right]\right]}{\mathbb{E}\left[\mathbb{E}\left[Y|\, T = 0, X\right]\right]}$$

$$\mathcal{P}_{X,T,Y} \approx \hat{\mathcal{P}}_{X,T,Y} \implies \Psi(\mathcal{P}) \approx \Psi(\hat{\mathcal{P}})$$

Observational Data

Efficiency Theory:

Suppose we want to estimate a function f in a value a:

$$a \approx \hat{a} \implies f(a) \approx f(\hat{a}) \qquad f(a) = f(\hat{a}) + \int_{\hat{a}}^{a} f'(t)\,dt$$

$$f(a) \implies \Psi(\mathcal{P}_{X,T,Y}) = \frac{\mathbb{E}\left[\mathbb{E}\left[Y|T=1,X\right]\right]}{\mathbb{E}\left[\mathbb{E}\left[Y|T=0,X\right]\right]}$$

$$\mathcal{P}_{X,T,Y} \approx \hat{\mathcal{P}}_{X,T,Y} \implies \Psi(\mathcal{P}) \approx \Psi(\hat{\mathcal{P}})$$

$$\Psi(\mathcal{P}) = \Psi(\hat{\mathcal{P}}) + \int \varphi(z;\hat{\mathcal{P}})d\mathcal{P}(z) + R_2(\hat{\mathcal{P}},\mathcal{P})$$

$$\simeq \Psi(\mathbb{P}_n) + \frac{1}{n}\sum_{i=1}^{n}\varphi(Z_i;\mathbb{P}_n) := \hat{\tau}_{\mathrm{RR,OS}}$$

φ can be seen a the derivative of the Ψ according to a distribution $\mathcal{P}$.

Simulations

Observational Data

We generate datasets $(X_i, T_i, Y_i^{(0)}, Y_i^{(1)})$ according to the model:

$$Y_i^{(1)} = m(X_i) + b(X_i) + \varepsilon_i^{(1)} \qquad e(X_i) = \mathbb{P}\left[T_i = 1 | X_i\right]$$
$$Y_i^{(0)} = b(X_i) + \varepsilon_i^{(0)} \qquad \varepsilon^{(t)} \sim \mathcal{N}\left(0, \sigma^2\right)$$

$$\hat{\tau}_{\mathrm{RR,IPW}} \quad \hat{\tau}_{\mathrm{RR,G,\ n}} \quad \hat{\tau}_{\mathrm{RR,AIPW}} \quad \hat{\tau}_{\mathrm{RR,OS}}$$

We need to estimate $\hat{\mu}_t(X) = \mathbb{E}\left[Y | T = t, X\right]$ and $\hat{e}(X) = \mathbb{P}\left[T = 1 | X\right]$:

Random Forest:

- RandomForestClassifier()
- RandomForestRegressor()

Regressions:

- LogisticRegression()
- LinearRegression()

Simulations

Linear and logistic setting:

$$
\begin{aligned}
m(X, V) &= 2, \\
b(X, V) &= \beta_0^\top [X, V], && \text{where } \beta_0 = (-1, 1, -1, -1, 1, 1) \\
e(X) &= (1 + \exp(-\beta_e X))^{-1}, && \text{where } \beta_e = (-0.6, 0.6, -0.6).
\end{aligned}
$$

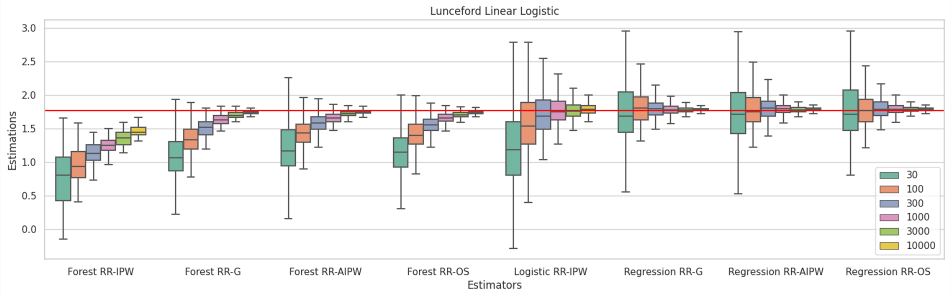


Figure 1: Estimation of the Risk Ratio in a linear and logistic setting

Non-linear and non-logistic setting:

$$m(X) = \sin(\pi X_1 X_2) + 2(X_3 - 0.5)^2 + X_4 + 0.5X_5 - (X_1 + X_2)/4$$
$$b(X) = (X_1 + X_2)/2$$
$$e(X) = \max\{0.1, \min(\sin(\pi X_1), 0.9)\}, \quad \text{where } X \sim \mathrm{Unif}(0,1)^6.$$

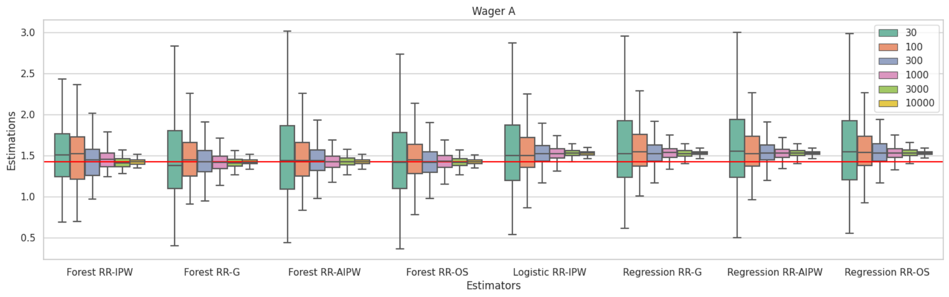


Figure 2: Estimation of the Risk Ratio in a Non-linear and non-logistic setting

Conclusion

Generalization of risk ratio using observational data

Generalization of risk ratio using observational data

- **RCT**

$\tau_{\mathrm{RR,N}}$ $\tau_{\mathrm{RR,HT}}$

Generalization of risk ratio using observational data

- **RCT**

$$\tau_{\mathrm{RR,N}} \qquad \tau_{\mathrm{RR,HT}}$$

- **Observational Data**

$$\hat{\tau}_{\mathrm{RR,IPW}} \quad \hat{\tau}_{\mathrm{RR,G,\ n}} \quad \hat{\tau}_{\mathrm{RR,AIPW}} \quad \hat{\tau}_{\mathrm{RR,OS}}$$

Generalization of risk ratio using observational data

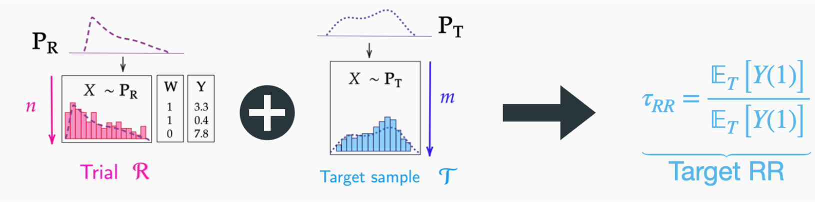
- **RCT**

$$\tau_{\mathrm{RR,N}} \qquad \tau_{\mathrm{RR,HT}}$$

- **Observational Data**

$$\hat{\tau}_{\mathrm{RR,IPW}} \qquad \hat{\tau}_{\mathrm{RR,G,\ n}} \qquad \hat{\tau}_{\mathrm{RR,AIPW}} \qquad \hat{\tau}_{\mathrm{RR,OS}}$$

Generalization:



Thanks for your attention!

Appendix

Causal Inference

$$Y(t) \textbf{ Vs } Y|T = t$$