EPE - Lecture 0
The Two Fundamental Problems of Inference

Sylvain Chabé-Ferret

Toulouse School of Economics, Inra

January 2017

In a nutshell

In this lecture, we are going to study the two fundamentals problems that we face when estimating the effect of an intervention on an outcome. We are also going to study the properties of two intuitive estimators.

Example

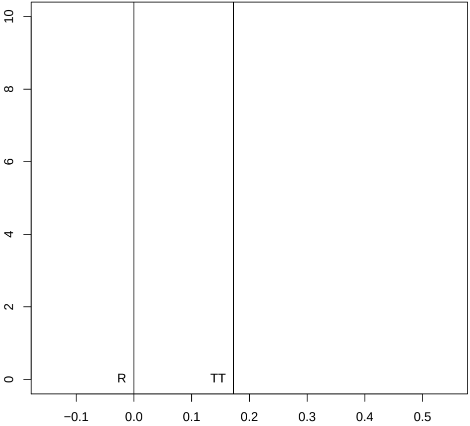


Figure: Our goal

Outline

Outline

The Fundamental Problem of Causal Inference

Intuitive Comparisons and Their Biases

The Fundamental Problem of Statistical Inference

The Fundamental Problem of Causal Inference

TT is unobserved even when $N = \infty$.

The Fundamental Problem of Causal Inference: illustration

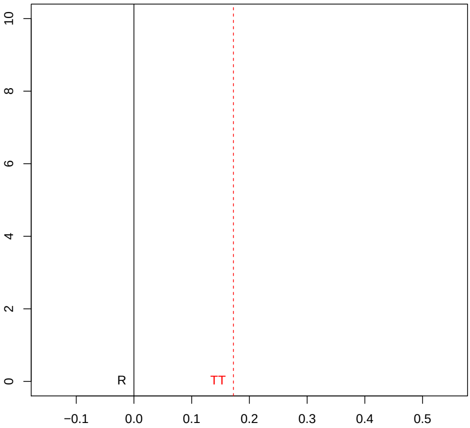


Figure: FPCI

Rubin Causal Model: Components

- Treatment allocation rule
- Potential outcomes
- Swithcing equation

Treatment Allocation Rule

- $D_i = 1$: if unit i receives the treatment
- $D_i = 0$: if unit i does NOT receive the treatment

Example: Sharp Cutoff Rule

$$D_i = \mathbb{1}[Y_i^B \leq \bar{Y}]$$

where $\mathbb{1}[A]$ is the indicator function, taking value 1 when A is true and 0 otherwise.

Numerical Example of Sharp Cutoff Rule

$$
\begin{aligned}
D_i &= \mathbb{1}[y_i^B \leq \bar{y}] \\
y_i^B &= \mu_i + U_i^B \\
\bar{y} &= \log \bar{Y} \\
\mu_i &\sim \mathcal{N}(\bar{\mu}, \sigma_\mu^2) \\
U_i^B &\sim \mathcal{N}(0, \sigma_U^2)
\end{aligned}
$$

The parameter values used in the simulations

	param
barmu	8.00
sigma2mu	0.50
sigma2U	0.28
barY	1500.00
rho	0.90
theta	0.01
sigma2epsilon	0.05
sigma2eta	0.05
delta	0.05
baralpha	0.10

Numerical Example of Sharp Cutoff Rule

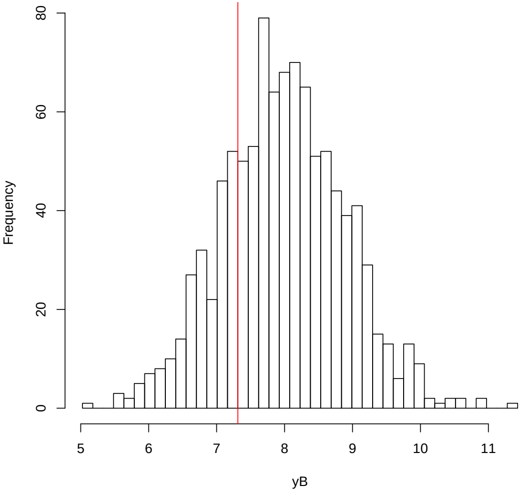


Figure: Histogram of y_B

Numerical Example of Sharp Cutoff Rule

	Ds
0	771
1	229

Table: Treatment allocation with sharp cutoff rule

Other Allocation Rules

Fuzzy cutoff rule $D_i = \mathbb{1}[Y_i^B + V_i \leq \bar{Y}]$

Self-selection rule $D_i = \mathbb{1}[\underbrace{Y_i^1 - Y_i^0 - C_i}_{D_i^*} \geq 0]$

Eligibility & self-select $D_i = \mathbb{1}[D_i^* \geq 0]E_i$

Other rules Awareness, eligibility, application, accepted, shows up

Potential Outcomes

- Y_i^1: outcome we would observe if unit i was given the treatment
- Y_i^0: outcome we would observe if unit i was NOT given the treatment

Potential Outcomes: Numerical Example

$$
\begin{aligned}
y_i^0 &= \mu_i + \delta + U_i^0 \\
U_i^0 &= \rho U_i^B + \epsilon_i \\
\epsilon_i &\sim \mathcal{N}(0, \sigma_\epsilon^2) \\
y_i^1 &= y_i^0 + \alpha_i \\
\alpha_i &= \bar{\alpha} + \theta \mu_i + \eta_i \\
\eta_i &\sim \mathcal{N}(0, \sigma_\eta^2)
\end{aligned}
$$

Potential Outcomes: Numerical Example

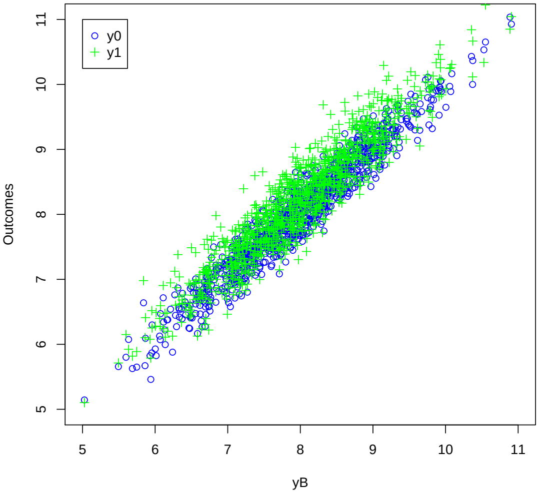


Figure: Potential outcomes

Individual Level Causal Effect

$$\Delta_i^Y = Y_i^1 - Y_i^0$$

Individual Level Causal Effect: Numerical Example

$$\Delta_i^y = \alpha_i = \bar{\alpha} + \theta\mu_i + \eta_i$$

Individual Level Causal Effect: Numerical Example

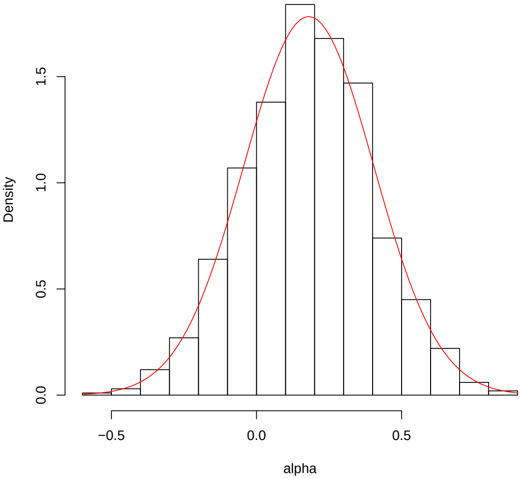


Figure: Histogram of Δ^y

TT: Average Treatment Effect on the Treated

$$\Delta^{Y}_{TT} = \mathbb{E}[\Delta^{Y}_{i} | D_i = 1]$$

$$\Delta^{Y}_{TT_s} = \frac{1}{\sum_{i=1}^{N} D_i} \sum_{i=1}^{N} (Y_i^1 - Y_i^0) D_i$$

TT: Numerical Example

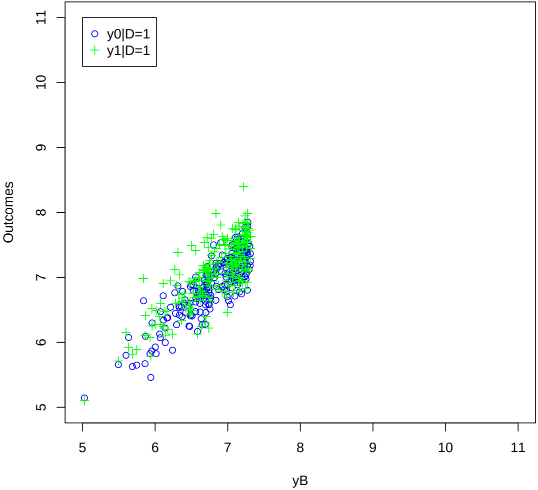


Figure: Potential outcomes

TT: Numerical Example

In the numerical example used in the class, we can derive the value of TT in the population:

$$\Delta^{y}_{TT} = \bar{\alpha} + \theta\bar{\mu} - \theta\frac{\sigma_{\mu}^{2}}{\sqrt{\sigma_{\mu}^{2}+\sigma_{U}^{2}}}\frac{\phi\left(\frac{\bar{y}-\bar{\mu}}{\sqrt{\sigma_{\mu}^{2}+\sigma_{U}^{2}}}\right)}{\Phi\left(\frac{\bar{y}-\bar{\mu}}{\sqrt{\sigma_{\mu}^{2}+\sigma_{U}^{2}}}\right)},$$

where $\bar{y} = \ln(\bar{Y})$ and where ϕ and Φ are respectively the density and the cumulative distribution functions of the standard normal.
The value of TT in our example is 0.17.
The value of TT_s in our example is 0.168

Proof

$$
\begin{aligned}
\Delta^y_{TT} &= \mathbb{E}[\Delta^Y_i | D_i = 1] \\
&= \bar{\alpha} + \theta \mathbb{E}[\mu_i | \mu_i + U^B_i \leq \bar{y}] \\
&= \bar{\alpha} + \theta \left(\bar{\mu} - \frac{\sigma^2_\mu}{\sqrt{\sigma^2_\mu + \sigma^2_U}} \frac{\phi\left(\frac{\bar{y}-\bar{\mu}}{\sqrt{\sigma^2_\mu + \sigma^2_U}}\right)}{\Phi\left(\frac{\bar{y}-\bar{\mu}}{\sqrt{\sigma^2_\mu + \sigma^2_U}}\right)} \right).
\end{aligned}
$$

The second equality follows from the definition of Δ^Y_i and D_i and from the fact that η_i is independent from μ_i and U^B_i. The third equality comes from the formula for the expectation of a censored bivariate normal random variable.

Treatment Effects: Numerical Example

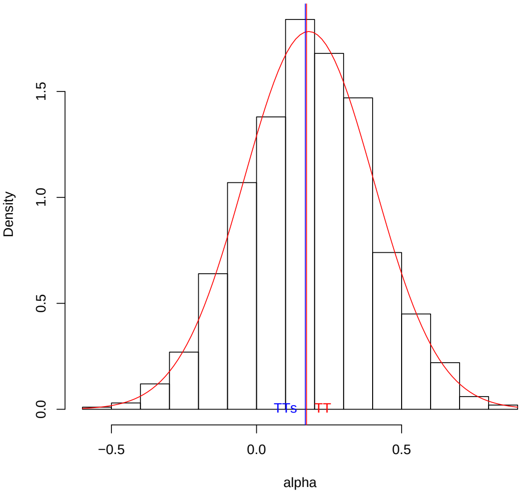


Figure: Histogram of Δ^y

Other Treatment Effects

$$\Delta^Y_{ATE} = \mathbb{E}[\Delta^Y_i]$$

$$\Delta^Y_{ATE_s} = \frac{1}{N}\sum_{i=1}^{N}(Y^1_i - Y^0_i)$$

Less interesting parameter.

ATE: Numerical Example

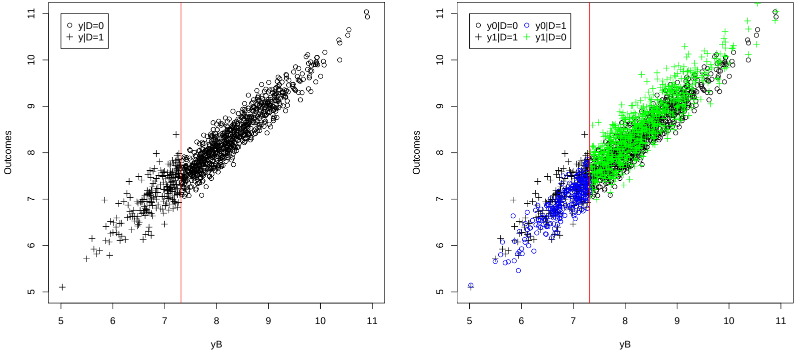
$$\begin{aligned}\Delta^y_{ATE} &= \mathbb{E}[\Delta^y_i] \\ &= \bar{\alpha} + \theta\bar{\mu}.\end{aligned}$$

The value of ATE in our example is 0.18.
The value of ATE_s in our example is 0.178

Switching Equation

$$\begin{aligned} Y_i &= \begin{cases} Y_i^1 & \text{if } D_i = 1 \\ Y_i^0 & \text{if } D_i = 0 \end{cases} \\ &= D_i Y_i^1 + (1 - D_i) Y_i^0 \end{aligned}$$

The Switching Equation: Numerical Example



(a) Observed outcomes (b) Potential outcomes

Figure: Observed and potential outcomes

The Fundamental Problem of Causal Inference

Theorem (Fundamental problem of causal inference)

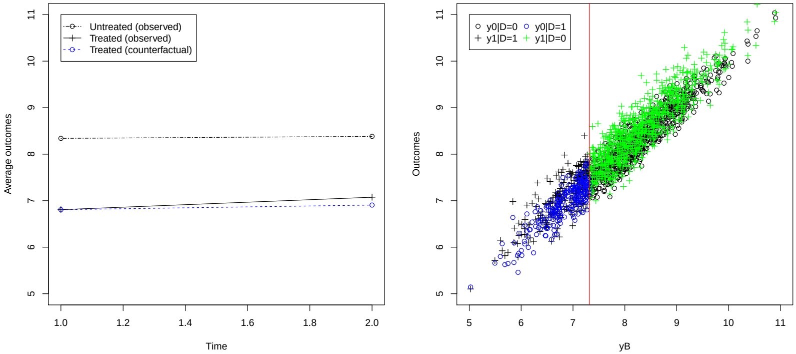
It is impossible to observe TT, either in the population or in the sample.

Why is TT Unobserved? Counterfactual

$$
\begin{aligned}
\Delta^Y_{TT} &= \mathbb{E}[\Delta^Y_i | D_i = 1] \\
&= \mathbb{E}[Y^1_i - Y^0_i | D_i = 1] \\
&= \mathbb{E}[Y^1_i | D_i = 1] - \mathbb{E}[Y^0_i | D_i = 1] \\
&= \mathbb{E}[Y_i | D_i = 1] - \underbrace{\mathbb{E}[Y^0_i | D_i = 1]}_{\textit{Counterfactual}}
\end{aligned}
$$

$$
\Delta^Y_{TT_s} = \frac{1}{\sum_{i=1}^N D_i} \sum_{i=1}^N Y_i D_i - \underbrace{\frac{1}{\sum_{i=1}^N D_i} \sum_{i=1}^N Y^0_i D_i}_{\textit{Counterfactual}}
$$

Why is TT Unobserved? Illustration



(a) Average outcomes

(b) Individual outcomes

Figure: Evolution of average outcomes in the treated and control group

Identification

What can we do to solve the fundamental Problem of Causal Inference?

- Use an observed quantity E (for Estimator) to recover TT
- When there exists E such that, under some assumptions, E=TT, we say that TT is identified under these assumptions
- When E$\neq$TT, we say that E is biased with B=E-TT.

Various Estimators

1. Intuitive comparisons
2. Observational methods
3. Natural experiments
4. Randomized Controlled Trials (RCTs)
5. Controlled experiments
6. Structural models

Outline

Intuitive Comparisons

- With/Without (WW):

$$\Delta^{Y}_{WW} = \mathbb{E}[Y_i | D_i = 1] - \mathbb{E}[Y_i | D_i = 0]$$

$$\Delta^{\hat{Y}}_{WW} = \frac{1}{\sum_{i=1}^{N} D_i} \sum_{i=1}^{N} Y_i D_i - \frac{1}{\sum_{i=1}^{N}(1 - D_i)} \sum_{i=1}^{N} Y_i (1 - D_i)$$

- Before/After (BA):

$$\Delta^{Y}_{BA} = \mathbb{E}[Y_i | D_i = 1] - \mathbb{E}[Y_i^B | D_i = 1]$$

$$\Delta^{\hat{Y}}_{BA} = \frac{1}{\sum_{i=1}^{N} D_i} \sum_{i=1}^{N} Y_i D_i - \frac{1}{\sum_{i=1}^{N} D_i} \sum_{i=1}^{N} Y_i^B D_i$$

Intuitive Comparisons: Illustration

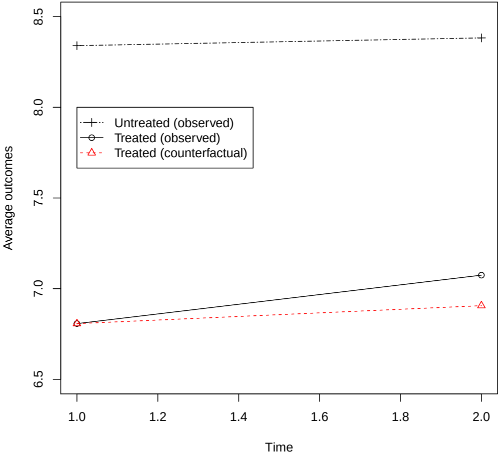


Figure: Evolution of average outcomes in the treated and control group

Intuitive Comparisons: Numerical Example

$$\Delta_{WW}^{\hat{y}} = -1.308$$
$$\Delta_{WW}^{y} = -1.298$$
$$\Delta_{BA}^{\hat{y}} = 0.267$$
$$\Delta_{BA}^{y} = 0.265$$

Intuitive Comparisons: Numerical Example

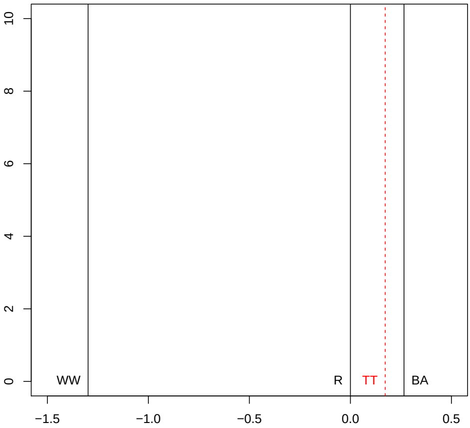


Figure: WW

Biases of Intuitive Comparisons: Selection Bias and Time Trend Bias

$$
\begin{aligned}
\Delta^Y_{SB} &= \Delta^Y_{WW} - \Delta^Y_{TT} \\
&= \mathbb{E}[Y_i^1|D_i = 1] - \mathbb{E}[Y_i^0|D_i = 0] - (\mathbb{E}[Y_i^1|D_i = 1] - \mathbb{E}[Y_i^0|D_i = 1]) \\
&= \mathbb{E}[Y_i^0|D_i = 1] - \mathbb{E}[Y_i^0|D_i = 0] \\
\Delta^Y_{TB} &= \Delta^Y_{BA} - \Delta^Y_{TT} \\
&= \mathbb{E}[Y_i^1|D_i = 1] - \mathbb{E}[Y_i^B|D_i = 1] - (\mathbb{E}[Y_i^1|D_i = 1] - \mathbb{E}[Y_i^0|D_i = 1]) \\
&= \mathbb{E}[Y_i^0|D_i = 1] - \mathbb{E}[Y_i^B|D_i = 1]
\end{aligned}
$$

Selection Bias: Numerical Example

- In the population, Δ^{y}_{SB} =-1.471
- In the sample, $\Delta^{\hat{y}}_{SB}$ =-1.476
- The counterfactual average outcome for the treated is $\frac{1}{\sum_{i=1}^{N} D_i} \sum_{i=1}^{N} D_i y_i^0 = 6.906$
- The average outcome for the untreated that we use to proxy for it is equal to $\frac{1}{\sum_{i=1}^{N}(1-D_i)} \sum_{i=1}^{N}(1-D_i) y_i$ =8.383

Proof

$$
\begin{aligned}
\Delta^{y}_{SB} &= \mathbb{E}[\mu_i | \mu_i + U_i^B \leq \bar{y}] - \mathbb{E}[\mu_i | \mu_i + U_i^B > \bar{y}] \\
&\quad + \rho \left(\mathbb{E}[U_i^B | \mu_i + U_i^B \leq \bar{y}] - \mathbb{E}[U_i^B | \mu_i + U_i^B > \bar{y}] \right) \\
&= \bar{\mu} - \frac{\sigma_\mu^2}{\sqrt{\sigma_\mu^2 + \sigma_U^2}} \frac{\phi\left(\frac{\bar{y}-\bar{\mu}}{\sqrt{\sigma_\mu^2+\sigma_U^2}}\right)}{\Phi\left(\frac{\bar{y}-\bar{\mu}}{\sqrt{\sigma_\mu^2+\sigma_U^2}}\right)} - \left(\bar{\mu} + \frac{\sigma_\mu^2}{\sqrt{\sigma_\mu^2 + \sigma_U^2}} \frac{\phi\left(\frac{\bar{y}-\bar{\mu}}{\sqrt{\sigma_\mu^2+\sigma_U^2}}\right)}{1 - \Phi\left(\frac{\bar{y}-\bar{\mu}}{\sqrt{\sigma_\mu^2+\sigma_U^2}}\right)} \right) \\
&\quad + \rho \left(- \frac{\sigma_U^2}{\sqrt{\sigma_\mu^2 + \sigma_U^2}} \frac{\phi\left(\frac{\bar{y}-\bar{\mu}}{\sqrt{\sigma_\mu^2+\sigma_U^2}}\right)}{\Phi\left(\frac{\bar{y}-\bar{\mu}}{\sqrt{\sigma_\mu^2+\sigma_U^2}}\right)} - \frac{\sigma_U^2}{\sqrt{\sigma_\mu^2 + \sigma_U^2}} \frac{\phi\left(\frac{\bar{y}-\bar{\mu}}{\sqrt{\sigma_\mu^2+\sigma_U^2}}\right)}{1 - \Phi\left(\frac{\bar{y}-\bar{\mu}}{\sqrt{\sigma_\mu^2+\sigma_U^2}}\right)} \right) \\
&= - \frac{\sigma_\mu^2 + \rho\sigma_U^2}{\sqrt{\sigma_\mu^2 + \sigma_U^2}} \left(\frac{\phi\left(\frac{\bar{y}-\bar{\mu}}{\sqrt{\sigma_\mu^2+\sigma_U^2}}\right)}{\Phi\left(\frac{\bar{y}-\bar{\mu}}{\sqrt{\sigma_\mu^2+\sigma_U^2}}\right)} + \frac{\phi\left(\frac{\bar{y}-\bar{\mu}}{\sqrt{\sigma_\mu^2+\sigma_U^2}}\right)}{1 - \Phi\left(\frac{\bar{y}-\bar{\mu}}{\sqrt{\sigma_\mu^2+\sigma_U^2}}\right)} \right)
\end{aligned}
$$

Time Trend Bias: Numerical Example

- In the population, Δ^{y}_{TB} =0.093
- In the sample, $\Delta^{\hat{y}}_{TB}$ =0.099
- The counterfactual average outcome for the treated is $\frac{1}{\sum_{i=1}^N D_i}\sum_{i=1}^N D_i y_i^0 = 6.906$
- The average outcome for the treated before the treatment that we use to proxy for it is equal to $\frac{1}{\sum_{i=1}^N D_i}\sum_{i=1}^N D_i y_i^B$ =6.807

Bias of Intuitive Estimators: Confounders

- Intuitive comparisons are biased because they generally fail to enforce the *ceteris paribus*, every else is held constant, condition.
- Generally, other influences are correlated with treatment allocation
- These influences are called confounders, as they confound the effect of the treatment.
- Because of confounders, correlation $\neq$ causation

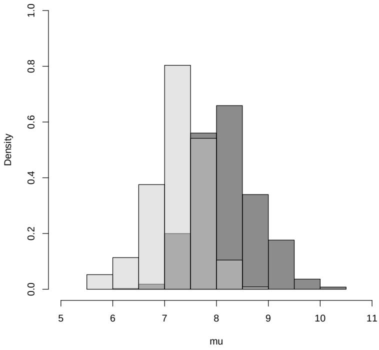
Why is There Selection Bias? Treatment Allocation

- The treatment allocation rule might generate selection bias.
- Example: the threshold eligibility rule conditions on pre-treatment outcomes
- Sicker individuals (or individuals with lower earnings) tend to enter the program
- As sickness and earnings persists, participants tend to exhibit lower outcomes than non participants

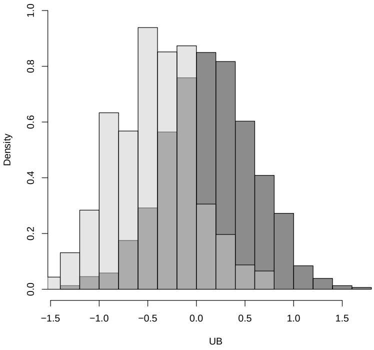
Confounders: Numerical Example

$$
\begin{aligned}
D_i &= \mathbb{1}[y_i^B \leq \bar{y}] \\
y_i^B &= \mu_i + U_i^B \\
y_i^0 &= \mu_i + \delta + \rho U_i^B + \epsilon_i \\
\Delta_{SB}^y &= \mathbb{E}[\mu_i | \mu_i + U_i^B \leq \bar{y}] - \mathbb{E}[\mu_i | \mu_i + U_i^B > \bar{y}] \\
&\quad + \rho \left(\mathbb{E}[U_i^B | \mu_i + U_i^B \leq \bar{y}] - \mathbb{E}[U_i^B | \mu_i + U_i^B > \bar{y}] \right)
\end{aligned}
$$

Selection Bias and Cross-Sectional Confounders



(a) μ_i



(b) U_i^B

Figure: Distribution of the confounding factors in the treated and control group

Selection Bias and Potential Outcomes

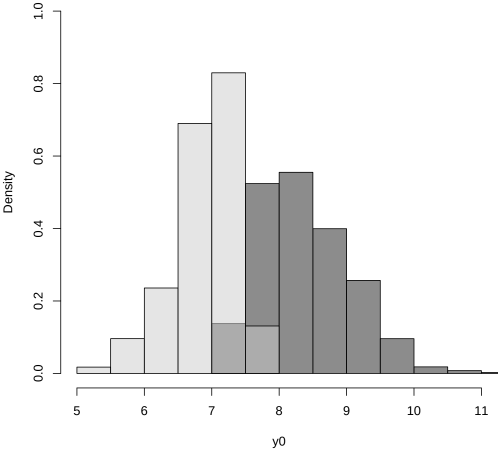


Figure: Distribution of y_i^0 in the treated and control group

Selection Bias and Confounders: Numerical Example

The contribution of μ_i to selection bias is -0.978 while that of U_i^0 is of -0.493.

Why is There Time Trend Bias? Temporal Confounders

- The treatment might be concomittant to other changes in the economy.
- Example: price changes, technology diffusion, business cycle, etc.
- The treatment allocation rule might interact with outcome dynamics and generate regression to the mean
- Example: initially sicker individuals eventually get better even without treatment

Why is There Time Trend Bias? Numerical Example

$$\begin{aligned}\Delta^{y}_{TB} &= \delta + \mathbb{E}[\mu_i|D_i = 1] - \mathbb{E}[\mu_i|D_i = 1] + (\rho - 1)\mathbb{E}[U_i^B|D_i = 1] \\ &= \delta + (\rho - 1)\mathbb{E}[U_i^B|\mu_i + U_i^B \leq \bar{y}]\end{aligned}$$

Time Trend Bias and Confounders: Numerical Example

The contribution of δ to selection bias is 0.05 while that of U_i^B is of 0.043.

Identifying TT Using WW: Assumption

Assumption (No selection bias)

$$\mathbb{E}[Y_i^0 | D_i = 1] = \mathbb{E}[Y_i^0 | D_i = 0].$$

Identifying TT Using WW: Theorem

Theorem

Under this assumption, WW identifies the TT parameter:

$$\Delta^Y_{WW} = \Delta^Y_{TT}.$$

Proof.

$$\begin{aligned}
\Delta^Y_{WW} &= \mathbb{E}[Y_i|D_i = 1] - \mathbb{E}[Y_i|D_i = 0] \\
&= \mathbb{E}[Y_i^1|D_i = 1] - \mathbb{E}[Y_i^0|D_i = 0] \\
&= \mathbb{E}[Y_i^1|D_i = 1] - \mathbb{E}[Y_i^0|D_i = 1] \\
&= \Delta^Y_{TT},
\end{aligned}$$

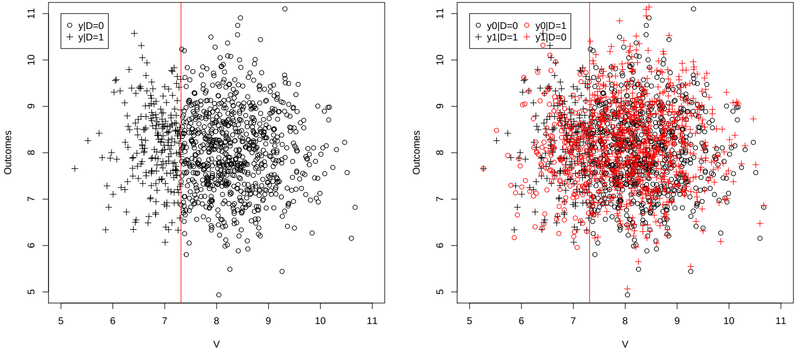
where the second equation uses the switching equation and the third uses the assumption. □

No Selection Bias in The Model Used in the Simulations

$$D_i = \mathbb{1}[V_i \leq \bar{y}]$$
$$V_i \sim \mathcal{N}(\bar{\mu}, \sigma_\mu^2 + \sigma_U^2),$$

where $\bar{y} = \ln(\bar{Y})$.

Absence of Selection Bias: Illustration



(a) Observed outcomes (b) Potential outcomes

Figure: Observed and potential outcomes

TT in the Example Without Selection Bias

In the numerical example used in the class, we can derive the value of TT in the absence of Selection Bias:

$$\begin{aligned}\Delta^{y}_{TT} &= \mathbb{E}[\Delta^{y}_{i}|D_i = 1] \\ &= \bar{\alpha} + \theta\mathbb{E}[\mu_i|V_i \leq \bar{y}] \\ &= \bar{\alpha} + \theta\mathbb{E}[\mu_i] \\ &= \bar{\alpha} + \theta\bar{\mu}.\end{aligned}$$

The value of TT in our example without selection bias is 0.18.
$\Delta^{\hat{y}}_{TT}$ =0.154≈0.133= $\Delta^{\hat{y}}_{WW}$

Placebo Test

In the absence of the treatment, no treatment effect should be detected before the program is implemented

$$\begin{aligned}\Delta_{WW}^{Y^B} &= \mathbb{E}[Y_i^B|D_i = 1] - \mathbb{E}[Y_i^B|D_i = 0] \\ &= 0\end{aligned}$$

Placebo Test: Numerical Example

$$\mathbb{E}[Y_i^B|\hat{D}_i = 1] = 7.965 \approx 7.996 = \mathbb{E}[Y_i^B|\hat{D}_i = 0]$$

Identifying TT Using BA: Assumption

Assumption (No time trend bias)

$$\mathbb{E}[Y_i^0 | D_i = 1] = \mathbb{E}[Y_i^B | D_i = 1].$$

Identifying TT Using BA: Theorem

Theorem

Under this assumption, BA identifies the TT parameter:

$$\Delta^Y_{BA} = \Delta^Y_{TT}.$$

Proof.

$$\begin{aligned}
\Delta^Y_{BA} &= \mathbb{E}[Y_i|D_i = 1] - \mathbb{E}[Y_i^B|D_i = 1] \\
&= \mathbb{E}[Y_i^1|D_i = 1] - \mathbb{E}[Y_i^0|D_i = 1] \\
&= \Delta^Y_{TT}
\end{aligned}$$

□

No Time Trend Bias in The Model Used in the Simulations

$$\delta = 0$$

$$\rho = 0$$

TT in the Example Without Time Trend Bias

$$\Delta^{\hat{y}}_{BA} = 0.173 \approx 0.168 = \Delta^{y}_{TT_s}.$$

Placebo test for the BA estimator

We cannot perform a placebo test using two periods of pre-treatment outcomes for the treated since we have generated only one period of pre-treatment outcome. We will be able to perform this test later in the DID lecture.
We can perfom the placebo test that applies the BA estimator to the untreated. $\hat{\Delta^y_{BA|D=0}}$ =0.007$\approx$ 0

Exercises

1. Install R, Miktex and Rstudio
2. Install knitr package
3. Configure Rstudio with knitr
4. Create a knitr file .Rnw
5. Generate the data with baseline parameter values
6. plot Figure 1 and Table 1
7. plot potential outcomes along y_i^B as in the slides
8. Compute TT_s, $\hat{WW}$ and $\hat{SB}$
9. Compute $\hat{BA}$ and $\hat{TB}$
10. Generate data without selection bias and compute $\hat{WW}$ and $\hat{SB}$
11. Compute placebo test
12. Generate data without time trend bias and compute $\hat{BA}$ and $\hat{TB}$
13. Compute placebo test

Outline

The Fundamental Problem of Statistical Inference

E is unobserved when $N < \infty$.

FPSI: Illustration

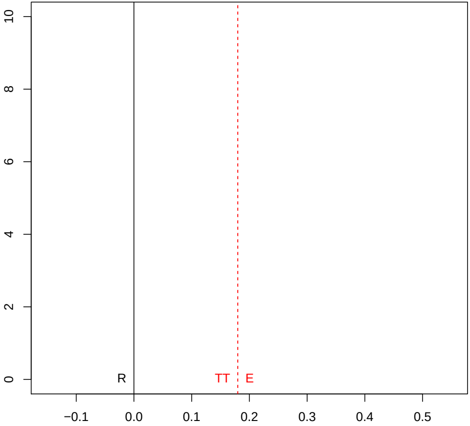


Figure: FPSI

What Do We Observe? Sample Estimator

From a sample of size N, we can form an estimator $\hat{E}$ analogous to the population estimator E.

Sample Estimator: Example of WW

$$\Delta^{Y}_{WW} = \mathbb{E}[Y_i|D_i = 1] - \mathbb{E}[Y_i|D_i = 0]$$

$$\Delta^{\hat{Y}}_{WW} = \frac{1}{\sum_{i=1}^{N} D_i} \sum_{i=1}^{N} Y_i D_i - \frac{1}{\sum_{i=1}^{N}(1 - D_i)} \sum_{i=1}^{N} Y_i(1 - D_i).$$

Illustration

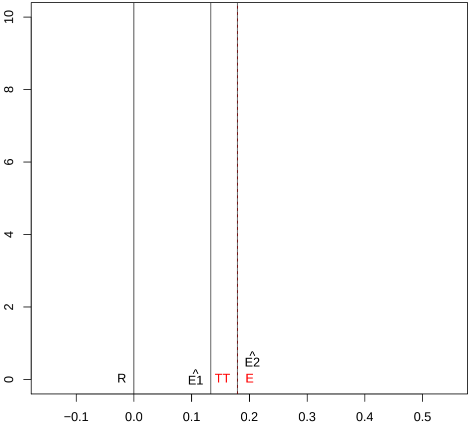


Figure: Sample Estimators

Why Does $\hat{E} \neq E$? Sampling Noise

Even with random sampling, a given sample does not perfectly represent the population. Some parts are overrepresented, some parts underrepresented:

- The treated and untreated groups might have different distributions of the confounders.
- The distribution of the individual treatment effect might differ from the population one.

For each given sample, we are going to make a mistake.
Because of sampling noise, $\hat{E}$ gives a blurry image of E.

Sampling Noise: Illustration

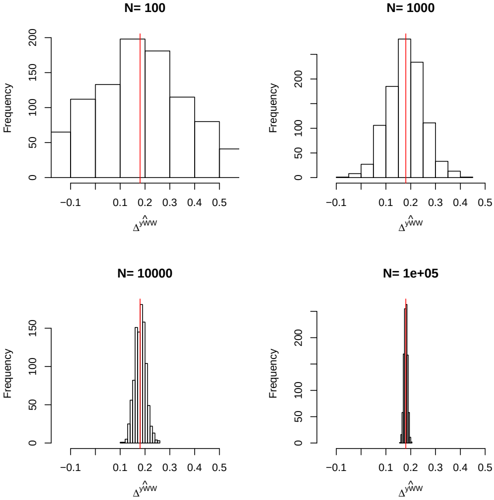


Figure: Distribution of the *WW* estimator over replications of samples of different sizes

What Can We Do to Solve the Fundamental Problem of Statistical Inference?

1. Estimate sampling noise and report it adequately (using confidence intervals)
2. Decrease sampling noise
   - Increasing sample size
   - Stratifying
   - Conditioning

Sampling Noise: a Definition

Definition (Sampling Noise (Symmetric))

Sampling noise 2ϵ is the width of the symmetric interval around TT within which $\delta * 100\%$ of the sample estimators fall, where δ is the confidence level:

$$\Pr(|\hat{E} - TT| \leq \epsilon) = \delta$$

Another Definition of Sampling Noise

Definition (Sampling Noise (Asymmetric))

Sampling noise 2ϵ is the width of the possibly asymmetric interval around TT such that each tail not in the interval contains $(1-\delta/2)*100\%$ of the sample estimators, where δ is the confidence level:

$$2\epsilon = \hat{E}_{\frac{1+\delta}{2}} - \hat{E}_{\frac{1-\delta}{2}},$$

where $\hat{E}_q$ is the q^{th} quantile of the distribution of $\hat{E}$.

What is Sampling Noise? Illustration

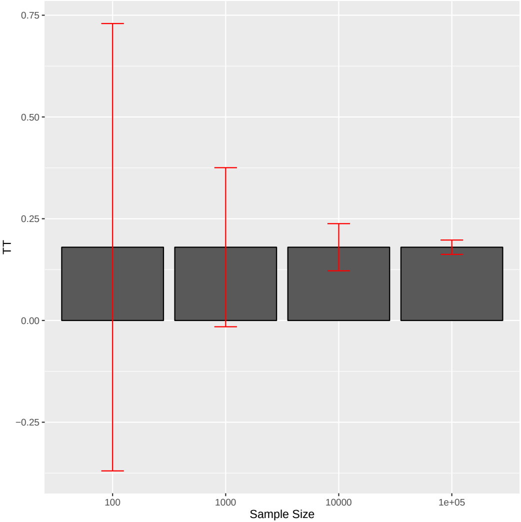
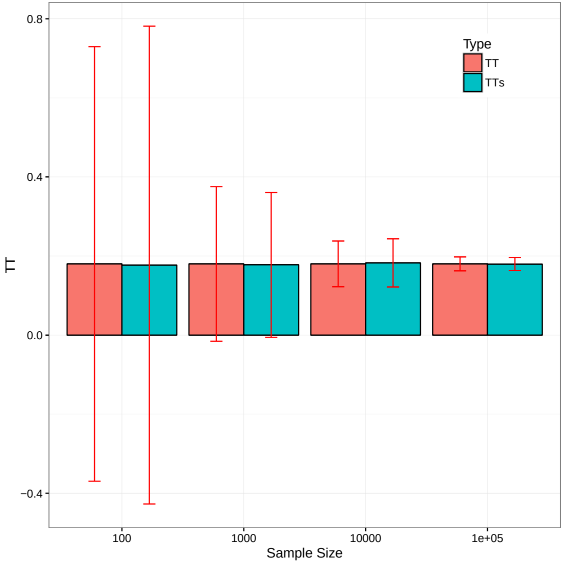


Figure: Symmetric sampling noise of the WW estimator (99% confidence) fot the population TT

Sampling Noise of the Sample Treatment Effect



Sampling Noise of the Sample Treatment Effect

Imbens and Rubin show that sampling noise for WW when estimating TT and TT_s is extremely close, up to a covariance term between potential outcomes that is in general impossible to estimate.

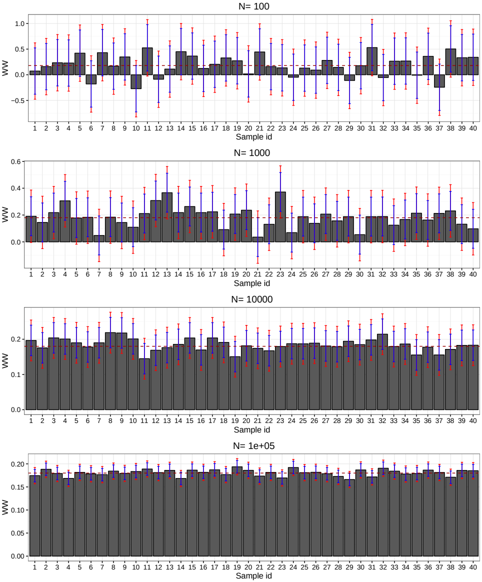
Reporting Sampling Noise Using Confidence Intervals

Theorem (Confidence interval)

For a given level of confidence δ and corresponding level of symmetric sampling noise 2ϵ of the estimator $\hat{E}$ of TT, the confidence interval $\left\{\hat{E}-\epsilon,\hat{E}+\epsilon\right\}$ is such that the probability that it contains TT is equal to δ over sample replications:

$$\Pr(\hat{E}-\epsilon \leq TT \leq \hat{E}+\epsilon) = \delta.$$

Reporting Sampling Noise Using Confidence Intervals: Illustration



How Can We Estimate Sampling Noise?

1. Upper bound using Chebyshev's inequality
2. Approximation using CLT (asymptotic)
3. Approximation using resampling methods (bootstrap)
4. Approximation using Fisher's permutation method

Assumptions: No selection bias

Assumption (No selection bias)

We assume the following:

$$\mathbb{E}[Y_i^0|D_i = 1] = \mathbb{E}[Y_i^0|D_i = 0].$$

Assumptions: Full rank

Assumption (Full rank)

We assume that there is at least one observation in the sample that receives the treatment and one observation that does not receive it:

$$\exists i, j \leq N \text{ such that } D_i = 1 \& D_j = 0.$$

Assumptions: i.i.d. sampling

Assumption (i.i.d. sampling)

We assume that the observations in the sample are identically and independently distributed:

$$\begin{aligned}\forall i, j \leq N,\ i \neq j,\ (Y_i, D_i) &\perp\!\!\!\perp (Y_j, D_j),\\ (Y_i, D_i) &\& (Y_j, D_j) \sim F_{Y,D}.\end{aligned}$$

Assumptions: Finite variances

Assumption (Finite variance of $\Delta^{\hat{Y}}_{WW}$)

We assume that $\mathbb{V}[Y^1|D_i = 1]$ *and* $\mathbb{V}[Y^0|D_i = 0]$ *are finite.*

Chebyshev's Inequality

Theorem (Chebyshev's inequality)

For any unbiased estimator $\hat{\theta}$, sampling noise level 2ϵ and confidence level δ, sampling noise is bounded from above:

$$2\epsilon \leq 2\sqrt{\frac{\mathbb{V}[\hat{\theta}]}{1-\delta}}.$$

In order to use this theorem to gauge the precision of WW, we need to recover values $\mathbb{V}[\Delta^{\hat{Y}}_{WW}]$.

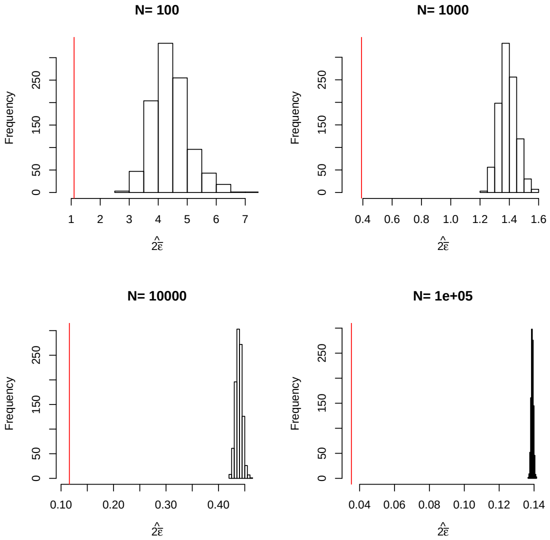
Chebyshev's Upper Bound on Sampling Noise of WW

Theorem (Chebyshev's Upper bound on the sampling noise of $\hat{WW}$)

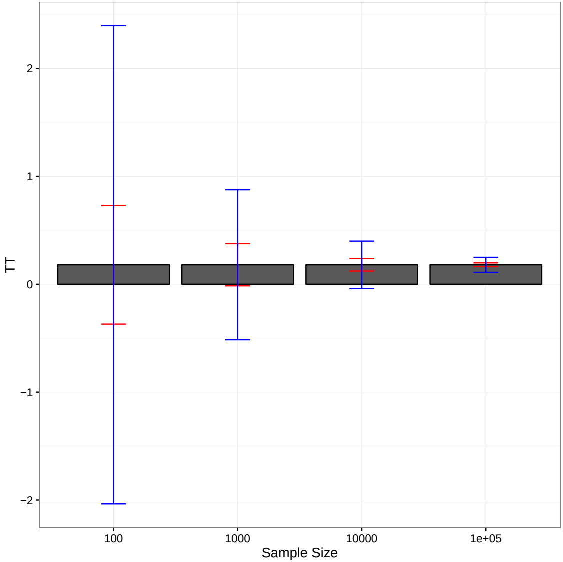
Under Assumption No Selection Bias, Full Rank, i.i.d. and Finite Variances, for a given confidence level δ, the sampling noise of the $\hat{WW}$ estimator is bounded from above:

$$2\epsilon \leq 2\sqrt{\frac{1}{N(1-\delta)}\left(\frac{\mathbb{V}[Y_i^1|D_i=1]}{\Pr(D_i=1)}+\frac{\mathbb{V}[Y_i^0|D_i=0]}{1-\Pr(D_i=1)}\right)} \equiv 2\bar{\epsilon}.$$

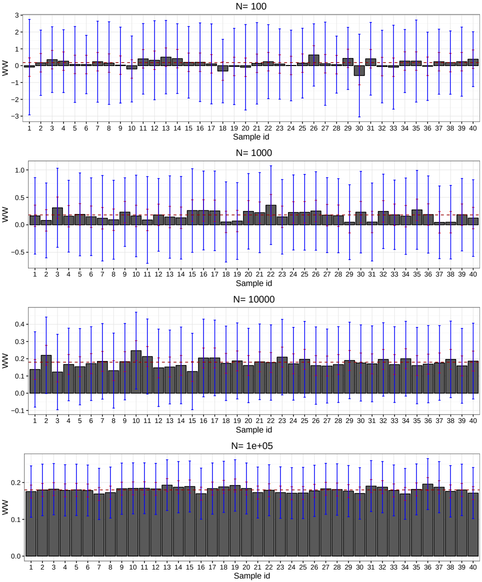
Chebyshev's Upper Bound on Sampling Noise of WW: Illustration



Chebyshev's Upper Bound on Sampling Noise of WW: Illustration



Chebyshev's Upper Bound on Confidence Intervals: Illustration



Problems with Chebyshev's Inequality

- $\bar{\epsilon} > \epsilon$ is too conservative (wide): overestimate sampling noise and underestimate precision
- $\bar{N} > N$ is too large: overestimate sample size

Instead of bounds, why not try to obtain approximations?

Central Limit Theorem

Theorem (Central Limit Theorem)

Let X_i be i.i.d. random variables with $\mathbb{E}[X_i] = \mu$ and $\mathbb{V}[X_i] = \sigma^2$, and define $Z_N = \frac{\frac{1}{N}\sum_{i=1}^{N} X_i - \mu}{\frac{\sigma}{\sqrt{N}}}$, then, for all z we have:

$$\lim_{N\to\infty} \Pr(Z_N \leq z) = \Phi(z),$$

where Φ is the cumulative distribution function of the centered standardized normal.

We say that Z_N converges in distribution to a standard normal random variable, and we denote: $Z_N \xrightarrow{d} \mathcal{N}(0,1)$.

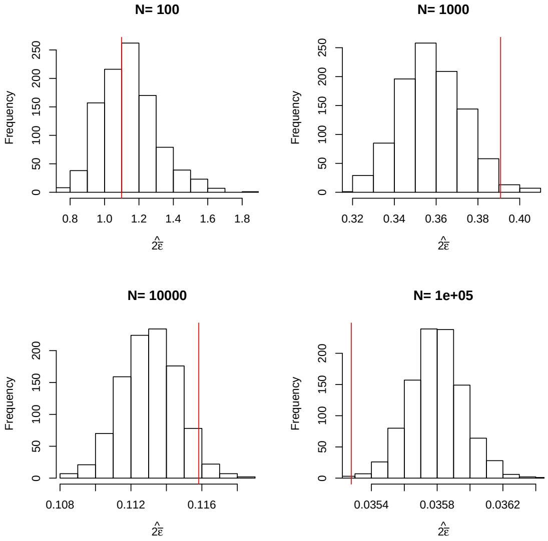
Asymptotic Distribution of WW

Theorem (CLT-based Estimate of Sampling Noise of WW)

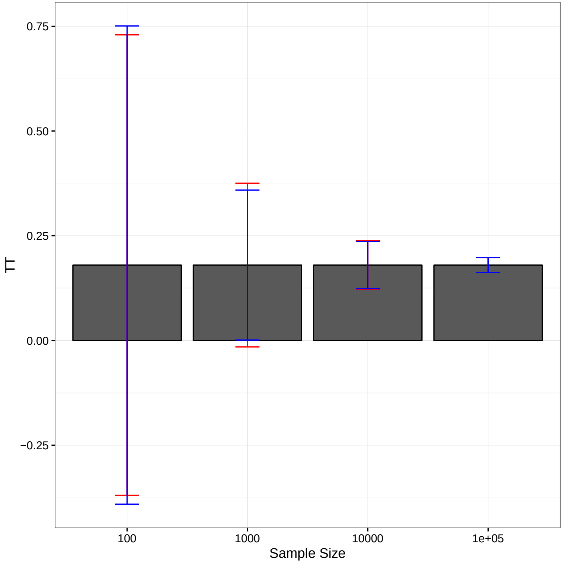
Under Assumptions No Selection Bias, Full Rank, i.i.d. and Finite Variances, for a given confidence level δ and sample size N, the sampling noise of $\hat{WW}$ can be approximated as follows:

$$2\epsilon \approx 2\Phi^{-1}\left(\frac{\delta+1}{2}\right)\frac{1}{\sqrt{N}}\sqrt{\frac{\mathbb{V}[Y_i^1|D_i=1]}{\Pr(D_i=1)}+\frac{\mathbb{V}[Y_i^0|D_i=0]}{1-\Pr(D_i=1)}} \equiv 2\tilde{\epsilon}.$$

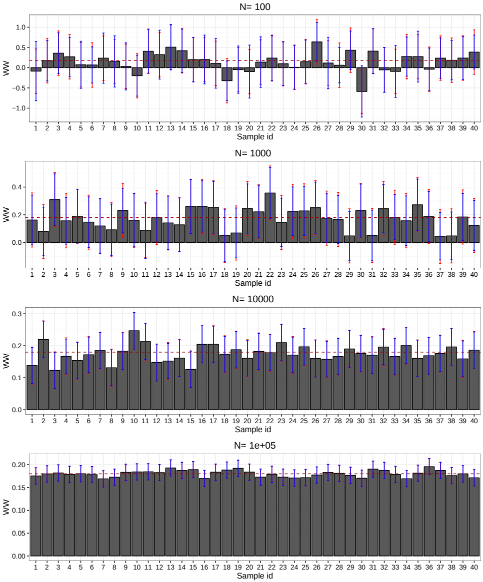
CLT approximation of Sampling Noise of WW: Illustration



CLT approximation of Sampling Noise of WW: Illustration



CLT approximation of Confidence Intervals: Illustration



Proof: outline

In order to prove this theorem, I follow the following procedure:

1. Prove that WW $=$ OLS
2. Prove that OLS is normally asymptotically distributed using
    - CLT
    - Slutsky's Theorem
    - Delta Method

Where it OLS Makes Sense: WW is OLS!

Lemma (WW is OLS)

Under the Full Rank Assumption, the OLS coefficient β in the following regression:

$$Y_i = \alpha + \beta D_i + U_i$$

is the WW estimator:

$$\begin{aligned}
\hat{\beta}_{OLS} &= \frac{\frac{1}{N}\sum_{i=1}^{N}\left(Y_i - \frac{1}{N}\sum_{i=1}^{N} Y_i\right)\left(D_i - \frac{1}{N}\sum_{i=1}^{N} D_i\right)}{\frac{1}{N}\sum_{i=1}^{N}\left(D_i - \frac{1}{N}\sum_{i=1}^{N} D_i\right)^2} \\
&= \Delta^{\hat{Y}}_{WW}.
\end{aligned}$$

From RCM to OLS

$$Y_i = \alpha + \beta D_i + U_i$$

Using RCM, we can also show that:

$$\begin{aligned}
\alpha &= \mathbb{E}[Y_i^0 | D_i = 0] \\
\beta &= \Delta_{TT}^Y \\
U_i &= Y_i^0 - \mathbb{E}[Y_i^0 | D_i = 0] + D_i(\Delta_i^Y - \Delta_{TT}^Y)
\end{aligned}$$

No Selection Bias and the Error Term

Under No Selection Bias, U_i is mean independent of D_i:

$$\begin{aligned}
\mathbb{E}[U_i|D_i = 0] &= \mathbb{E}[Y_i^0|D_i = 0] - \mathbb{E}[Y_i^0|D_i = 0] = 0 \\
\mathbb{E}[U_i|D_i = 1] &= \mathbb{E}[Y_i^0|D_i = 1] - \mathbb{E}[Y_i^0|D_i = 0] \\
&= \mathbb{E}[Y_i^0|D_i = 0] - \mathbb{E}[Y_i^0|D_i = 0] = 0
\end{aligned}$$

RCM, OLS and Heteroskedasticity

Under No Selection Bias, we have:

$$U_i = (1 - D_i)(Y_i^0 - \mathbb{E}[Y_i^0|D_i = 0]) + D_i(Y_i^1 - \mathbb{E}[Y_i^1|D_i = 1])$$

There is heteroskedasticity because the outcomes of the treated and of the untreated have different variances:

$$\begin{aligned}
\mathbb{V}[U_i|D_i = d] &= \mathbb{E}[U_i^2|D_i = d] \\
&= \mathbb{E}[(Y_i^d - \mathbb{E}[Y_i^d|D_i = d])^2|D_i = d] \\
&= \mathbb{V}[Y_i^d|D_i = d]
\end{aligned}$$

Using OLS Heteroskedasticity-Robust Strandard Errors

Use sandwich library and vcovHC command.
Rough estimate

- 95% Sampling Noise $\approx$ 4*s.e.
- 99% Sampling Noise $\approx$ 5*s.e.

Problems with the CLT

- Sometimes imprecise in small samples (if non normal errors)
- Bad in the tails (nonuniform approximation)
- Asymptotic variance sometimes difficult to compute (more complex estimators, more complex autocorrelation structure)

Resampling Methods

Idea: use the sample as a population, draw samples from it, apply the estimator and assess its precision

Jackknife: leave-one-out samples

Bootstrap: sampling with replacement, might increase precision, less robust

Subsampling: sampling without replacement, generally conservative, very robust

Randomization inference: for RCTs, reshuffle the treatment dummy

The Bootstrap

Percentile method

1. Draw a sample k with replacement from the original sample
2. Compute the estimator on the bootstrapped sample: $\hat{E}^*_k$
3. Repeat N_{sim} times
4. Compute the estimate of sampling noise as follows:
   $\hat{E}^*_{\frac{1+\delta}{2}} - \hat{E}^*_{\frac{1-\delta}{2}}$

Other methods (asymptotically pivotal statistics) bring asymptotic refinements.

Validity of the Bootstrap

Theorem (Mammen (1992))

Let $\{X_i : i = 1, \ldots, N\}$ be a random sample from a population. For a sequence of functions g_N and sequences of numbers t_N and σ_N, define $\bar{g}_N = \frac{1}{N}\sum_{i=1}^{N} g_N(X_i)$ and $T_N = (\bar{g}_N - t_N)/\sigma_N$. For the bootstrap sample $\{X_i^* : i = 1, \ldots, N\}$, define $\bar{g}_N^* = \frac{1}{N}\sum_{i=1}^{N} g_N(X_i^*)$ and $T_N^* = (\bar{g}_N^* - \bar{g}_N)/\sigma_N$. Let $G_N(\tau) = \Pr(T_N \leq \tau)$ and $G_N^*(\tau) = \Pr(T_N^* \leq \tau)$, where this last probability distribution is taken over bootstrap sampling replications. Then G_N^* consistently estimates G_N if and only if $T_N \xrightarrow{d} \mathcal{N}(0, 1)$.

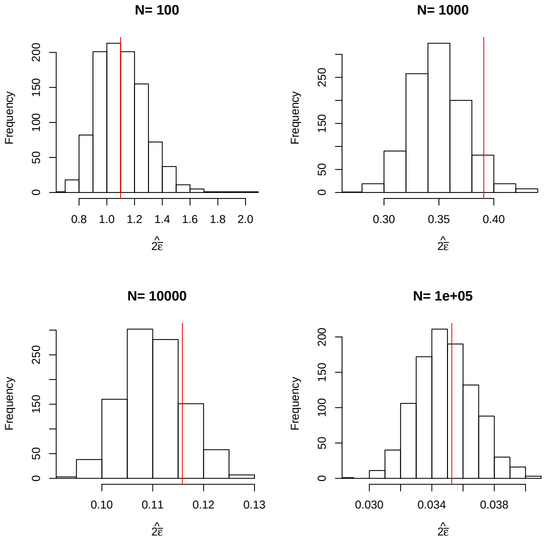
Bootstrapped Estimate of Sampling Noise of WW

Theorem (Bootstrapped Estimate of Sampling Noise of WW)

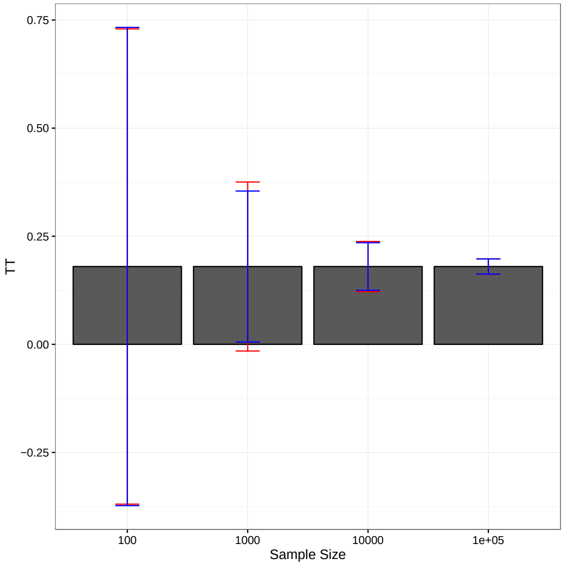
Under Assumptions No Selection Bias, Full Rank, i.i.d. and Finite Variances, for a given confidence level δ and sample size N, the sampling noise of $\hat{WW}$ can be approximated as follows:

$$2\epsilon \approx \hat{E}^{*}_{\frac{1+\delta}{2}} - \hat{E}^{*}_{\frac{1-\delta}{2}} \equiv 2\tilde{\epsilon}^{b}.$$

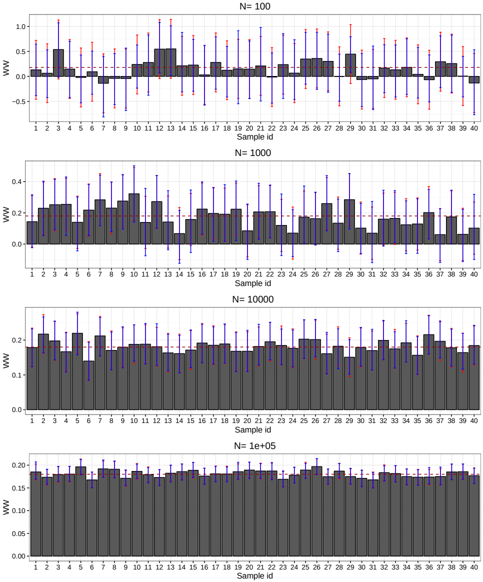
Bootstrapped Estimate of Sampling Noise of WW: Illustration



Bootstrapped Estimate of Sampling Noise of WW: Illustration



Bootstrapped Estimate of Confidence Intervals of WW: Illustration

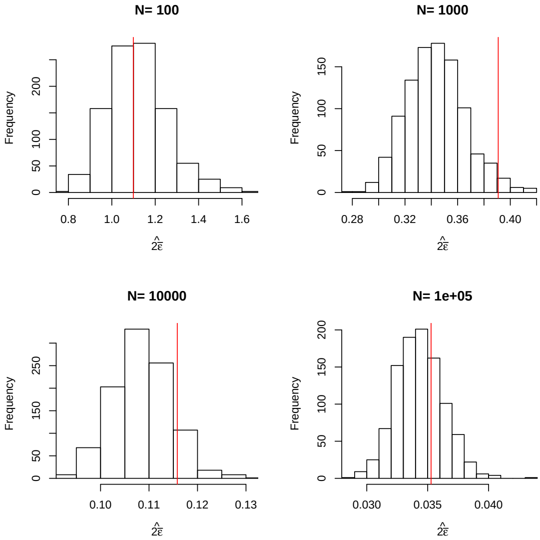


Fisher's Exact Permutation Method

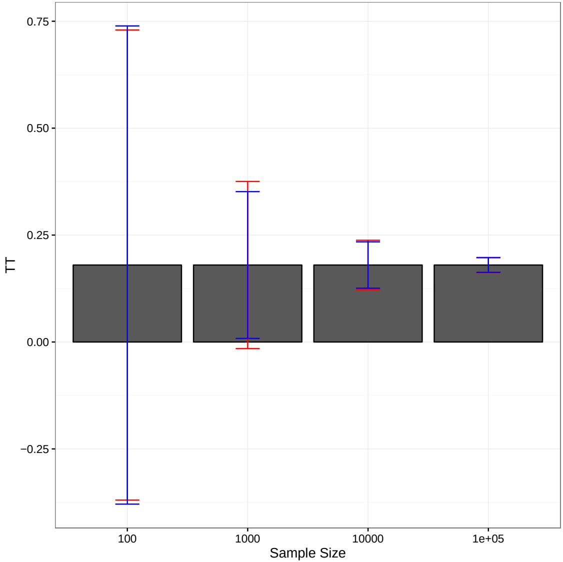
1. Draw a vector k of treatment indicators at random
2. Compute the estimator $\hat{E}$ on the original sample using the new treatment allocation: $\hat{E}^*_k$
3. Repeat N_{sim} times
4. Compute the estimate of sampling noise as follows: $\hat{E}^*_{\frac{1+\delta}{2}} - \hat{E}^*_{\frac{1-\delta}{2}}$

Provides valid exact (finite sample) distribution of any test statistics under the sharp null assumption of all individual treatment effects are zero.

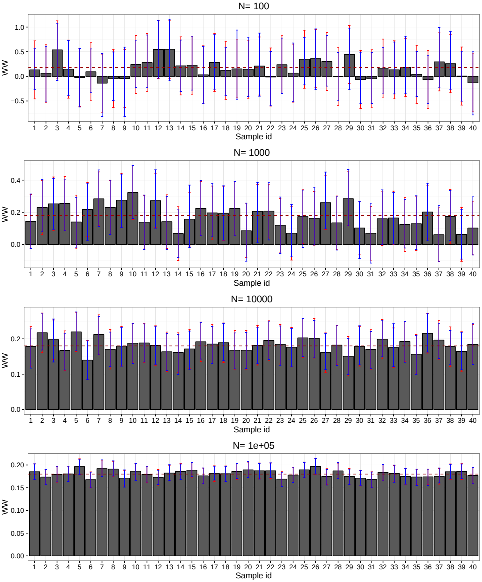
Fisher's Based Estimate of Sampling Noise of WW: Illustration



Fisher's Based Estimate of Sampling Noise of WW: Illustration



Fisher's Based Estimate of Confidence Intervals of WW: Illustration



Decreasing Sampling Noise

- Increasing sample size
- Stratifying
- Conditioning

Increasing Sample Size

Corollary (CLT-based Estimate of Sample Size of WW)

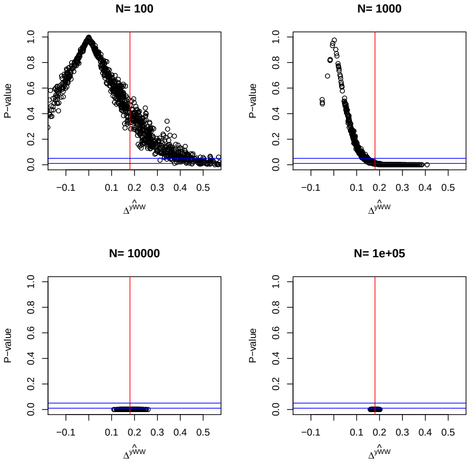
Under Assumptions No Selection Bias, Full Rank, i.i.d. and Finite Variances, for a given confidence level δ, the sample size needed to reach a level of sampling noise 2ϵ with the $\hat{WW}$ estimator can be approximated as follows:

$$N \approx 4\left(\frac{\Phi^{-1}\left(\frac{\delta+1}{2}\right)}{2\epsilon}\right)^2\left(\frac{\mathbb{V}[Y_i^1|D_i=1]}{\Pr(D_i=1)}+\frac{\mathbb{V}[Y_i^0|D_i=0]}{1-\Pr(D_i=1)}\right)\equiv\tilde{N}.$$

The Perils of p-values and Test Statistics

1. Test statistics and p-values are not designed for scientific inquiry but for industrial decisions
2. Test statistics and p-values give a false cutoff sense of confidence
3. Statistically significant treatment effects are biased, all the more so as sampling noise is large
4. Marginally significant results have very low signal to noise ratio

Statistically Significant Results Are Biased



Ioannidis and coauthors show that 80% of results in economics are overestimated y a factor of 2.

Marginally Statistically Significant Results Are Very Noisy

$$\left|\frac{\Delta_{WW}^{\hat{Y}}}{\sigma_{\Delta_{WW}^{\hat{Y}}}}\right| \geq 1.96$$

$$\Rightarrow \left|\frac{\Delta_{WW}^{\hat{Y}}}{2\tilde{\epsilon}}\right| \geq 0.38$$

Typically Powered Studies Are Very Noisy

$$
\begin{aligned}
\frac{\beta_A}{2\epsilon} &\approx \frac{\left(\Phi^{-1}\left(\kappa\right) + \Phi^{-1}\left(1-\alpha\right)\right)\sqrt{\mathbb{V}[\hat{E}]}}{2\Phi^{-1}\left(\frac{\delta+1}{2}\right)\sqrt{\mathbb{V}[\hat{E}]}} \\
&= \frac{\left(\Phi^{-1}\left(\kappa\right) + \Phi^{-1}\left(1-\alpha\right)\right)}{2\Phi^{-1}\left(\frac{\delta+1}{2}\right)}
\end{aligned}
$$

For the usual values for α (0.05) and κ (0.8) and a two-sided t-test, the signal to noise ratio for $\delta = 0.99$ is of 0.54.

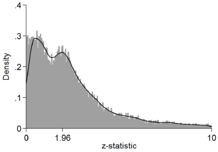
The Consequences of Using pvalues on Science

1. Publication bias and replication crisis
2. Low-powered studies and imprecise estimates

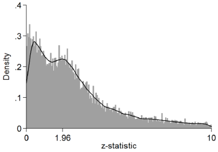
What Are the Margins of Manipulation?

- Choice of specification
- choice of controls
- choice of method
- choice of data
- choice of outcome
- Multiple research teams

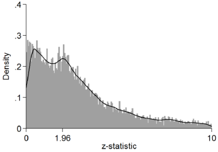
Publication Bias



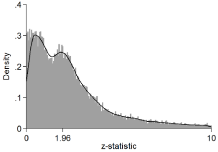
(a) Eye-catchers.



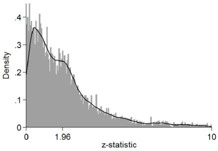
(b) No eye-catchers.



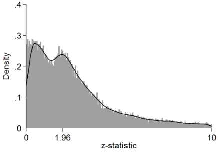
(c) Model.



(d) No model.



(e) Lab. experiments or RCT data.



(f) Other data.

From Brodeur et al., "Star Wars: the Empirics Strike Back", forthcoming, AEJ: Applied.

Publication Bias (continued)

Essay

Why Most Published Research Findings Are False

John P. A. Ioannidis

General Article

PSYCHOLOGICAL SCIENCE

Psychological Science
22(11) 1359–1366

DOI: 10.1177/0956797611417632
http://pss.sagepub.com
SAGE

False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant

Joseph P. Simmons[1], Leif D. Nelson[2], and Uri Simonsohn[1]
[1]The Wharton School, University of Pennsylvania, and [2]Haas School of Business, University of California, Berkeley

See also the excellent discussion on the Marginal Revolution blog.

Replication problem

NATURE | NEWS

Nobel laureate challenges psychologists to clean up their act

Social-priming research needs "daisy chain" of replication.

Ed Yong

03 October 2012

Rights & Permissions

Nobel prize-winner Daniel Kahneman has issued a strongly worded call to one group of psychologists to restore the credibility of their field by creating a replication ring to check each others' results.

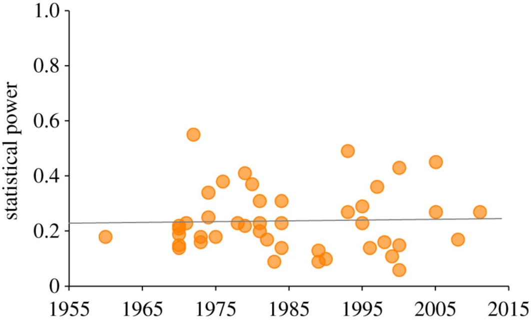
Kahneman, a psychologist at Princeton University in New Jersey, addressed his open e-mail to researchers who work on social priming, the study of how subtle cues can unconsciously influence our thoughts or behaviour. For example, volunteers might walk more slowly down a corridor after seeing words related to old age[1], or fare better in general-knowledge tests after writing down the attributes of a typical professor[2].

Jon Roemer

Daniel Kahneman wants psychologists to spend more time replicating each others' work.

Such tests are widely used in psychology, and Kahneman counts himself as a "general believer" in priming effects. But in his e-mail, seen by *Nature*, he writes that there is a "train wreck looming" for the field, due to a "storm of doubt" about the robustness of priming results.

Imprecise studies



What to do?

1. Ban p-values and significance testing
2. Decrease samling noise
3. Register pre-analysis plans, for example on the AEA website.
4. Use blind data analysis (see this excellent article)
5. Do robustness checks
6. Do a Meta-Analysis
7. Reproduce results

Ban pvalues

NATURE | RESEARCH HIGHLIGHTS: SOCIAL SELECTION

Psychology journal bans *P* values

Test for reliability of results 'too easy to pass', say editors.

Chris Woolston

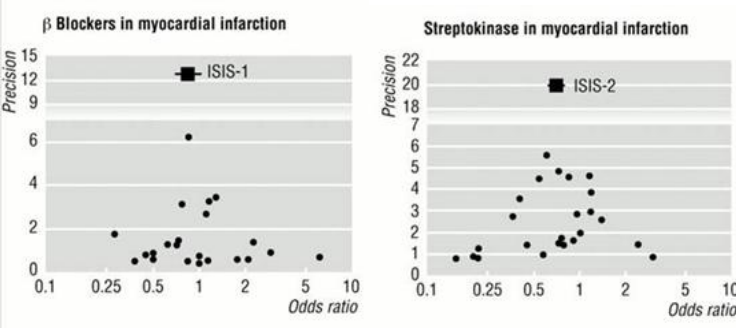
26 February 2015 | Clarified: 09 March 2015

PDF | Rights & Permissions

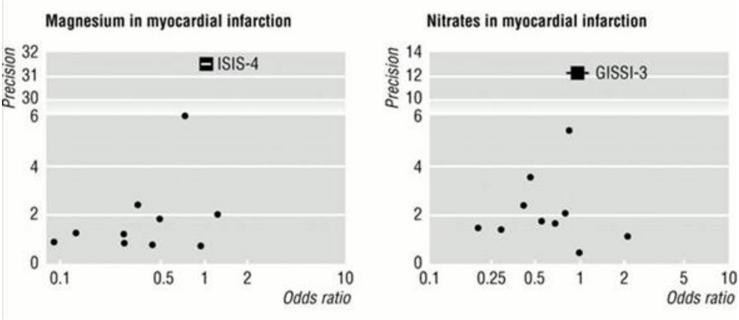
A controversial statistical test has finally met its end, at least in one journal. Earlier this month, the editors of *Basic and Applied Social Psychology* (*BASP*) announced that the journal would no longer publish papers containing *P* values because the statistics were too often used to support lower-quality research[1].

At least, report confidence intervals and compute signal to noise ratio for your results.

Meta-analysis Funnel Graphs Without publication Bias



Meta-analysis Funnel Graphs With publication Bias



Exercises

1. Estimate Cheyshev's upper bound on precision in generated data
2. Estimate CLT-based approximation to sampling noise in generated data
3. Estimate bootstrapped approximation to sampling noise in generated data
4. Estimate Fisher-based approximation to sampling noise in generated data
5. Follow the same steps in the treatment arm of your RCT
6. Advanced: for those who fill like it, look at what happens when the treatment effect is in levels, not logs (Use Monte Carlos). CLT should perform less well in small samples.