

Chapitre 3 : Causalité. Une introduction

Économétrie (ECON0212)

Natalia Bermúdez

HEC Liège

October 2026

Aujourd'hui — introduction à l'inférence causale

- *Causalité* versus *corrélation*
- Les *réalisations potentielles* et le *modèle de Rubin*
- *Expérimentations aléatoires contrôlées* (*Randomized Controlled Trials* = RCT)
- Retour sur l'application empirique reliant *taille d'une classe* et *performance des élèves*

Plan du cours

1. **Corrélation versus causalité**
2. Réalisations potentielles et effet moyen du traitement
3. Le problème de l'inférence causale
4. Expérimentations aléatoires (RCT)
5. Conclusion

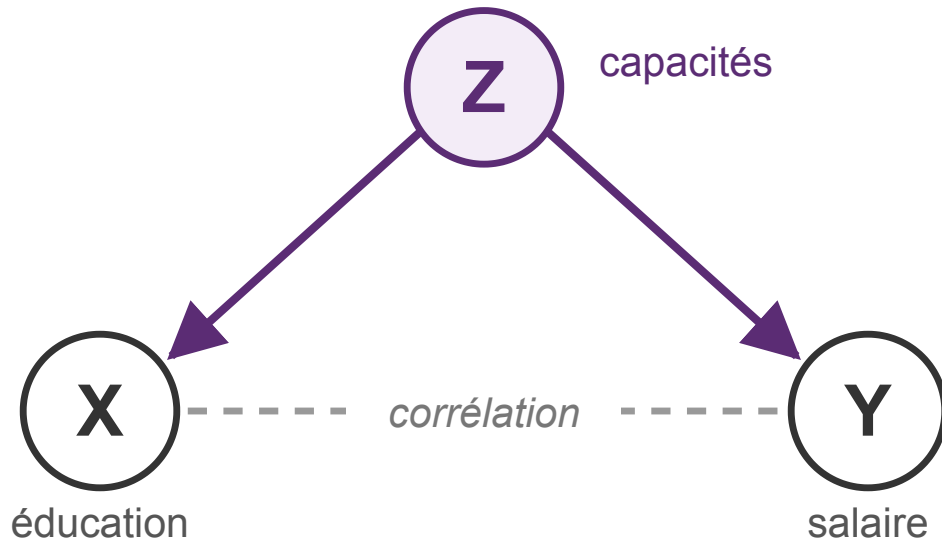
Le concept de causalité

Causalité : de quoi parle-t-on ?

- On dit que **X** cause **Y** si...
- ... nous devons intervenir et *changer* la valeur de **X**, **sans que rien d'autre ne change**...
- ... alors **Y** changerait également, **en conséquence**.
- Le point clé est ce « **sans que rien d'autre ne change** »
 - souvent appelé l'hypothèse **ceteris paribus** (🧐) = *toutes choses égales par ailleurs*.
- ⚠️ Cela ne signifie **PAS** que **X** est le seul facteur qui cause **Y**.

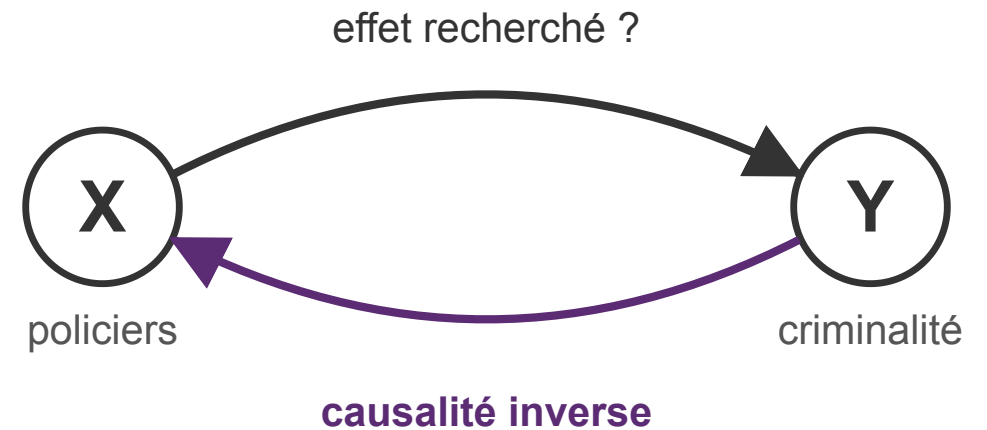
Pourquoi corrélation $\neq$ causalité ?

1. Facteur confondant



Une variable omise Z influence **à la fois** X et Y .

2. Causalité inverse



Y influence aussi X : plus de crimes $\rightarrow$ plus de policiers recrutés.

! Important

À retenir : une corrélation seule n'identifie pas l'effet causal de X sur Y .

Plan du cours

1. Corrélation versus causalité
2. **Réalisations potentielles et effet moyen du traitement**
3. Le problème de l'inférence causale
4. Expérimentations aléatoires (RCT)
5. Conclusion

Réalisations potentielles (*Potential Outcomes Framework*)

Souvent appelé le **modèle causal de Rubin**, en l'honneur du statisticien **Donald Rubin**, qui l'a généralisé et formalisé dans les années 1970.

Idée clé : chaque individu peut être exposé à **plusieurs états de traitement alternatifs** :

- fumer des cigarettes, fumer des cigares, ou ne pas fumer,
- grandir dans un quartier pauvre, de classe moyenne, ou riche,
- être dans une petite ou une grande classe.

Notons la variable de traitement D_i (binaire) :

$$D_i = \begin{cases} 1 & \text{si l'individu } i \text{ est traité} \\ 0 & \text{si l'individu } i \text{ n'est pas traité} \end{cases}$$

Groupe de traitement : tous les individus tels que $D_i = 1$.

Groupe de contrôle : tous les individus tels que $D_i = 0$.

Réalisations potentielles

- Dans ce cadre, chaque individu a deux **réalisations potentielles**, mais une seule **réalisation observée** Y_i :
 - Y_i^1 : réalisation potentielle si l'individu i reçoit le traitement ($D_i = 1$),
 - Y_i^0 : réalisation potentielle si l'individu i ne reçoit pas le traitement ($D_i = 0$).
- Dans la vie réelle, on n'observe que Y_i , qui s'écrit :

$$Y_i = D_i \times Y_i^1 + (1 - D_i) \times Y_i^0$$

→ **Problème fondamental de l'inférence causale** :

- pour n'importe quel individu i , on n'observe qu'**une seule** des deux réalisations potentielles (Holland, 1986).

Réalisations potentielles : le contrefactuel

- La réalisation potentielle non observée existe en principe : on l'appelle le **résultat contrefactuel**.

Groupe	Y_i^1	Y_i^0
Groupe de traitement ($D_i = 1$)	Observé en tant que Y_i	Contrefactuel
Groupe de contrôle ($D_i = 0$)	Contrefactuel	Observé en tant que Y_i

- À partir de ces éléments, on définit l'**effet individuel du traitement** δ_i :

$$\delta_i = Y_i^1 - Y_i^0$$

- δ_i mesure l'**effet causal du traitement** (D_i) sur Y , pour l'individu i .
- Comme cet effet ne peut être observé au niveau individuel, on estime des **moyennes** sur la population.

L'effet moyen du traitement

En anglais : *Average Treatment Effect*, ou **ATE**. C'est l'effet moyen le plus large possible :

$$\begin{aligned}ATE &= \mathbb{E}(\delta_i) \\&= \mathbb{E}(Y_i^1 - Y_i^0) \\&= \mathbb{E}(Y_i^1) - \mathbb{E}(Y_i^0)\end{aligned}$$

- L'ATE mesure simplement l'**effet moyen du traitement sur la population totale**.
- (Voir **annexe**) Il existe d'autres effets causaux d'intérêt.

Exemple : petites vs. grandes classes

Réalisations potentielles de la note (0-10) pour des élèves en petite (Y^1) ou en grande (Y^0) classe :

Élève	Y^1	Y^0	δ
1	5	2	3
2	6	4	2
3	3	6	-3
4	5	4	1
5	10	8	2
6	2	4	-2
7	5	2	3
8	6	4	2
9	2	9	-7
10	8	2	6
Moyenne	5,2	4,5	0,7

$$\begin{aligned}ATE &= \mathbb{E}(\delta) \\&= \mathbb{E}(Y^1) - \mathbb{E}(Y^0) \\&= 5,2 - 4,5 = 0,7\end{aligned}$$

→ en **moyenne**, être dans une classe à petit effectif plutôt qu'à grand effectif améliore la note de 0,7 point.

⚠ l'effet du traitement varie d'un élève à l'autre !

Plan du cours

1. Corrélation versus causalité
2. Réalisations potentielles et effet moyen du traitement
3. **Le problème de l'inférence causale**
4. Expérimentations aléatoires (RCT)
5. Conclusion

Le problème de l'inférence causale

- Pour chaque individu i , on n'observe qu'**un seul des deux résultats potentiels** : Y_i^1 ou Y_i^0 . On ne peut donc pas observer directement δ_i , ni calculer directement l'ATE.
- En revanche, on peut comparer les **résultats observés** des individus traités et non traités. Dans la population :

$$SDR = \mathbb{E}(Y_i \mid D_i = 1) - \mathbb{E}(Y_i \mid D_i = 0)$$

C'est la **Simple Différence des moyennes des Réalisations (SDR)**.

- Dans un échantillon, on estime cette différence par :

$$\widehat{SDR} = \underbrace{\frac{1}{N_T} \sum_{i:D_i=1} Y_i}_{\text{moyenne, groupe traité}} - \underbrace{\frac{1}{N_C} \sum_{i:D_i=0} Y_i}_{\text{moyenne, groupe contrôle}}$$

SDR : quand l'assignation n'est pas aléatoire

Imaginons que le·la directeur·ice observe les deux résultats potentiels de chaque élève et choisisse le traitement qui **maximise sa note** :

Élève	Y^1	Y^0	δ	Y (observé)	D
1	5	2	3	5	1
2	6	4	2	6	1
3	3	6	-3	6	0
4	5	4	1	5	1
5	10	8	2	10	1
6	2	4	-2	4	0
7	5	2	3	5	1
8	6	4	2	6	1
9	2	9	-7	9	0
10	8	2	6	8	1

Le·la directeur·ice assigne le traitement ($D = 1$) quand il **augmente la note** ($Y^1 > Y^0$).

$\Rightarrow$ L'assignation dépend directement des **résultats potentiels**.

SDR : quand l'assignation n'est pas aléatoire

Élève	Y	D	δ
1	5	1	3
2	6	1	2
3	6	0	-3
4	5	1	1
5	10	1	2
6	4	0	-2
7	5	1	3
8	6	1	2
9	9	0	-7
10	8	1	6
Moyenne			0,7

La simple différence des réalisations moyennes :

$$\begin{aligned} SDR &= \frac{5 + 6 + 5 + 10 + 5 + 6 + 8}{7} \\ &\quad - \frac{6 + 4 + 9}{3} \\ &\approx 6,43 - 6,33 \approx 0,1 \end{aligned}$$

- La SDR est **inférieure** à l'ATE : $0,1 < 0,7$!
- La comparaison « **naïve** » (SDR) n'identifie pas, en général, **l'effet causal** du traitement.
- Souvent fait par des journalistes, des politiciens, de mauvais scientifiques (mais pas vous 😊).

Problème des comparaisons naïves

Réécrivons la SDR pour faire apparaître l'effet individuel du traitement (δ_i) :

$$\begin{aligned} SDR &= \mathbb{E}(Y_i^1 \mid D_i = 1) - \mathbb{E}(Y_i^0 \mid D_i = 0) \\ &= \mathbb{E}(Y_i^0 + \delta_i \mid D_i = 1) - \mathbb{E}(Y_i^0 \mid D_i = 0) \end{aligned}$$

Simplifions en supposant que **l'effet du traitement est constant** entre individus : pour tout i , $\delta_i = \delta$.
Alors :

$$SDR = \delta + \mathbb{E}(Y_i^0 \mid D_i = 1) - \mathbb{E}(Y_i^0 \mid D_i = 0)$$

Et comme $ATE = \mathbb{E}(\delta_i) = \mathbb{E}(\delta) = \delta$ (par hypothèse) :

$$SDR = ATE + \underbrace{\mathbb{E}(Y_i^0 \mid D_i = 1) - \mathbb{E}(Y_i^0 \mid D_i = 0)}_{\text{biais de sélection}}$$

Annexe : Quand on relâche l'hypothèse d'effet du traitement constant, un **second biais** apparaît (**sélection sur les gains** ou **hétérogénéité de l'effet**). Ex: Directeur·ice assigne le traitement en fonction de δ_i .



Plan du cours

1. Corrélacion versus causalité
2. Réalisations potentielles et effet moyen du traitement
3. **Le problème de l'inférence causale: solution ?**
4. Expérimentations aléatoires (RCT)
5. Conclusion

La randomisation résout le problème de l'inférence causale !

- **Expérimentation aléatoire** : vous assignez les individus **aléatoirement** à un groupe de traitement et de contrôle.
- Dans ce cas, le traitement est **indépendant** des réalisations potentielles : $(Y_i^1, Y_i^0) \perp D_i$.
- Conséquences :
 - $\mathbb{E}(Y_i^0 \mid D_i = 1) = \mathbb{E}(Y_i^0 \mid D_i = 0) \Rightarrow$ le **biais de sélection est nul** ;
 - $\mathbb{E}(Y_i^1 \mid D_i = 1) = \mathbb{E}(Y_i^1 \mid D_i = 0) \Rightarrow$ pas de **sélection sur les gains** (annexe), même si les effets δ_i sont hétérogènes.
- Avec une assignation aléatoire,

$$SDR = \mathbb{E}(Y_i \mid D_i = 1) - \mathbb{E}(Y_i \mid D_i = 0) = \mathbb{E}(Y_i^1) - \mathbb{E}(Y_i^0) = ATE$$

👉 On peut directement estimer l'ATE à partir des données !

Exercice 1 : pourquoi randomiser ?

15:00

Mêmes individus, mêmes réalisations potentielles (Y_i^0, Y_i^1) : seule l'assignation au traitement change.

Base A : assignation non aléatoire

1. Importer `toy_data_2.csv` (dossier `exercices/`, sur LOLA)

```
import delimited "toy_data_2.csv", clear
```

Variables : `group_dummy` (D_i), `y0` (Y_i^0), `y1` (Y_i^1)

2. Construire :

- réalisation observée : $Y_i = D_i Y_i^1 + (1 - D_i) Y_i^0$
- effet individuel : $\delta_i = Y_i^1 - Y_i^0$

3. Calculer :

- ATE : `summarize delta`
- SDR : `tabstat y, by(group_dummy)`

Base B : assignation aléatoire

1. Importer `toy_data_random.csv`

```
import delimited "toy_data_random.csv", clear
```

2. Refaire l'étape 2 de la base A.

⚠ Recalculer Y_i , car D_i a changé.

3. Refaire l'étape 3 et comparer les résultats avec la base A.

Exercice 1 : pourquoi randomiser ?

Optionnel : pour A et B, calculer le **biais de sélection** et la **sélection sur les gains**, puis vérifier la décomposition de la SDR.

	A : non aléatoire	B : aléatoire
ATE	?	?
SDR	?	?
Biais = SDR – ATE	?	?

À discuter :

- Y a-t-il un biais dans les bases A et B ?
- Le biais est-il important ? Pourquoi ou pourquoi pas ?

Plan du cours

1. Corrélation versus causalité
2. Réalisations potentielles et effet moyen du traitement
3. Le problème de l'inférence causale
4. **Expérimentations aléatoires (RCT)**
5. Conclusion

Expérimentations aléatoires

- Souvent appelées « **R**andomized **C**ontrolled **T**rials » (RCT).
- Les premières RCT sont menées en médecine (18^e et 19^e siècles).
- Popularisées au début du 20^e siècle par les statisticiens **J. Neyman** et **R. A. Fisher**.
- Depuis, méthode de plus en plus influente, devenue progressivement une **méthode d'évaluation des politiques publiques**.
- En économie, le **prix Nobel 2019** a été attribué à trois promoteurs des RCT : **Abhijit Banerjee, Esther Duflo et Michael Kremer**, « for their experimental approach to alleviating global poverty ».

Retour sur la taille de la classe et les performances des élèves

$$\text{note en math}_i = \beta_0 + \beta_1 \text{ taille de la classe}_i + e_i$$

Mais pourquoi b_1^{OLS} ne permet-il d'établir qu'une corrélation, et non une relation causale ? **Trois sources de biais :**

1. **Tri des élèves** : sélection dans les écoles à classes de tailles différentes — si les parents pensent que les petites classes sont meilleures, ils y inscriront leurs enfants.
2. **Tri des enseignants** : ils/elles peuvent préférer les écoles à classes plus petites (plus facile d'y enseigner) ; s'il y a de la concurrence pour ces postes, les enseignants les plus qualifiés y seront avantagés.
3. **Effet de localisation** : les grandes classes sont plus courantes en zones urbaines riches et denses ; les petites classes, en zones rurales plus pauvres.

Une RCT prendrait en compte tous ces biais !

Le projet d'expérimentation STAR

Tennessee **S**tudent/**T**eacher **A**chievement **R**atio Experiment (cf. Krueger, 1999)

- Financé par l'État du Tennessee, pour un coût total d'environ 12 millions de dollars.
- L'expérience a débuté lors de l'année scolaire 1985-1986 et a duré quatre ans.
- 11 600 **élèves et leurs enseignants** ont été **assignés au hasard** à l'un de ces groupes:
 1. **Petite classe** (groupe traité): 13 à 17 élèves par enseignant,
 2. **Classe normale** (groupe de contrôle) : 22 à 25 élèves,
- La randomisation des classes a eu lieu **au sein des écoles**.
- Les compétences en mathématiques et en lecture ont été testées.

L'effet du traitement sous forme de régression

$$Y_i = D_i Y_i^1 + (1 - D_i) Y_i^0$$

En développant et en regroupant, avec $\delta_i = Y_i^1 - Y_i^0$:

$$Y_i = Y_i^0 + D_i (Y_i^1 - Y_i^0) = Y_i^0 + D_i \delta_i$$

En supposant un effet constant, $\delta_i = \delta$ pour tout i , puis en ajoutant $\mathbb{E}[Y_i^0] - \mathbb{E}[Y_i^0] = 0$:

$$Y_i = \underbrace{\mathbb{E}[Y_i^0]}_{\beta_0} + \delta D_i + \underbrace{Y_i^0 - \mathbb{E}[Y_i^0]}_{e_i}$$

Pourquoi l'estimateur MCO de δ est-il sans biais ? Il faut que : $\mathbb{E}[e_i \mid D_i] = 0$ Cela signifie que Y_i^0 est, en moyenne, le même dans le **groupe traité** et le **groupe de contrôle** :

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

⇒ **pas de biais de sélection (L'assignation aléatoire garantit cette condition).**

Le projet d'expérimentation STAR : régression

Modèle estimé : $\text{score en math}_i = \beta_0 + \delta \text{small}_i + e_i$

Source	SS	df	MS	Number of obs	=	3,363
				F(1, 3361)	=	28.31
Model	66685.1688	1	66685.1688	Prob > F	=	0.0000
Residual	7916753.58	3,361	2355.47563	R-squared	=	0.0084
				Adj R-squared	=	0.0081
Total	7983438.75	3,362	2374.60998	Root MSE	=	48.533

math	Coefficient	Std. err.	t	P> t	[95% conf. interval]	
small	8.922239	1.676867	5.32	0.000	5.634457	12.21002
_cons	484.4145	1.14938	421.46	0.000	482.1609	486.668

Le projet d'expérimentation STAR : régression

$$b_0 = \mathbb{E}[Y_i^0]$$

```
tabstat math if small==0, stat(mean)
```

Variable	Mean
math	484.4145

```
tabstat math if small==1, stat(mean)
```

Variable	Mean
math	493.3367

$$\hat{\delta} = \mathbb{E}[Y_i \mid D_i = 1] - \mathbb{E}[Y_i \mid D_i = 0]$$

$$\hat{\delta} = 493,3367 - 484,4145 = 8,9222$$

Limites des expérimentations aléatoires

Les RCT ont une forte **validité interne**, c'est-à-dire qu'elles permettent d'établir le lien causal de manière très convaincante.

Elles présentent toutefois certaines limites :

- Les RCT sont souvent **infaisables** :
 - **coûteuses**,
 - **questions éthiques** : certains *traitements* ne peuvent simplement pas être assignés,
 - elles **prennent du temps**.
- **Interprétation** des résultats :
 - **validité externe** : dans quelle mesure les résultats d'une RCT se généralisent-ils à d'autres contextes (pays, populations, ...) ?
 - il peut être difficile de mettre au jour les mécanismes en jeu,
 - randomisation imparfaite, attrition, ...

Exercice 2

08:00

1. Sélectionner les élèves de CE1 (*first grader*) dans une classe à taille petite ou normale.
2. Calculer la note moyenne en mathématiques pour les classes de taille normale et petite, ainsi que la différence entre les deux.
3. Créer une variable binaire **treatment**, égale à 1 si l'élève est dans le groupe de traitement (petite classe), et à 0 s'il/elle est dans le groupe de contrôle (classe normale).
4. Régresser la note en mathématiques sur la variable de traitement. Les résultats sont-ils cohérents avec la question 2 ?
5. Comment interpréter les coefficients ?

Plan du cours

1. Corrélation versus causalité
2. Réalisations potentielles et effet moyen du traitement
3. Le problème de l'inférence causale
4. Expérimentations aléatoires (RCT)
5. **Conclusion**

La suite ?

On ne peut pas toujours s'appuyer sur des RCT pour se simplifier la vie.

⇒ il faut trouver une manière de faire de l'inférence causale à partir de *données réelles* (et non de *données expérimentales*).

Deux sources de sélection à traiter différemment :

- **caractéristiques observables** : régression multivariée (prochain chapitre !).
- **caractéristiques non observables** : *différences-de-différences* (à la fin de ce cours !) et *Regression Discontinuity Design* (en master en économie !)

Où en sommes-nous de notre quête de la causalité ?

- ✅ Comment gérer les données ? Ouvrez-les, ordonnez-les, visualisez-les...
- ⚠️ Comment résumer une relation entre plusieurs variables ? Régression linéaire simple... *à suivre*
- ✅ **Qu'est-ce que la causalité ?**
- ❌ Comment faire si nous n'observons qu'une partie de la population ?
- ❌ Nos résultats sont-ils uniquement dus au hasard ?
- ❌ Comment trouver de l'exogénéité en pratique ?

À la semaine prochaine !

CONTACT

natalia.bermudez@uliege.be

Un grand merci à [Malka Guillot](#) pour avoir partagé les supports de ce cours, ainsi qu'à [Florian Oswald](#) et à toute l'équipe de ScPoEconometrics pour leur livre et leurs [ressources](#).

Annexe

Effet moyen sur les traités et sur les non-traités

D'autres effets moyens conditionnels peuvent être intéressants :

Average **T**reatment on the **T**reated (**ATT**)

$$\begin{aligned} ATT &= \mathbb{E}(\delta_i \mid D_i = 1) \\ &= \mathbb{E}(Y_i^1 - Y_i^0 \mid D_i = 1) \\ &= \mathbb{E}(Y_i^1 \mid D_i = 1) - \mathbb{E}(Y_i^0 \mid D_i = 1) \end{aligned}$$

L'ATT mesure l'**effet moyen du traitement, conditionnel au fait d'être dans le groupe de traitement**.

Exemple : l'effet de participer à un programme de formation (*traitement*) pour celles et ceux qui y ont participé (*groupe de traitement*).

Remarque : dans la majorité des cas, $ATE \neq ATT \neq ATU$! *retour*

Average **T**reatment on the **U**ntreated (**ATU**)

$$\begin{aligned} ATU &= \mathbb{E}(\delta_i \mid D_i = 0) \\ &= \mathbb{E}(Y_i^1 - Y_i^0 \mid D_i = 0) \\ &= \mathbb{E}(Y_i^1 \mid D_i = 0) - \mathbb{E}(Y_i^0 \mid D_i = 0) \end{aligned}$$

L'ATU mesure l'**effet moyen du traitement, conditionnel au fait d'être dans le groupe de contrôle**.

Exemple : l'effet d'aller dans une école privée (*traitement*) pour des élèves venant d'une école publique (*groupe de contrôle*).

Problèmes des comparaisons naïves (suite)

On relâche maintenant l'hypothèse que l'effet du traitement est constant entre individus.

Après **de longs calculs** que l'on peut sauter, la SDR se décompose ainsi :

$$SDR = ATE + \underbrace{\mathbb{E}(Y_i^0 \mid D_i = 1) - \mathbb{E}(Y_i^0 \mid D_i = 0)}_{\text{biais de sélection}} + \underbrace{(1 - \pi)(ATT - ATU)}_{\text{effet hétérogène du traitement}}$$

où $1 - \pi$ est la part des individus dans le groupe de contrôle.

Une nouvelle source de biais apparaît, liée à une possible **hétérogénéité de l'effet individuel du traitement** δ_i (**sélection sur les gains**).

retour

Problèmes des comparaisons naïves (suite)

Deux sources de biais coexistent dans la SDR dès que l'effet du traitement est hétérogène :

- **Biais de sélection** : celles et ceux qui vont à l'université ont probablement de meilleures performances cognitives de base (qu'ils y aillent ou non).
- **Hétérogénéité de l'effet du traitement** ou **sélection sur les gains** : l'effet d'être dans une petite classe peut différer d'un élève à l'autre. Celles et ceux qui vont à l'université amélioreraient probablement plus leurs compétences cognitives en y étudiant que les autres, car plus motivé·e·s.

Dans l'exemple du/de la directeur·ice ($\pi = 0,7$), le biais de sélection seul n'explique pas l'écart entre SDR et ATE :

$$\underbrace{0,1}_{SDR} \approx \underbrace{0,7}_{ATE} + \underbrace{(3,71 - 6,33)}_{\text{biais de sélection}} + \underbrace{0,3 \times (2,71 - (-4))}_{\text{effet hétérogène}}$$

[retour](#)

Exercice 3 : données STAR

10:00

1. Importer les données `star_data.dta` (`use "star_data.dta", replace`).
 - Cette documentation explique les variables (les données ont été *reformatées* : ne pas tenir compte des « k », « 1 », etc. dans les noms de variables).
2. Quelle est l'unité d'observation ? Quelle variable contient : (i) l'affectation (aléatoire) de classe, (ii) le niveau scolaire de l'élève, (iii) la variable d'intérêt ?
3. Combien y a-t-il d'observations ? Pourquoi autant, si 11 598 élèves ont participé ? Pourquoi y a-t-il tant de valeurs manquantes ? À quoi correspondent-elles ?
4. Ne gardez que les observations pour lesquelles `read`, `math`, `ladder` et `tethnicity` ne sont pas manquants.
5. Vérifions que la randomisation a bien été faite, en construisant une table de **balancing checks** : calculez, par groupe de niveau et de traitement, les pourcentages de filles, d'afro-américains, et de bénéficiaires de repas gratuits (*free lunch*).

Exercice 3 : Le projet d'expérimentation STAR

Nous venons de voir que, dans une RCT, l'effet moyen du traitement (ATE) s'obtient en calculant la différence de résultats entre groupe de traitement et groupe de contrôle.

Concentrons-nous sur :

- un groupe de traitement : **petites classes**,
- un groupe de contrôle : **classes normales**,
- une classe : **maternelle** (k).

test	moyenne (classe normale)	moyenne (petite classe)	ATE
math	484.41	493.34	8.93
read	435.76	441.15	5.39

Quelle est l'interprétation de ces ATE ?

Régression avec une variable binaire : formellement

Rappel du modèle : $\text{score en math}_i = b_0 + \delta \text{small}_i + e_i$

$$\begin{aligned}\mathbb{E}[\text{score} \mid \text{small}_i = 0] &= \mathbb{E}[b_0 + \delta \text{small}_i + e_i \mid \text{small}_i = 0] \\ &= b_0 + \delta \mathbb{E}[\text{small}_i \mid \text{small}_i = 0] + \mathbb{E}[e_i \mid \text{small}_i = 0] = b_0\end{aligned}$$

$$\begin{aligned}\mathbb{E}[\text{score} \mid \text{small}_i = 1] &= \mathbb{E}[b_0 + \delta \text{small}_i + e_i \mid \text{small}_i = 1] \\ &= b_0 + \delta \mathbb{E}[\text{small}_i \mid \text{small}_i = 1] + \mathbb{E}[e_i \mid \text{small}_i = 1] = b_0 + \delta\end{aligned}$$

$$\begin{aligned}ATE &= \mathbb{E}[\text{score} \mid \text{small}_i = 1] - \mathbb{E}[\text{score} \mid \text{small}_i = 0] \\ &= b_0 + \delta - b_0 = \delta\end{aligned}$$

On le savait déjà, mais on comprend maintenant *pourquoi* c'est vrai 🙌.