Causalité

Analyse tabulaire multivariée

Visseho Adjiwanou, PhD.

Département de Sociologie - UQAM

28 Janvier 2019

Plan de présentation

1. Questions causales en sciences sociales et terminologie
2. Effets causaux et contrefactuel
3. Essais contrôlés randomisés (*Randomized controlled trials*) et causalité
4. Causalité à partir des données observationnelles
5. Exercices

Introduction

Introduction

- Dans ce chapitre, nous considérons la causalité, l'un des concepts les plus centraux des sciences sociales quantitatives.
- Une grande partie de la recherche en sciences sociales s'intéresse aux effets causaux de diverses politiques et autres facteurs sociétaux.
- Par exemple:
  - Est-ce que le vaccin A protège contre la maladie X?
  - Les classes de petite taille augmentent-elles les résultats des tests standardisés des élèves?
  - Les soins de santé universels amélioreraient-ils la santé et les finances des pauvres?
  - L'éducation réduit-elle le nombre d'enfants?
  - La rémunération des gens sur Wikipedia augmentera-elle leur productivité?
  - Est-ce que l'augmentation du salaire minimum réduit l'activité économique?

Questions de recherche

Questions de recherche

- Une question de recherche est au cœur d'un projet de recherche, d'une étude ou d'une revue de littérature.
- Il concentre l'étude, détermine la méthodologie et guide toutes les étapes de la recherche, de l'analyse et de la production de rapports.
- Peut être **associatif** ou **causal**

Exemple 1

1. Le salaire minimum augmente-t-il le taux de chômage?
   - Le taux de chômage a augmenté après l'augmentation du salaire minimum.
   - Le taux de chômage aurait-il augmenté si l'augmentation du salaire minimum n'avait pas eu lieu?

Exemple 2

2. La race/l'ethnie a-t-elle une incidence sur les perspectives d'emploi?
   - Mohamed a postulé pour un emploi mais ne l'a pas obtenu.
   - Mohamed aurait-il trouvé un travail s'il était blanc (avait un nom européen)?

Exemple 3

3. Est-ce que fumer cause une maladie coronarienne?
   - Jean, fumeur, a eu une maladie coronarienne.
   - Est-ce que Jean aurait eu la même maladie s'il n'était pas fumeur?

Exemple 4

4. Quelle est l'importance des questions souverainistes dans la victoire de François Legault?
   - Au cours de ces élections, la question souverainiste a été laissé de côté et François Legault a gagné.
   - François Legault aurait-il gagné les élections si ces questions étaient présentes?

Terminologie

1. **Réponse ou variable dépendante**, *outcome*

- C'est ce que nous voulons expliquer.
- *Exemples*:
  - Taux de chômage
  - Perspective d'emploi
  - Maladie coronarienne
  - Victoire de François Legault

Terminologie

2. **Variable indépendante, facteur de risque**
   - Tout facteur pouvant influencer la variable de réponse
   - Peut être de différents niveaux
   - Leur choix dépend de la théorie
   - *Exemples*:
     - Salaire minimum
     - Ethnie / Race
     - Fumer
     - Questions souverainistes
3. **Variables de contrôle**

Type de relation

Association

- On dit que deux variables A et B sont **associées** quand l'une se trouve plus communément en présence de l'autre.
- Se détecte souvent à partir d'un tableau dit de **contingence** ou **tableau croisé** ou d'un graphique
- Exemple - Existe-il une association entre le degré d'ouverture d'un pays et l'attitude face a la violence contre les femmes?

Pierotti, Rachel. (2013). "Increasing Rejection of Intimate Partner Violence: Evidence of Global Cultural Diffusion." *American Sociological Review*, 78: 240-265.

Nous utilisons les données des enquêtes démographiques et de santé (EDS), qui représentent un ensemble de plus de 300 enquêtes représentatives à l'échelle nationale, régionale et résidentielle menées dans des pays en développement du monde entier depuis 1992.

Association

Name	Description
beat_burnfood	Pourcentage de femmes dans chaque pays qui pensent qu'un mari a le droit de battre sa femme si elle brûle la nourriture (quantitative)
beat_burnfood_cat	Variable beat_bunfood en 4 catégories (qualitativ
no_media	Pourcentage de femmes dans chaque pays qui ont rarement accès un journal, une radio ou une télévision (quantitative)
no_media_cat	Variable no_media en 3 catégories (qualitative)
country	pays

Association

```
## [1] "/Users/visseho/OneDrive - UQAM/Cours/SOC2206_UQAM/S
```

```
## # A tibble: 6 x 11
##     ...1 beat_burnfood beat_goesout sec_school no_media c
##   <dbl>          <dbl>        <dbl>      <dbl>    <dbl> <
## 1     1            4.4         18.6       25.2      1.5 A
## 2     4            4.9         19.9       67.7      8.7 A
## 3     5            2.1         10.3       67.6      2.2 A
## 4     6            0.3          3.1       46        6.4 A
## 5     7           12.1         42.5       74.6      7.4 A
## 6     8           NA           NA         24       41.9 P
## # i 3 more variables: beat_burnfood_cat <fct>, beat_goes
## #   no_media_cat <fct>
```

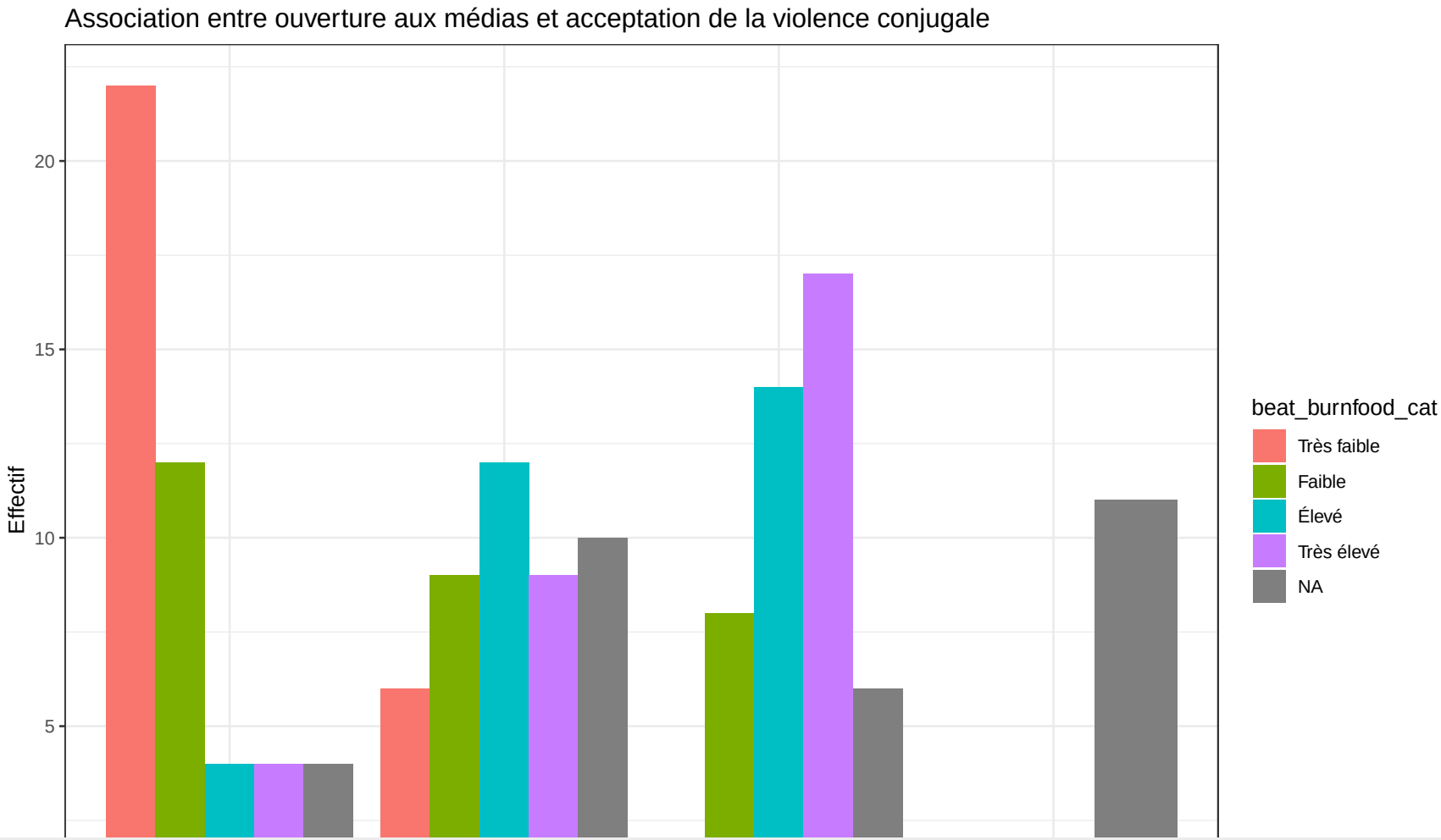
Association

- Résumé de l'information contenue dans la base de donnée

```
##       ...1        beat_burnfood    beat_goesout      sec_
##  Min.   :  1.00   Min.   : 0.10   Min.   : 0.30   Min.
##  1st Qu.: 40.50   1st Qu.: 4.50   1st Qu.:11.85   1st Qu
##  Median : 79.00   Median :11.85   Median :28.10   Median
##  Mean   : 80.53   Mean   :15.04   Mean   :28.60   Mean
##  3rd Qu.:119.50   3rd Qu.:22.25   3rd Qu.:42.08   3rd Qu
##  Max.   :160.00   Max.   :64.50   Max.   :82.70   Max.
##                   NA's   :31      NA's   :27      NA's
##     no_media         country                year         re
##  Min.   : 0.80   Length:151         Min.   :1999   Lengt
##  1st Qu.:11.25   Class :character   1st Qu.:2004   Class
##  Median :29.15   Mode  :character   Median :2007   Mode
##  Mean   :28.40                      Mean   :2007
```

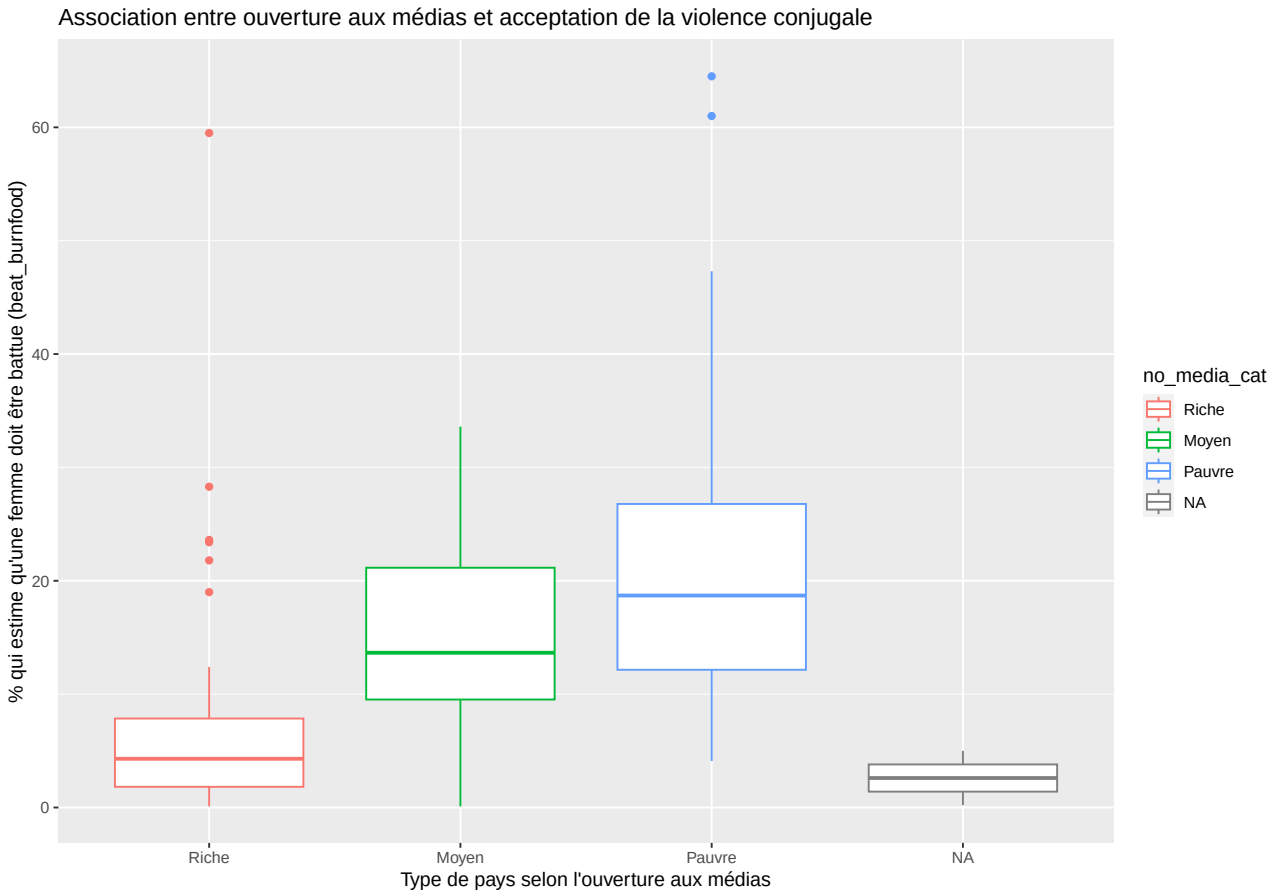
Association

- Association entre deux variables qualitatives



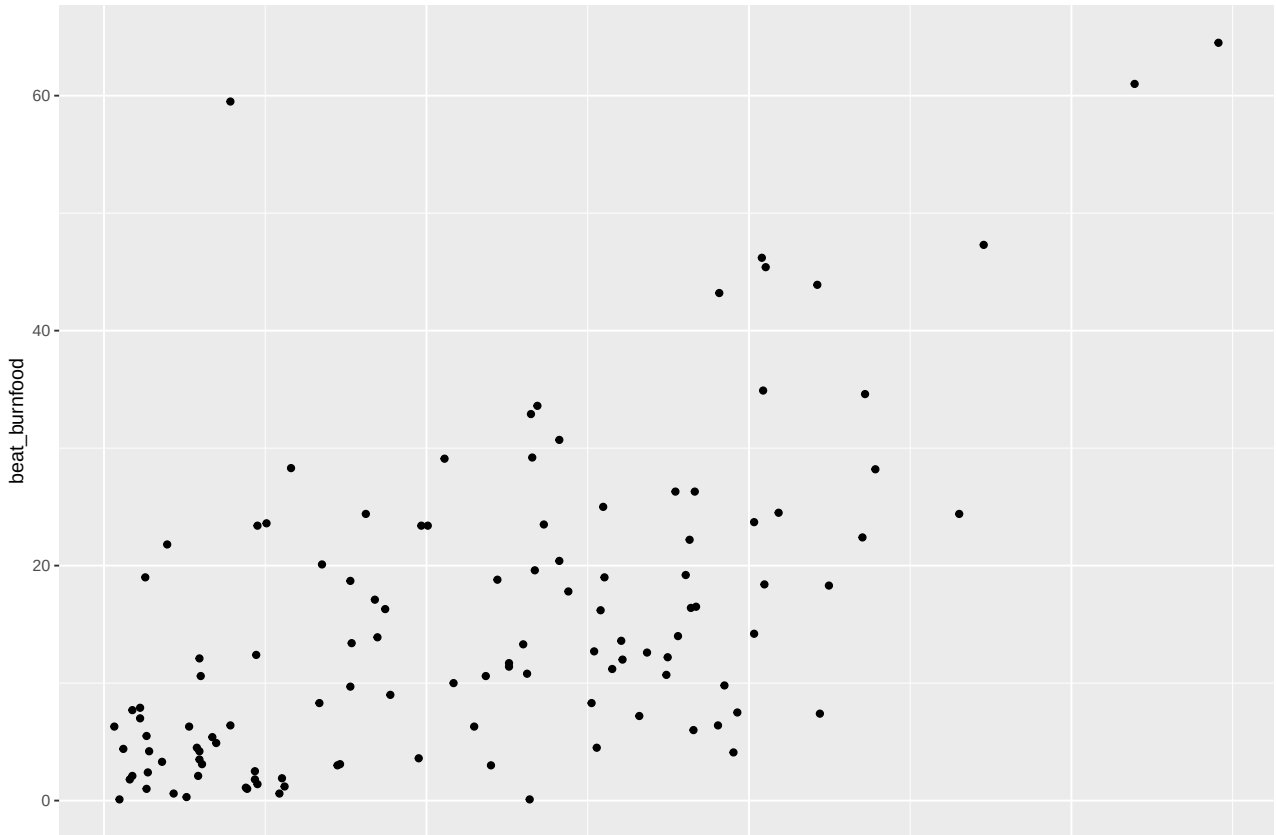
Association

- Représentation graphique (boxplot) dans le cas d'une variable qualitative et d'une variable quantitative

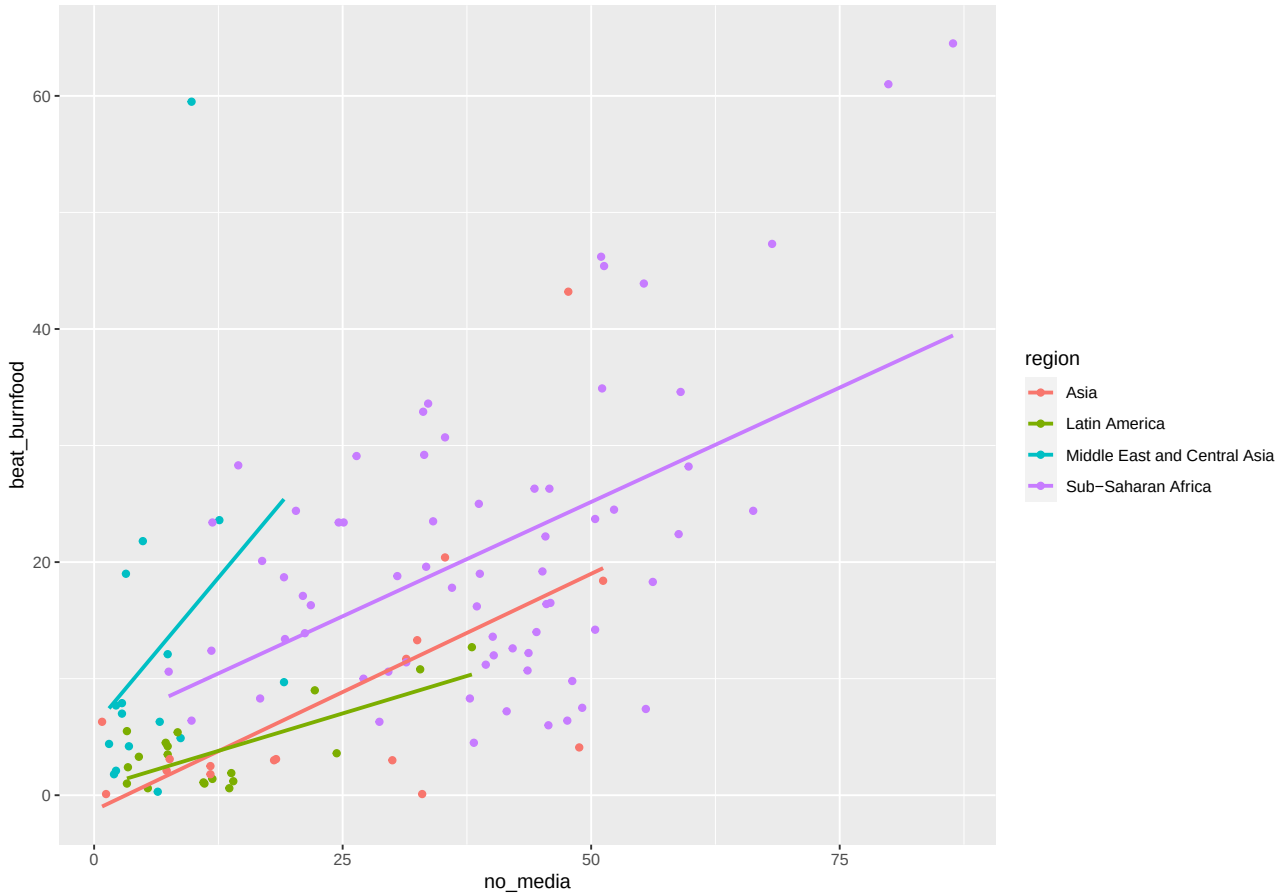


Association

- Représentation linéaire (scatterplot) et de calcul d'indicateurs (corrélation de Pearson): dans le cas de deux variables quantitatives.



Association



Relation associative

- Une association (linéaire) peut être
  - **positive** si les deux variables vont dans le même sens (une augmentation de l'un est associée à une augmentation de l'autre);
    - *Exemple* : éducation et revenu, durée de résidence et emploi
  - **négative** si les deux variables vont dans des sens opposés (une augmentation de l'un est associée à une diminution de l'autre);
    - *Exemple* : scolarisation et racisme, revenu et obésité, niveau de développement d'un pays et niveau de mortalité infantile
  - **nulle** (Absence d'association).
    - *Exemple*:

Relation causale

- L'association est une **condition nécessaire** à la causalité (Mais elle **n'est pas suffisante**).
- Toutes les associations ne sont pas causales. L'association peut arriver par hasard.
- **L'analyse statistique à elle seule ne peut constituer une preuve d'un lien de causalité**
- Comparaison entre *factuel* et *contrefactuel*
- Problème fondamental de l'inférence causale:
  - Il faut déduire des résultats contrefactuels
  - Il n'y a pas de causalité sans manipulation: caractéristiques immuables

Relation causale

- La clé pour comprendre la causalité est de penser au contrefactuel. L'inférence causale est une comparaison entre le factuel (ce qui s'est réellement passé) et le contrefactuel (ce qui se serait passé si une condition était différente).
- Contrefactuels ne sont pas observés, sauf dans les films.
- https://www.youtube.com/watch?v=BvUbv4iwbDs&rel=0&modestbranding=1&autohide=1&showinfo=0

Essais contrôlés randomisés (Expérimentation)

Essais contrôlés randomisés

- Idée clé: la **randomisation** du traitement rend les groupes de **traitement** et de **contrôle** en moyenne «identiques»
- Les deux groupes sont similaires en termes de toutes les caractéristiques (**observées et non observées**)
- Peut attribuer les différences moyennes de résultats à la différence de traitement
- Effet du Traitement Moyen (Sample Average Treatment Effect, SATE)

$$SATE = \frac{1}{n}\sum_{i=1}^{n}\{Y_i(1) - Y_i(0)\}$$

- Essais contrôlés randomisés comme **norme d'excellence** (*Gold standard*)

Essais contrôlés randomisés

- La SATE n'est pas directement observable.
- Pour le groupe de traitement qui a reçu le traitement, nous avons observé le résultat moyen sous le traitement, mais nous ne savons pas quel aurait été leur résultat moyen sans le traitement.
- Le même problème existe pour le groupe témoin car ce groupe ne reçoit pas le traitement et, par conséquent, nous n'observons pas le résultat moyen qui se produirait dans les conditions de traitement.
- Pour estimer le résultat contrefactuel moyen du traitement, nous pouvons utiliser le résultat moyen observé du groupe témoin.
- De même, nous pouvons utiliser le résultat moyen observé du groupe de traitement comme une estimation du résultat contrefactuel moyen pour le groupe de contrôle

Essais contrôlés randomisés

Dans un essai contrôlé randomisé (ECR), chaque unité est assignée de manière aléatoire au groupe de traitement ou au groupe de contrôle. La randomisation de l'assignation de traitement garantit que la différence moyenne de résultats entre les groupes de traitement et de contrôle peut être attribuée uniquement au traitement, car les deux groupes sont en moyenne identiques pour toutes les caractéristiques de prétraitement (observées et non observées).

Essais contrôlés randomisés

1. **Forces**
   - **Validité interne** - mesure dans laquelle les hypothèses de causalité sont satisfaites dans l'étude
2. **Limites**
   - **Validité externe** - mesure dans laquelle les conclusions peuvent être généralisées au-delà d'une étude particulière
   - Explication causale faible
   - Considérations éthiques
   - Possibilité de contamination

Applications

Exemple 1 discrimination raciale sur le marché du travail

1. **Question de recherche**

- La discrimination raciale existe-t-elle sur le marché du travail?
- Ou bien les disparités raciales dans le taux de chômage devraient-elles être attribuées à d'autres facteurs tels que les écarts raciaux dans le niveau d'instruction?

2. **Expérimentation**

- En réponse aux annonces dans les journaux, les chercheurs ont envoyé les CV de candidats fictifs à des employeurs potentiels.
- Changé seulement le nom du demandeur d'emploi
  - Noms afro-américains
  - Noms à consonance caucasienne
- Les autres informations sont inchangées

Exemple 1 discrimination raciale sur le marché du travail

- **Unité d'analyse**: Individus
- **Variable de traitement** (variable d'intérêt causal) **T**: Nom à consonance afro-américain
- **Groupe de traitement** (unités traitées): Afro-américains
- **Groupe de contrôle** (unités non traitées): Caucasiens
- **Réponse** (variable de réponse) **Y**: si un rappel a été effectué
  - Que signifie **"T cause Y"**?
  - Contrefactuels, **"Quoi si"** : Les Afro-Américains auraient-ils été rappelés s'ils n'avaient pas de noms afro-américains?

Exemple 1 discrimination raciale sur le marché du travail

- **Deux résultats possibles**: Y(1) et Y(0)
- **Effet causal**: `Y(1) - Y(0)`
- **Problème fondamental d'inférence causale**: un seul des deux résultats potentiels est observable

Exemple 1 discrimination raciale sur le marché du travail

- Comment pouvons-nous comprendre les contrefactuels?
  - L'association n'est pas un lien de causalité
  - Trouvez une unité similaire! ==> **Matching**
  - Est-ce-que Jamal n'a été rappelé à cause de sa race?
  - Trouver une personne blanche qui ressemble à Jamal
- Le problème: on ne peut pas correspondre sur tout
- Facteurs de **confusion non observés**: variables associées au traitement et au résultat ==> **biais de sélection**

Exemple 1 discrimination raciale sur le marché du travail

- La clé pour comprendre la causalité est de penser au contrefactuel. L'inférence causale est une comparaison entre le factuel (ce qui s'est réellement passé) et le contrefactuel (ce qui se serait passé si une condition était différente).

CV (i)	Noms à consonnance afro-américain (Ti)	Appellé pour interview?		Age	Niveau d'éducation
		Yi(1)	Yi(0)		
1	1	1	?	25	Collège
2	0	?	0	55	Secondaire
3	0	?	1	40	Collège
n	1	0	?	62	Secondaire

Exemple 1 discrimination raciale sur le marché du travail

```
##   firstname    sex  race call
## 1   Allison female white    0
## 2   Kristen female white    0
## 3   Lakisha female black    0
## 4   Latonya female black    0
## 5    Carrie female white    0
## 6       Jay   male white    0
```

Exemple 1 discrimination raciale sur le marché du travail

```
freq(resume$sex)
```

```
## Frequencies
## resume$sex
## Type: Character
##
##                  Freq   % Valid   % Valid Cum.   % Total
## ------------ ------ --------- -------------- ---------
##       female   3746     76.92          76.92     76.92
##         male   1124     23.08         100.00     23.08
##         <NA>      0                               0.00
##        Total   4870    100.00         100.00    100.00
```

Exemple 1 discrimination raciale sur le marché du travail

```
freq(resume$race)
```

```
## Frequencies
## resume$race
## Type: Character
##
##                Freq   % Valid   % Valid Cum.   % Total
## ----------- ------ --------- -------------- --------- --
##       black   2435     50.00          50.00     50.00
##       white   2435     50.00         100.00     50.00
##        <NA>      0                               0.00
##       Total   4870    100.00         100.00    100.00
```

Exemple 1 discrimination raciale sur le marché du travail

```
freq(resume$call)
```

```
## Frequencies
## resume$call
## Type: Integer
##
##                Freq   % Valid   % Valid Cum.   % Total
## ----------- ------ --------- -------------- --------- --
##           0   4478     91.95          91.95     91.95
##           1    392      8.05         100.00      8.05
##        <NA>      0                                0.00
##       Total   4870    100.00         100.00    100.00
```

Y'a-t-il discrimination ou pas?

```
#ctable(resume$race, resume$call)
tab <- table(resume$race, resume$call)

round(prop.table(tab, 1)*100, 2)
```

```
##
##                0     1
##    black 93.55  6.45
##    white 90.35  9.65
```

- SATE = 9,65 - 6,45 = 3,2%

Est-ce que les deux groupes étaient similaires au début?

```
ctable(resume$race, resume$sex)
```

```
## Cross-Tabulation, Row Proportions
## race * sex
## Data Frame: resume
##
## ------- ----- ------------- ------------- -----------
##             sex        female          male          To
##    race
##   black          1886 (77.5%)   549 (22.5%)   2435 (100
##   white          1860 (76.4%)   575 (23.6%)   2435 (100
##   Total          3746 (76.9%)  1124 (23.1%)   4870 (100
## ------- ----- ------------- ------------- -----------
```

Exercices de groupes

Causalité à partir des données observationnelles

Données observationnelles

- Souvent, nous ne pouvons pas randomiser le traitement pour des raisons éthiques et logistiques:
- par exemple, tabagisme et cancer du poumon
- Études observationnelles: traitement naturellement attribué
- Plans d'observation passifs ou plans corrélationnels
- Pas d'assignation aléatoire, pas de groupe de contrôle. . .

Données observationnelles

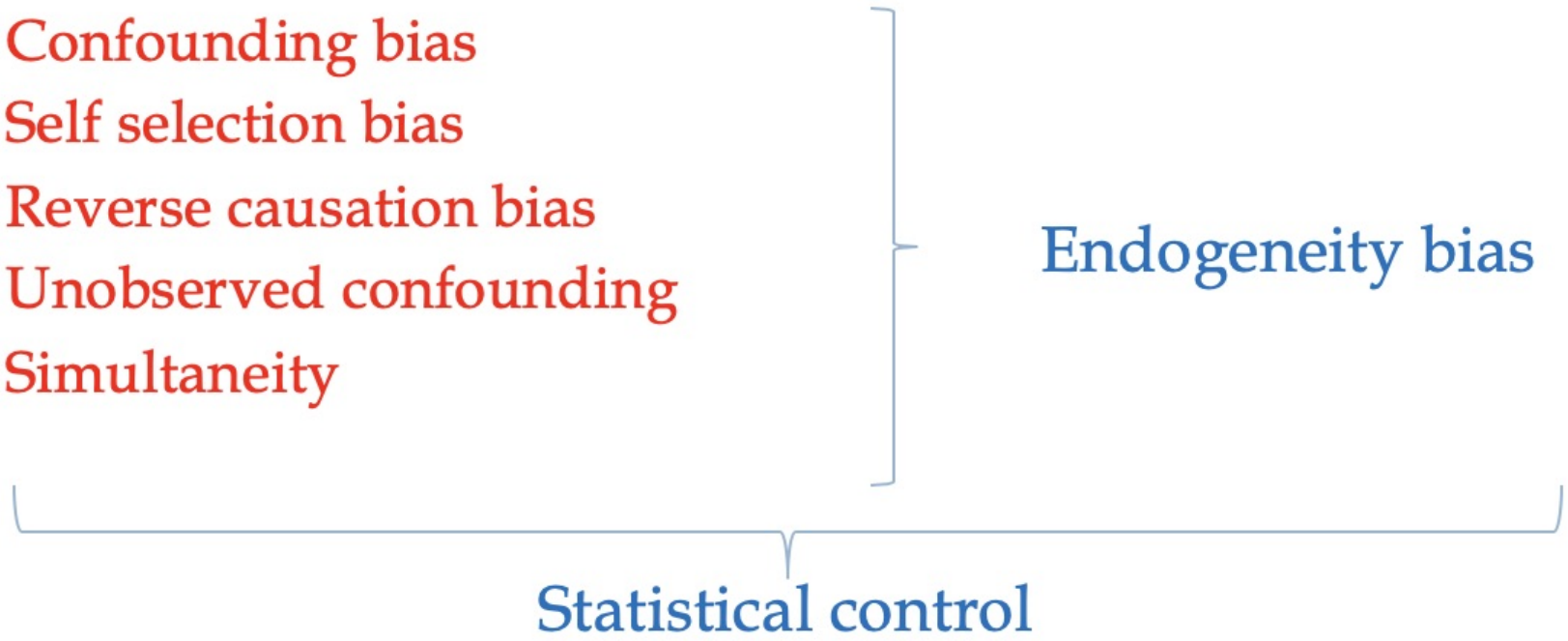
- Meilleure validité externe pour la généralisation au-delà de l'expérience
- Validité interne plus faible:
  - les variables pré-traitement peuvent différer entre les groupes (traitement et contrôle)
  1. **biais de confusion (Confounding bias)** dû à ces différences : Une variable de prétraitement associée aux variables de traitement et de résultat s'appelle un facteur de confusion et constitue une source de biais de confusion dans l'estimation de l'effet du traitement.
  2. **biais de confusion non observée (Unobserved confounding)** constitue la ménace la plus importante car il est inobservé.

Données observationnelles

3. **biais de sélection (selection bias)** de l'auto-sélection au traitement: Le biais de confusion dû à l'auto-sélection dans le groupe de traitement s'appelle un biais de sélection. Un biais de sélection apparaît souvent dans les études d'observation car les chercheurs n'ont aucun contrôle sur le destinataire du traitement.

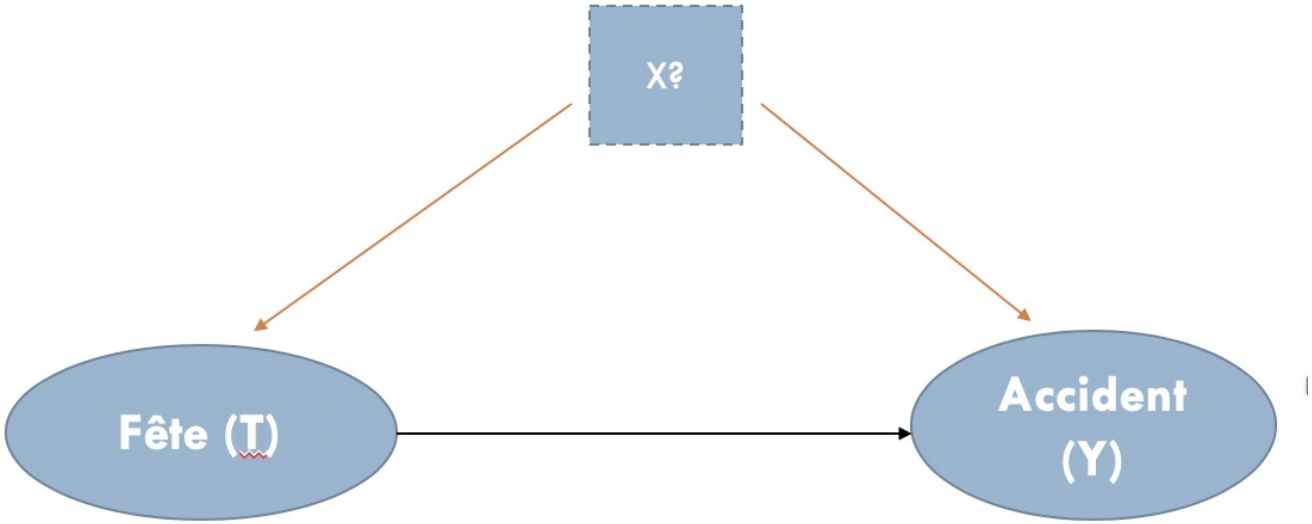
- Contrôle statistique devient alors nécessaire

Données observationnelles



Exemples

Il y a beaucoup d'accidents pendant les périodes de fête de Noël, donc la fête de Noël cause des accidents.



Problèmes?

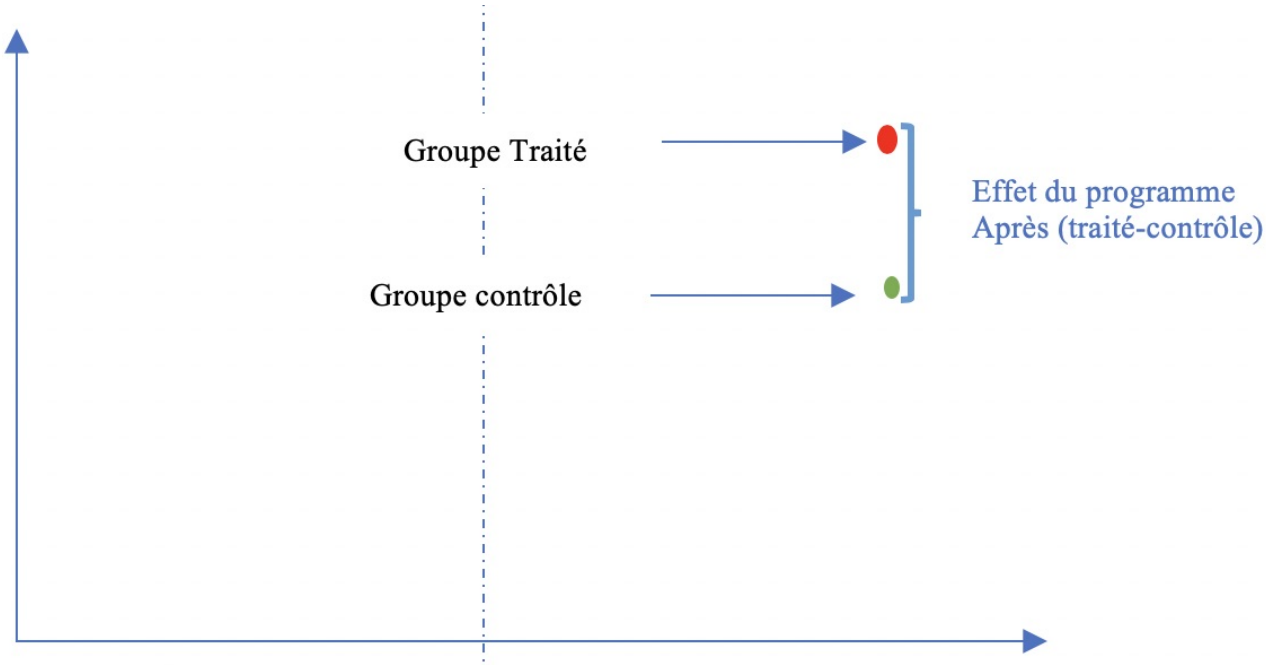
Exemples

- Adjiwanou, V. et LeGrand, T. (2013). Does antenatal care matter in the use of skilled birth attendance in rural Africa: A multi-country analysis, Social Science & Medicine 86: 26-34.
  - Est-ce que le fait d'avoir des consultations prénatales entraîne un accouchement à l'hopital?
- Adjiwanou, V. (En revision). Stepfamilies in sub-Saharan Africa and their consequences in terms of children's well-being, Presented at the Population Association of America (PAA) 2017.
  - Est-ce que le fait de vivre avec son beau-père réduit les chances de scolarisation?

Solutions

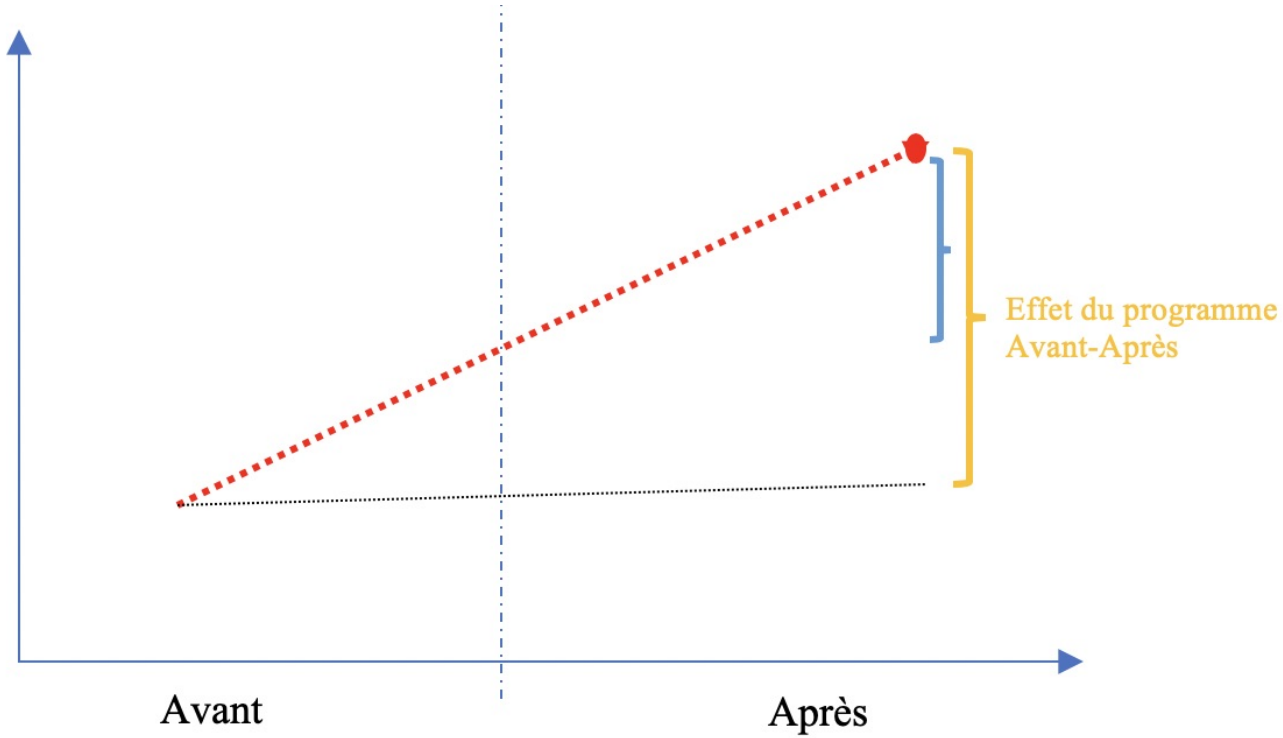
Solution 1

1. Comparaison transversale (Cross-section comparison)
   - Comparez les unités traitées avec les unités de contrôle après le traitement
   - Hypothèse: les unités traitées et les unités de contrôle sont comparables
   - Possibilité de confusion



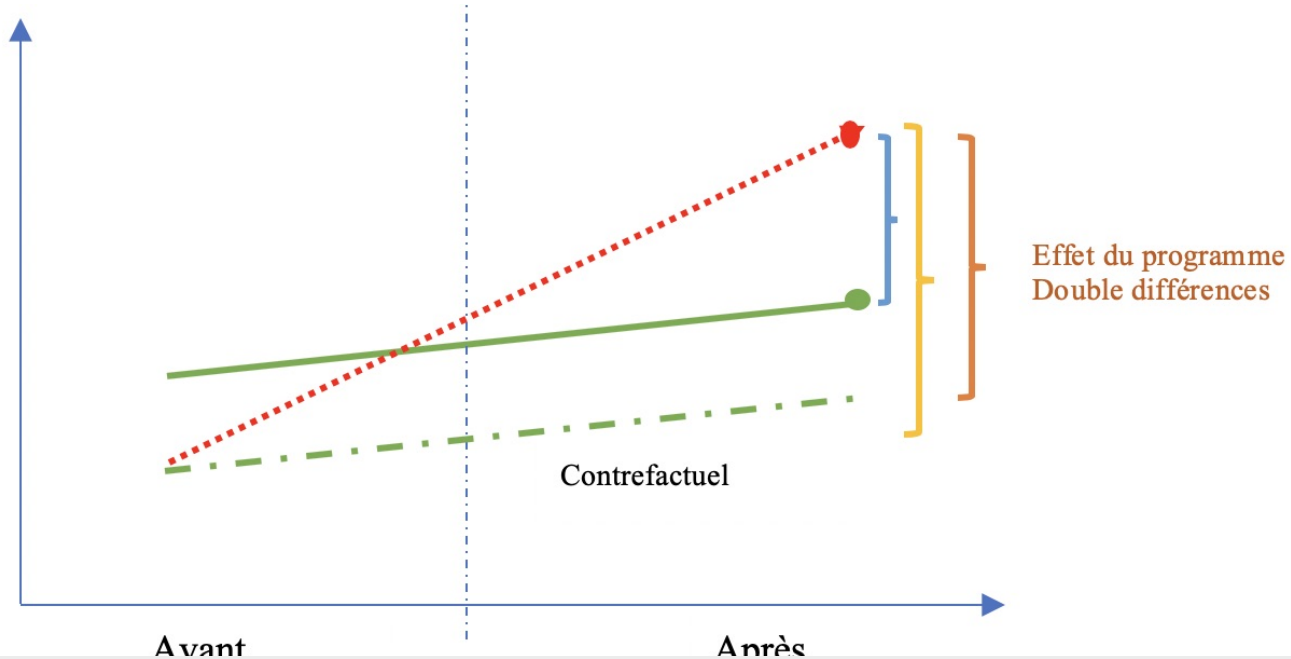
Solution 2

2. Comparaison avant-après (Before_and_after comparison)
   - Comparez les mêmes unités avant et après le traitement
   - Hypothèse: pas de variable de confusion qui change dans le temps



Solution 3

3. Double différence (DD) - Difference-in-differences
   - Compare les individus entre périodes et entre traitement et contrôle
   - Hypothèse: tendance temporelle parallèle
   - Tient compte à la fois des facteurs de confusion spécifiques aux unités et variables dans le temps.



Exercices de groupes

Comment l'augmentation du salaire minimum affecte-t-elle l'emploi?

- Débat actuel: augmentation du salaire minimum fédéral
- De nombreux économistes estiment que cet effet sera négatif:
  - surtout pour les pauvres
  - aussi pour toute l'économie
- Difficile de randomiser l'augmentation du salaire minimum
- Deux chercheurs en sciences sociales ont testé cette technique en utilisant des chaînes de restauration rapide au New-Jersey (NJ) et en Pennsylvanie (PA).
  - En 1992, le salaire minimum dans le New Jersey a augmenté de 4,25 dollars à 5,05 dollars
  - En Pennsylvanie, il est demeuré à 4,25 $
- NJ et PA (est) sont similaires
- Les chaînes de restauration rapide au NJ et en PA sont similaires: prix, salaires, produits, etc.

Données

Name	Description
chain	name of fastfood restaurant chain
location	location of restaurants (centralNJ, northNJ, PA, shoreNJ, southNJ)
wageBefore	wage before the minimum wage increase
wageAfter	wage after the minimum wage increase
fullBefore	number of fulltime employees before the minimum wage increase
fullAfter	number of fulltime employees before the minimum wage increase
partBefore	number of parttime employees before the minimum wage increase
partAfter	number of parttime employees before the minimum wage increase

Données

```
##         chain location wageBefore wageAfter fullBefore fu
## 1     wendys       PA       5.00      5.25         20
## 2     wendys       PA       5.50      4.75          6
## 3 burgerking       PA       5.00      4.75         50
## 4 burgerking       PA       5.00      5.00         10
## 5         kfc       PA       5.25      5.00          2
## 6         kfc       PA       5.00      5.00          2
##   partAfter
## 1        36
## 2         3
## 3        18
## 4         9
## 5        12
## 6         9
```

Réponse

- Variable dépendante: proportion de la main-d'oeuvre à temps plein.
- Comparaison transversale

```
## # A tibble: 2 x 2
##   state fullPropAfter
##   <chr>         <dbl>
## 1 NJ            0.320
## 2 PA            0.272
```

réponse

- Avant et après

```
##   PropBefore PropAfter diff_bef_aft_NJ
## 1  0.2965262  0.320401      0.02387474
```

réponse

- Double différences

```
## # A tibble: 1 x 3
##       NJ      PA diff_in_diff
##    <dbl>   <dbl>         <dbl>
## 1 0.0239 -0.0377        0.0616
```

Pour la suite

Pour la suite

- R possède le meilleur package de graphique appellé ggplot
- https://www.datacamp.com/courses/data-visualization-with-ggplot2-1
- Vos travaux m'intéressent!