

# Exercices Module 1

## Introduction à l'évaluation d'impact

Master GPE – UCA FERDI IHEDD

2026-03-24

### Comment utiliser ce document ?

Ce document contient des blocs de code **interactifs** que vous pouvez modifier et exécuter directement dans votre navigateur.

- Cliquez sur le bouton ► **Run Code** pour exécuter un bloc
- Modifiez le code et ré-exécutez pour voir l'effet
- Les blocs avec # TODO contiennent des exercices à compléter
- Les solutions sont cachées : cliquez sur « **Voir la solution** » pour les révéler

### Travailler dans RStudio ?

Téléchargez le script R complet du module avec tous les exercices à compléter :

Ouvrez-le dans RStudio et complétez les \_\_\_ au fur et à mesure. Exécutez bloc par bloc avec **Ctrl+Enter**.

### Pourquoi ce concept est-il crucial pour un fonctionnaire ?

Dans votre travail quotidien, vous prenez des décisions d'allocation de ressources publiques limitées (subventions agricoles, programmes de formation, infrastructures sanitaires). Sans cadre rigoureux d'évaluation, il est impossible de distinguer l'effet réel d'une politique de tendances pré-existantes ou d'autres facteurs. Maîtriser la logique de la différence-en-différences vous permettra de défendre vos choix budgétaires avec des preuves solides et d'éviter de financer des programmes inefficaces sur des décennies.

---

## Le problème fondamental de l'inférence causale

### Exercice 1.1 Identifier le contrefactuel

**Contexte** : Le gouvernement du Sénégal souhaite évaluer l'impact de son programme de microfinance *Yiriwa* lancé en 2022 dans la région de Kaolack. Ce programme cible des femmes entrepreneures dans 50 villages sélectionnés, leur accordant des prêts à taux réduit (4 % annuel) pour développer leurs activités de transformation agroalimentaire. Le Ministère de l'Économie veut savoir si ce programme

justifie son extension nationale (budget estimé : 12 milliards de FCFA). Vous disposez des données suivantes sur le revenu moyen annuel (en milliers de FCFA) :

Groupe	2021 (avant)	2023 (après)
Villages avec programme	850	1 050
Villages sans programme	820	870


**Question :** Calculez manuellement l'estimateur DiD.

```
{webr-r}
# Entrez les moyennes observées
Y_traite_avant <- ___ # Villages avec programme, avant
Y_traite_apres <- ___ # Villages avec programme, après
Y_controle_avant <- ___ # Villages sans programme, avant
Y_controle_apres <- ___ # Villages sans programme, après

# Calcul des différences
delta_traite <- Y_traite_apres - Y_traite_avant
delta_controle <- Y_controle_apres - Y_controle_avant

# Estimateur DiD
did_estime <- delta_traite - delta_controle

cat("=== Résultats ===\n")
cat("Changement chez les traités :", delta_traite, "k FCFA\n")
cat("Changement chez les contrôles:", delta_controle, "k FCFA\n")
cat("Estimateur DiD :", did_estime, "k FCFA\n")
```

 Voir la solution

```
Y_traite_avant <- 850
Y_traite_apres <- 1050
Y_controle_avant <- 820
Y_controle_apres <- 870

delta_traite <- 1050 - 850 # = 200
delta_controle <- 870 - 820 # = 50

did_estime <- 200 - 50 # = 150 k FCFA

# Interprétation : Le programme a augmenté le revenu moyen de 150 000 FCFA
# SANS le programme, les villages traités auraient eu une hausse de 50 k FCFA
# (comme les villages contrôle). La hausse SUPPLÉMENTAIRE de 150 k FCFA = effet
# du programme.
```

**Interprétation :** Sans programme *Yiriwa*, le revenu des villages traités aurait probablement augmenté de 50 000 FCFA (comme les contrôles) – c’est le contrefactuel. La différence supplémentaire de 150 000 FCFA par ménage est attribuable au programme de microfinance. Pour une décision d’extension nationale, cette hausse de 17,6 % du revenu de base constitue un argument solide en faveur du programme.

**Point clé :** L’estimateur DiD =  $(\Delta \text{ traités}) - (\Delta \text{ contrôles})$ . On soustrait la tendance naturelle pour isoler le pur effet causal du programme.

**Référence :** Card & Krueger (1994) ont popularisé cette logique pour évaluer l’impact du salaire minimum ; la même structure s’applique à tout programme public.

---

## Exercice 1.2 Biais de sélection


**Contexte :** Dans l’exercice précédent, que se serait-il passé si un conseiller pressé avait simplement comparé les villages avec et sans programme *Yiriwa* **après** l’intervention, sans tenir compte de la période initiale ? C’est une erreur fréquente dans les rapports d’évaluation au sein des ministères.

```
{webr-r}
# Remplissez les valeurs de l'exercice précédent
Y_traite_apres <- ___
Y_controle_apres <- ___

# Comparaison naïve (sans DiD)
comparaison_naive <- Y_traite_apres - Y_controle_apres

# Vrai effet DiD (de l'exercice précédent)
did_estime <- 150

cat("Comparaison naïve (après seulement) :", comparaison_naive, "k FCFA\n")
cat("Estimateur DiD (correct)           :", did_estime,           "k FCFA\n")
cat("\nBiais de sélection estimé         :", comparaison_naive - did_estime, "k
FCFA\n")
```

 Voir la solution

```
Y_traite_apres <- 1050
Y_controle_apres <- 870

comparaison_naive <- 1050 - 870 # = 180 k FCFA
did_estime <- 150 # k FCFA

# Biais de sélection = 180 - 150 = +30 k FCFA
# Les villages qui ont reçu le programme avaient déjà des revenus plus élevés.
# La comparaison naïve SURESTIME l'effet de 30 k FCFA.
```

**Interprétation :** La comparaison naïve affiche un effet de 180 000 FCFA, soit 20 % de plus que l'effet réel de 150 000 FCFA. Le biais de sélection de +30 000 FCFA vient du fait que le programme ciblait probablement des villages déjà plus dynamiques économiquement. Dans un rapport ministériel, cette surestimation de 20 % pourrait conduire à des décisions d'extension budgétaire non justifiées.

**Point clé :** Comparer uniquement « après » sans groupe de contrôle mélange l'effet du programme et les différences préexistantes entre groupes. C'est le biais de sélection.

---

## Exploration des données du cours

### Exercice 1.3 Explorer les données mpdta

Le dataset mpdta (Callaway & Sant'Anna) provient d'une étude sur l'impact du salaire minimum aux États-Unis. Bien que ce contexte soit américain, la structure des données est exactement celle que vous rencontrerez dans vos évaluations : des unités administratives (ici des comtés, mais cela pourrait être des communes sénégalaises, des districts ivoiriens, ou des provinces cambodgiennes), observées sur plusieurs années, certaines bénéficiant d'une réforme à des dates différentes. Ce dataset sera utilisé tout au long du cours.

```
{webr-r}
library(tidyverse)

# Chargement des données (disponibles dans le package 'did' sur CRAN)
# Pour cet exercice, nous simulons une version simplifiée
set.seed(42)
mpdta_sim <- tibble(
  countyreal = rep(1:50, each = 5),
  year       = rep(2003:2007, 50),
  lemp      = rnorm(250, mean = 5.5, sd = 0.3),
  first.treat = rep(sample(c(2004, 2005, 2006, 0), 50, replace = TRUE,
                          prob = c(0.25, 0.25, 0.25, 0.25)), each = 5),
  treat      = 0
) |>
mutate(treat = if_else(first.treat > 0 & year >= first.treat, 1, 0))

# TODO : Calculez le nombre d'unités par groupe de traitement
# Complétez le code ci-dessous :
mpdta_sim |>
```

```
filter(year == __) |>      # Filtrage sur la première année
count(__)                 # Comptez par première année de traitement
```

### **i** Indice 1 Si vous êtes bloqué

Pour filtrer sur la première année des données, utilisez `min()` ou regardez directement dans le tableau : quelle est la valeur minimale de `year` dans `mpdta_sim` ? Pour compter, la fonction `count()` prend comme argument le nom de la variable de regroupement.

### **i** Indice 2 Indice plus détaillé

```
# Étape 1 : filtrer sur year == 2003 (la première année)
# Étape 2 : compter par first.treat (la variable qui indique l'année du
# traitement)
mpdta_sim |>
  filter(year == 2003) |>
  count(first.treat) # ← remplacez __ par le nom de cette variable
```

### Voir la solution

```
mpdta_sim |>
  filter(year == 2003) |>
  count(first.treat)
```

**Interprétation :** Vous obtenez la répartition des unités entre les différentes vagues de traitement. Dans un contexte réel africain, ce tableau vous dirait par exemple : « 12 communes ont bénéficié de la réforme en 2004, 13 en 2005, 12 en 2006, et 13 n'ont jamais été traitées. » Cette répartition est cruciale pour évaluer la représentativité de votre groupe de contrôle.

**Point clé :** Toujours commencer par comprendre la structure de vos données — combien d'unités par groupe, par vague — avant toute estimation.

---

## Exercice 1.4 Visualiser les tendances

```
{webr-r}
library(tidyverse)

# Recréer les données simulées
set.seed(42)
mpdta_sim <- tibble(
  countyreal = rep(1:50, each = 5),
  year       = rep(2003:2007, 50),
  lemp      = rnorm(250, mean = 5.5, sd = 0.3),
  first.treat = rep(sample(c(2004, 2005, 2006, 0), 50, replace = TRUE,
                          prob = c(0.25, 0.25, 0.25, 0.25)), each = 5)
) |>
mutate(
```

```

groupe = case_when(
  first.treat == 0 ~ "Jamais traité",
  first.treat == 2004 ~ "Traité en 2004",
  first.treat == 2005 ~ "Traité en 2005",
  first.treat == 2006 ~ "Traité en 2006"
)
)

# TODO : Créez un graphique des tendances moyennes par groupe
# Calculez la moyenne de lemp par année et par groupe, puis tracez
mpdta_sim |>
  group_by(___, ___) |> # Grouper par année et groupe
  summarise(lemp_moy = mean(lemp), .groups = "drop") |>
  ggplot(aes(x = ___, y = ___, color = ___)) + # Compléter l'esthétique
  geom_line(linewidth = 1.2) +
  geom_point(size = 3) +
  labs(title = "Tendances moyennes par groupe de traitement",
        x = "Année", y = "Log emploi moyen", color = "Groupe") +
  theme_minimal()

```

### **i** Indice 1 Si vous êtes bloqué

Pour créer un graphique de tendances, il faut : (1) calculer une statistique résumée (la moyenne) pour chaque combinaison année × groupe, puis (2) passer ces moyennes à `ggplot()`. Le `group_by()` doit inclure les deux variables qui définissent chaque courbe.

### **i** Indice 2 Indice plus détaillé

```

# Les deux variables à utiliser dans group_by() sont : year et groupe
# Dans aes(), x = year, y = lemp_moy, color = groupe
mpdta_sim |>
  group_by(year, groupe) |> # ← remplacez ___ par year et groupe
  summarise(lemp_moy = mean(lemp), .groups = "drop") |>
  ggplot(aes(x = year, y = lemp_moy, color = groupe)) + ...

```

💡 Voir la solution

```
mpdta_sim |>
  group_by(year, groupe) |>
  summarise(lemp_moy = mean(lemp), .groups = "drop") |>
  ggplot(aes(x = year, y = lemp_moy, color = groupe)) +
  geom_line(linewidth = 1.2) +
  geom_point(size = 3) +
  labs(title = "Tendances moyennes par groupe de traitement",
       x = "Année", y = "Log emploi moyen", color = "Groupe") +
  theme_minimal()
```

**Interprétation** : Observez si les courbes des différents groupes sont à peu près parallèles avant leurs dates de traitement respectives. C'est ce que l'on appelle le « test visuel » des tendances parallèles – un premier indicateur de la validité de l'hypothèse centrale de la DiD.

**Point clé** : Un graphique de tendances est TOUJOURS la première analyse à produire. Il vous donne une intuition visuelle avant même de lancer les régressions.

---

## Question de réflexion

### Exercice 1.5 Discussion

Répondez aux questions suivantes (pas de code R – réflexion théorique) :

```
{webr-r}
# Espace pour tester vos idées en code si nécessaire
# Question 1 : Dans quel cas l'hypothèse de tendances parallèles
# serait-elle clairement violée dans l'exemple du programme de microfinance ?
#
# Exemple de réponse à compléter :
# "L'hypothèse serait violée si les villages sélectionnés pour le programme
# étaient des villages qui ____ avant le programme."

# Question 2 : Si on ne peut pas observer le contrefactuel, quelle condition
# sur les données permet à la DiD de l'approximer correctement ?
#
# "La DiD approxime le contrefactuel si les deux groupes avaient des
# tendances ____ avant le traitement."

cat("Remplacez les ____ dans les commentaires ci-dessus.\n")
cat("Puis discutez avec votre groupe.\n")
```

## 💡 Éléments de réponse

**Question 1 :** L'hypothèse serait violée si les villages sélectionnés pour le programme *Yiriwa* étaient des villages qui **progressaient déjà plus vite** avant le programme (biais de sélection positive). Par exemple, si les agents de terrain ont ciblé les villages les plus dynamiques ou les plus accessibles, ils auraient eu une croissance plus forte même sans programme. Dans les évaluations de politiques publiques africaines, ce problème est très fréquent car les programmes ciblés visent souvent les zones les « plus prêtes ».

**Question 2 :** La DiD approxime le contrefactuel si les deux groupes avaient des tendances **parallèles** (identiques) avant le traitement. C'est l'hypothèse de tendances parallèles (« parallel trends assumption »). Elle signifie : en l'absence de programme, les revenus des villages traités auraient évolué comme ceux des villages contrôles.

**Point clé :** Cette hypothèse est une *hypothèse*, pas un fait démontrable pour la période post-traitement. On peut la rendre crédible en montrant qu'elle tenait dans le passé (test des pré-tendances – modules 3 et 4).

---

## i Fin du Module 1

### Récapitulatif des compétences acquises :

- Calculer un estimateur DiD manuellement à partir de 4 moyennes
- Identifier et quantifier le biais de sélection
- Explorer des données de panel en R avec tidyverse
- Visualiser les tendances par groupe

**Prochain exercice :** Module 2 – Implémenter la DiD en régression avec `lm()`