CASSION · FORMATION

Méthodes d'évaluation d'impact

Le raisonnement contrefactuel et les protocoles qui le rendent possible — randomisation, différences de différences, appariement et régression sur discontinuité — et comment énoncer les limites d'une comparaison avant-après.

Formateur	Pierre Robentz Cassion
Niveau	Avancé
Heures apprenant	24 h
Enseignée en	Python · R
Mise à jour	29 juillet 2026
Secteurs	Éducation · Nutrition · Santé publique
Normes et méthodologies	Critères d'évaluation du CAD de l'OCDE · Définitions d'indicateurs de l'UNICEF · Enquête SMART · Théorie du changement

Consulter et exécuter la formation en ligne sur data-analysis.cassion.dev/fr/cours/impact-evaluation-methods

Ce que vous saurez faire

— Énoncer le contrefactuel qu'exige une affirmation, et dire pourquoi une évolution avant-après n'en est pas un

— Lire un tableau d'équilibre et juger si un groupe de comparaison en est un

— Ajuster une estimation en différences de différences et argumenter l'hypothèse sur laquelle elle repose

— Apparier sur un score de propension et rendre compte des unités que l'appariement a écartées

— Reconnaître un seuil d'éligibilité comme un protocole, et vérifier les trois choses qu'il exige

— Calculer un effet minimal détectable avant l'évaluation, et dire ce qu'une étude sous-dimensionnée a le droit de conclure

Prérequis

— Régression appliquée aux données de programme

Sommaire

À propos de cette formation

Chaque bloc « rapportez-le en entier » de *Régression appliquée aux données de programme* se termine par la même phrase : *observationnel, les groupes n'ont pas été randomisés.* Voici le cours qui dit ce qu'il faudrait pour la supprimer.

Tout tient dans une question posée trois fois au même fichier.

Les scores de littératie ont monté de 7,0 points sur l'année scolaire. En appariement, sur 585 enfants, intervalle +6,1 à +7,9. Un document de programme appellerait cela l'effet de la cantine scolaire.

Les écoles avec cantine terminent l'année 2,1 points devant les écoles sans. Intervalle −1,2 à +5,4. Un autre document appellerait *cela* l'effet.

Les deux groupes ont progressé, et les écoles sans cantine ont progressé davantage. Avec cantine +6,7, sans +7,6, si bien que l'estimation en différences de différences vaut **−0,96 point**, avec un intervalle groupé par école de −3,5 à +1,6. Il n'y a ici aucun effet à rapporter, et les deux premiers nombres sont l'année scolaire et un écart préexistant déguisés en effet.

Trois autres résultats structurent le cours.

Quinze écoles ont le programme et neuf non, et elles n'étaient pas comparables au départ. La différence standardisée de littératie initiale vaut 0,85, contre le 0,1 que tolérerait un essai randomisé. Les six caractéristiques sont toutes déséquilibrées, et le Centre comme le Nord-Ouest comptent cinq écoles avec cantine pour une sans.

L'appariement améliore l'équilibre en jetant 40 % du groupe traité. Neuf écoles témoins peuvent absorber neuf écoles traitées ; les six écartées sont les six aux scores initiaux les plus élevés, et l'estimation qui survit ne vaut que pour les écoles restantes.

Ce protocole n'aurait jamais pu détecter moins de 4,9 points. Avec une corrélation intra-classe de 0,065 et cinquante enfants par école, 24 grappes achètent un effet minimal détectable de 4,9 points là où les mêmes 1 200 enfants randomisés individuellement en auraient acheté 2,3. Ce nombre était fixé avant qu'un seul enfant soit mesuré.

Python et R tout du long. Il vous faut *Régression appliquée aux données de programme* d'abord — ce cours ne consacre aucun temps à la façon d'ajuster un modèle et tout son temps à la comparaison que le modèle a le droit d'être.

Première partie

Le contrefactuel

CHAPITRE 1

Un gain de sept points qui était l'année scolaire

Leçon 1 sur 8 · 180 min

La littératie a monté de 7,0 points entre le passage initial et le final, et une demande de financement l'appelle l'effet de la cantine scolaire. Les écoles sans cantine ont monté de 7,6 points. Tout le gain, et un peu plus, appartient à l'année et non aux repas.

1.1 Une question, trois réponses

```python
import pandas as pd
import numpy as np

assessment = pd.read_csv("learning-assessment-2024.v1.csv")
assessment["pct"] = assessment["raw_score"] / assessment["items"]
literacy = (assessment[assessment["domain"] == "literacy"]
            .pivot_table(index="student_id", columns="assessment_round",
                         values="pct").dropna())

roster = pd.read_csv("school-roster-2024.v1.csv").drop_duplicates(
    "student_id", keep="last")
d = literacy.join(roster.set_index("student_id"), how="inner")

print(d.groupby("feeding_programme")[["baseline", "endline"]].mean().round(4))
print(f"overall gain: {(d['endline'] - d['baseline']).mean():+.4f}")
```

```r
library(dplyr)

d |> summarise(baseline = mean(baseline), endline = mean(endline),
               .by = feeding_programme)
```

	Initial	Final	Gain	n
Avec cantine	57,75 %	64,41 %	**+6,66 pts**	398
Sans cantine	54,67 %	62,29 %	**+7,62 pts**	187
Tous les élèves	56,76 %	63,73 %	**+6,97 pts**	585

Trois affirmations peuvent se bâtir sur ce tableau et deux sont fausses.

« Le programme a produit un gain de 7,0 points. » Avant-après, apparié, IC à 95 % +6,1 à +7,9, t = 15,2. Tranché, reproductible, et cela attribue aux repas une année de scolarité dont chaque enfant a bénéficié.

« Les écoles avec cantine ont 2,1 points d'avance. » Coupe transversale finale, IC à 95 % −1,2 à +5,4. Cela compare deux groupes qui étaient distants de 3,1 points avant même que le programme soit mesuré.

« Le gain a été inférieur de 0,96 point dans les écoles avec cantine. » Chaque groupe contre son propre point de départ, IC à 95 % −3,5 à +1,6 une fois les 24 écoles prises en compte. **C'est celle qui tente de répondre à la question**, et sa réponse est qu'il n'y a rien ici.

1.2 Le contrefactuel, énoncé en une phrase

Toute affirmation d'impact a une seconde moitié cachée.

> La littératie dans les écoles avec cantine a monté de 6,66 points **par rapport à ce qui se serait passé sans le programme.**

Cette seconde proposition est le contrefactuel, et il n'est jamais observé. Les mêmes enfants, la même année, sans repas, n'est pas une chose que contient un jeu de données ; une évaluation est donc toujours une argumentation sur ce qu'il faut y substituer.

Affirmation	Contrefactuel supposé	Est-ce plausible ?
Avant-après	Les mêmes enfants n'auraient pas changé du tout	**Non** — une année scolaire a eu lieu
Coupe transversale finale	Les deux groupes auraient été identiques	**Non** — ils différaient de 3,1 points a
Différences de différences	Les deux groupes auraient changé d'autant	Discutable, et la leçon 3 est cette dis

Nommer le contrefactuel est toute la méthode. Une fois écrit, l'affirmation avant-après s'effondre sans la moindre statistique : elle suppose que les enfants n'apprennent rien dans une année passée à l'école.

1.3 Pourquoi le nombre avant-après est si convaincant

Il est précis, il est grand, et il est facile à calculer — trois propriétés qui n'ont rien à voir avec le fait d'être juste.

```python
gain = d["endline"] - d["baseline"]
se = gain.std(ddof=1) / np.sqrt(len(gain))
print(f"{gain.mean():+.4f}  95% CI [{gain.mean() - 1.96*se:+.4f},"
      f" {gain.mean() + 1.96*se:+.4f}]  t = {gain.mean()/se:.1f}")
```

```r
t.test(d$endline - d$baseline)
```

t = 15,2. Toutes les barrières du cours de statistiques passent. L'intervalle est étroit, la taille d'effet est grande, l'échantillon suffit, et la valeur p est très loin sous n'importe quel seuil.

Aucune de ces vérifications ne peut détecter un groupe de comparaison manquant, et c'est ce qu'il faut emporter de cette leçon. La rigueur statistique appliquée au mauvais contrefactuel produit une réponse fausse et assurée, et c'est l'assurance qui la fait entrer dans une demande de financement.

1.4 Trois choses que contient une évolution avant-après

Chaque fois que vous en voyez une, c'est une somme, et seul le dernier terme est le programme.

Le passage du temps. Les enfants grandissent et sont enseignés. Les prix montent. Les points d'eau vieillissent. Ici, c'est la totalité des 7 points.

La régression vers la moyenne. Ceux qui ont été sélectionnés parce qu'ils étaient les plus mal lotis paraîtront mieux au passage suivant même si rien n'a été fait, parce qu'une part de « le plus mal loti » était du bruit de mesure. Le PB à l'admission du cours de nutrition en est l'exemple type.

Tout le reste de ce qui est arrivé. Une sécheresse, un mouvement de change, le programme d'une autre agence dans le même district, un changement dans la façon dont l'indicateur est enregistré.

```python
by_school = d.groupby(["school_id", "feeding_programme"])[
    ["baseline", "endline"]].mean()
by_school["gain"] = by_school["endline"] - by_school["baseline"]
print(by_school.groupby("feeding_programme")["gain"].describe()[
    ["mean", "min", "max"]].round(4))
```

```r
by_school |> summarise(mean = mean(gain), min = min(gain), max = max(gain),
                       .by = feeding_programme)
```

Chacune des 24 écoles a progressé, de +1,2 point à +12,2. Un programme qui n'aurait rien fait du tout aurait produit un tableau de nombres positifs.

1.5 Quand une comparaison avant-après est légitime

Elle n'est pas toujours fausse, et il vaut la peine d'être précis sur les cas où elle tient.

Quand le contrefactuel est réellement plat et que vous pouvez le montrer. Trois passages ou plus avant le programme sans tendance constituent un argument ; un seul point de départ non.

Quand le résultat ne peut pas changer tout seul. Une latrine ne se construit pas elle-même. Un point d'eau ne devient pas chloré sans que quelqu'un le fasse. La couverture d'une chose que seul votre programme fournit a un contrefactuel nul défendable.

Quand vous faites du suivi, non de l'évaluation. « La présence est passée de 84 % à 88 % » est une affirmation vraie sur le monde et utile à suivre. Elle devient une affirmation sur le programme quand quelqu'un écrit « grâce à ».

La formule à surveiller est « grâce à ». C'est là qu'un nombre de suivi devient une affirmation d'impact, et elle est généralement ajoutée par quelqu'un qui n'était pas dans l'analyse.

1.6 Rapportez-le en entier

```
Literacy outcomes, baseline to endline 2024

  All students        56.8% to 63.7%    +6.97 points   95% CI +6.1 to +7.9
    With feeding      57.8% to 64.4%    +6.66 points
    Without feeding   54.7% to 62.3%    +7.62 points

  The overall change is reported as a monitoring result and is not
  attributed to the feeding programme. Schools without the programme gained
  slightly more, so the difference-in-differences estimate of the programme
  effect is -0.96 points (95% CI -3.5 to +1.6, clustered on 24 schools).

  The before-after change assumes children would have learned nothing over
  a school year. That counterfactual is not defensible here and the
  comparison group is what replaces it.
```

Le dernier paragraphe fait trois lignes et c'est la différence entre un rapport de suivi et une affirmation d'impact. Écrire le contrefactuel est ce qui force le choix, et un rapport qui n'en écrit jamais a fait le choix par défaut.

1.7 La suite

L'estimation en différences de différences ci-dessus repose sur le fait que les deux groupes d'écoles sont comparables sur ce qui compte. La leçon suivante teste cela directement, sur les quinze et les neuf, et trouve une différence standardisée huit fois plus grande que ce que tolérerait un essai randomisé.

CHAPITRE 2

Une différence standardisée de 0,85, quand la limite est 0,1

Leçon 2 sur 8 · 180 min

Quinze écoles ont la cantine et neuf non. Elles diffèrent de 0,85 écart-type sur la littératie initiale et de 0,68 sur la taille, et le Centre compte cinq écoles avec cantine pour une sans. Voilà à quoi ressemble un groupe de comparaison quand personne n'a randomisé.

2.1 Construisez le tableau d'équilibre avant tout le reste

```python
import pandas as pd
import numpy as np

enrolment = pd.read_csv("school-enrolment-2024.v1.csv")
current = enrolment[(enrolment["school_year"] == 2024)
                    & (enrolment["age_years"] <= 20)]   # drop the age outliers
roster = pd.read_csv("school-roster-2024.v1.csv").drop_duplicates(
    "student_id", keep="last")
joined = current.merge(roster[["student_id", "school_id", "feeding_programme"]],
                       on="student_id", suffixes=("", "_r"))
joined["over_age"] = joined["age_years"] > joined["grade"] + 5
joined["displaced"] = joined["displacement_status"] != "resident"

by = joined.groupby("school_id")
schools = pd.DataFrame({
    "feeding_programme": by["feeding_programme"].first(),
    "baseline": by["baseline"].mean(),      # from the assessment file
    "age_years": by["age_years"].mean(),
    "size": by.size(),
    "over_age": by["over_age"].mean(),
    "disability": by["disability_reported"].mean(),
    "displaced": by["displaced"].mean(),
    "district": by["admin2"].first(),
}).reset_index()

def std_diff(treated, control):
    pooled = np.sqrt((treated.var(ddof=1) + control.var(ddof=1)) / 2)
    return (treated.mean() - control.mean()) / pooled

for column in ["baseline", "age_years", "size", "over_age",
               "disability", "displaced"]:
    t = schools[schools["feeding_programme"]][column].dropna()
    c = schools[~schools["feeding_programme"]][column].dropna()
    print(f"{column:12} {t.mean():8.4f} {c.mean():8.4f}"
          f"  std.diff {std_diff(t, c):+.3f}")
```

```r
library(dplyr)

schools |>
  summarise(across(c(baseline, age_years, size, over_age), mean),
            .by = feeding_programme)
```

Caractéristique	Avec cantine (15)	Sans (9)	Différence standardisée
Littératie initiale	54,10 %	51,05 %	**+0,85**
Âge moyen	9,65	9,42	**+0,79**
Taille de l'école	50,3	44,8	**+0,68**
Handicap signalé	7,6 %	9,7 %	−0,52
Déplacé ou retourné	23,7 %	26,9 %	−0,51
Retard d'âge	39,2 %	36,8 %	+0,43

Le seuil conventionnel vaut 0,10, et chacune des six le dépasse — quatre d'entre elles d'un facteur cinq ou plus.

```python
print(pd.crosstab(schools["district"], schools["feeding_programme"]))
```

```r
table(schools$district, schools$feeding_programme)
```

District	Sans cantine	Avec cantine
Artibonite	3	3
Centre	**1**	**5**
Nord-Ouest	**1**	**5**
Sud	4	2

Deux districts sont à cinq contre un en faveur du programme et un est à deux contre quatre. Tout ce qui diffère par ailleurs entre le Centre et le Sud voyage avec chaque comparaison d'écoles avec et sans cantine.

2.2 Ce qu'est la différence standardisée, et pourquoi pas une valeur p

```python
t = schools[schools["feeding_programme"]]["baseline"]
c = schools[~schools["feeding_programme"]]["baseline"]
from scipy import stats
print(f"std.diff {std_diff(t, c):+.3f}   p = {stats.ttest_ind(t, c).pvalue:.3f}")
```

```r
# Both, and only one of them is the right tool.
```

Différence standardisée +0,85, p = 0,064. Un relecteur ne lisant que la valeur p conclurait que les écoles sont équilibrées sur la littératie initiale. Elles ne le sont pas ; l'échantillon compte 24 écoles et le test n'a presque aucune puissance.

La valeur p d'un test d'équilibre confond le déséquilibre et la taille d'échantillon, ce qui est l'inverse de ce dont une vérification d'équilibre a besoin. Vingt-quatre unités passeront tous les tests d'équilibre jamais écrits et deux mille en échoueront à certains par hasard.

Rapportez la différence standardisée, qui ne dépend pas de n. Au-delà de 0,10 c'est un déséquilibre qu'il faut traiter ; au-delà de 0,25 c'est un déséquilibre qu'un ajustement par régression ne sauvera pas.

2.3 Ce que la randomisation aurait acheté

Non pas l'équilibre sur tout — l'équilibre *en espérance*, sur tout, y compris les variables que vous n'avez pas mesurées.

Cette dernière proposition est toute la valeur. L'ajustement peut corriger un déséquilibre de littératie initiale parce qu'elle est dans le fichier. Il ne peut pas corriger un déséquilibre de motivation du directeur, du fait que l'école a été choisie parce que quelqu'un a plaidé pour elle, ou de sa distance au bureau du district — dont rien n'est enregistré nulle part.

```
rng = np.random.default_rng(20260729)
draws = []
for _ in range(2000):
    assigned = rng.permutation(schools["feeding_programme"].values)
    t = schools["baseline"][assigned]
    c = schools["baseline"][~assigned]
    draws.append(std_diff(t, c))
print(f"under random assignment, |std.diff| > 0.85 in "
      f"{np.mean(np.abs(draws) > 0.85):.1%} of draws")
```

```
# Permute the assignment 2,000 times and see how unusual the real split is.
```

Réaffecter au hasard les mêmes 24 écoles produit un déséquilibre de cette ampleur dans 5,3 % de 2 000 tirages. Le partage observé est donc inhabituel sans être impossible — ce qui est exactement la mauvaise conclusion à en tirer. **La question n'est pas de savoir si ce déséquilibre aurait pu survenir par hasard ; c'est qu'il n'est pas survenu par hasard, parce que personne n'a randomisé.**

2.4 Quand la randomisation est refusée, et que dire

Elle l'est généralement, et les raisons sont le plus souvent bonnes.

Objection	La réponse honnête
« On ne peut pas priver des écoles dans le besoin »	Randomisez l'**ordre** d'un déploiement par vagues ; tout le mon
« C'est le ministère qui choisit les écoles »	Randomisez **à l'intérieur** de la liste éligible du ministère, ou é
« C'est déjà en cours »	Dites-le, employez un protocole de comparaison, et publiez le t
« Le bailleur veut des résultats cette année »	Un essai randomisé sous-dimensionné et des différences de diffé

Le déploiement par vagues est le protocole le plus sous-employé de ce secteur. Tout programme qui s'étend sur trois ans a un ordre randomisable, et le randomiser ne coûte rien et n'écarte aucune des objections ci-dessus.

Là où rien de tout cela n'est possible, le tableau d'équilibre est ce que vous devez au lecteur. Ce n'est pas une formalité : c'est la preuve sur laquelle il décide quelle part de l'ajustement il doit croire.

2.5 Les trois questions à poser avant un protocole d'évaluation

Qui a décidé qui reçoit le programme, et sur quoi ? Si la réponse est une règle, vous avez peut-être un protocole à seuil. Si c'est le jugement d'une personne, vous avez une confusion que vous ne pouvez pas énumérer.

Y a-t-il quelque chose de mesuré avant le début du programme ? Sans point de départ, les différences de différences de la leçon suivante sont indisponibles et il ne vous reste qu'une coupe transversale.

Combien d'unités ont été affectées ? Non pas combien de personnes ont été mesurées — combien d'écoles, de centres, de villages, de camps. Ce nombre fixe la précision, et la leçon 6 montre ce qu'achètent 24.

2.6 Rapportez-le en entier

```
Comparability of feeding and non-feeding schools

  15 schools with the programme, 9 without. Assignment was not randomised.

  Standardised differences at baseline:
    Baseline literacy     +0.85      School size        +0.68
    Mean age              +0.79      Disability         -0.52
    Displacement          -0.51      Over-age           +0.43

  All six characteristics exceed the 0.10 conventional threshold and all six
  exceed 0.25. Programme schools are concentrated in Centre and
  Nord-Ouest (5 of 6 schools in each) and scarce in Sud (2 of 6).

  Balance tests are not reported as p-values: with 24 units they would show
  balance regardless. The standardised difference does not depend on n.

  The comparison is adjusted for baseline literacy and district. Adjustment
  cannot address unmeasured differences in why these schools were selected,
  and no result in this report should be read as an effect of the programme
  alone.
```

Le dernier paragraphe est là où une évaluation gagne ou perd un relecteur, et sa longueur est la bonne : deux phrases, énonçant la limite sans s'en excuser.

2.7 La suite

Si les deux groupes différaient au départ, la façon d'employer ce point de départ est de le soustraire. La leçon suivante le fait formellement, obtient une estimation de $-0{,}96$ point, et consacre l'essentiel de sa longueur à l'hypothèse qui la rend signifiante — hypothèse que deux passages de mesure ne peuvent pas tester.

Deuxième partie

Quand personne n'a randomisé

CHAPITRE 3

Moins 0,96 point, et l'hypothèse qu'on ne peut pas tester

Leçon 3 sur 8 · 180 min

Soustraire à chaque groupe son propre point de départ donne des différences de différences de −0,96 point. Cela repose sur le fait que les deux groupes étaient sur le point de changer d'autant, hypothèse intestable avec deux passages — la leçon porte donc sur la façon de l'argumenter, non de la vérifier.

3.1 Deux différences, soustraites

```python
import pandas as pd
import numpy as np

cells = d.groupby("feeding_programme")[["baseline", "endline"]].mean()
print(cells.round(4))

did = ((cells.loc[True, "endline"] - cells.loc[True, "baseline"])
       - (cells.loc[False, "endline"] - cells.loc[False, "baseline"]))
print(f"difference-in-differences: {did:+.4f}")
```

```r
library(dplyr)

d |> summarise(baseline = mean(baseline), endline = mean(endline),
                .by = feeding_programme) |>
  mutate(gain = endline - baseline)
```

	Initial	Final	Évolution
Avec cantine	57,75 %	64,41 %	+6,66
Sans cantine	54,67 %	62,29 %	+7,62
Différence	+3,09	+2,12	**−0,96**

L'estimation se lit dans la dernière colonne ou dans la dernière ligne et c'est le même nombre. Cette symétrie est ce que décrit le nom : la différence entre les évolutions des deux groupes, égale à l'évolution de la différence entre les deux groupes.

L'écart initial de 3,09 points est ce que le protocole retire. Une comparaison transversale en fin d'année aurait rapporté +2,12 et l'aurait appelé un effet ; les différences de différences disent que l'essentiel de cet écart précède le programme.

3.2 Ajustez-le comme une régression, parce qu'il vous faut l'intervalle

```python
import statsmodels.formula.api as smf

d = d.assign(gain=d["endline"] - d["baseline"])

naive = smf.ols("gain ~ feeding_programme", data=d).fit()
```

```
clustered = naive.get_robustcov_results(cov_type="cluster",
                                        groups=d["school_id"])

by_school = d.groupby(["school_id", "feeding_programme"])["gain"].mean().reset_index()
aggregated = smf.ols("gain ~ feeding_programme", data=by_school).fit()
```

```
lm(gain ~ feeding_programme, data = d)                          # naive
lm(gain ~ feeding_programme, data = by_school)                  # school level
```

Approche	Estimation	ET	IC à 95 %
Niveau élève, naïf	−0,96 pt	0,98	−2,89 à +0,96
Niveau élève, groupé par école	**−0,96 pt**	**1,29**	**−3,48 à +1,56**
Niveau école, 15 contre 9	−0,68 pt	1,18	−2,99 à +1,63

L'intervalle groupé est celui à rapporter, pour la raison qu'a établie le cours de régression : le programme a été attribué à 24 écoles, non à 585 enfants.

Chaque version traverse zéro largement. Il n'y a aucun effet à rapporter, et l'intervalle dit que l'étude n'a pas pu écarter quoi que ce soit entre une perte de 3,5 points et un gain de 1,6.

3.3 L'hypothèse, énoncée exactement

Tendances parallèles : en l'absence du programme, les résultats des deux groupes auraient évolué d'autant.

Trois choses que cette hypothèse n'exige *pas*, et chacune est une mélecture courante :

Elle n'exige pas que les groupes se ressemblent. Ils partent avec 3,09 points d'écart et c'est sans importance — le protocole le soustrait. L'équilibre importe pour la plausibilité des tendances parallèles, non pour le fonctionnement de l'estimateur.

Elle n'exige pas que les tendances soient plates. Les deux groupes peuvent progresser ; l'hypothèse est qu'ils auraient progressé *également*.

Elle n'exige ni même variance, ni même effectif, ni même composition. Cela agit sur l'intervalle, non sur l'identification.

Ce qu'elle exige est inobservable, parce que c'est une affirmation sur un monde où le programme n'a pas eu lieu. C'est pourquoi cette leçon porte sur l'argumentation plutôt que sur le test.

3.4 Pourquoi deux passages ne peuvent pas la tester

```
rounds = assessment["assessment_round"].unique()
print(rounds)
```

```
unique(assessment$assessment_round)
```

Il y en a deux : initial et final. Avec deux points, vous ne pouvez tracer qu'une droite par groupe, si bien que toute divergence antérieure au programme est inmesurable par construction.

Avec trois passages ou plus avant le programme, vous pouvez regarder. Tracez le résultat de chaque groupe sur la période antérieure et voyez si les lignes bougent ensemble.

Ce n'est pas une preuve — un parallélisme passé ne garantit pas un parallélisme futur — mais c'est une pièce à conviction, et c'est la plus convaincante qu'une évaluation de ce type puisse porter.

L'implication pour le protocole est une décision de collecte prise des années plus tôt. Si vous comptez évaluer par différences de différences, collectez plus d'un passage antérieur. Le coût est une enquête de plus ; l'alternative est une hypothèse que personne ne peut examiner.

3.5 Quatre arguments à faire valoir quand on ne peut pas la tester

Chacun est disponible ici, et ensemble ils sont ce que le rapport offre à la place d'un test.

Nommez pourquoi les écoles du programme ont été choisies. Si la sélection a porté sur quelque chose d'invariant dans le temps — leur emplacement, qui les dirige —, les tendances parallèles sont plus plausibles que si elle a porté sur quelque chose de tendanciel, comme une baisse récente d'effectifs.

Montrez que d'autres résultats ont bougé en parallèle. La numératie est mesurée sur les mêmes enfants et n'est pas ce qu'une cantine vise en premier.

```
def panel(domain):
    wide = (assessment[assessment["domain"] == domain]
            .pivot_table(index="student_id", columns="assessment_round",
                         values="pct").dropna())
    out = wide.join(roster.set_index("student_id"), how="inner").reset_index()
    return out.assign(gain=out["endline"] - out["baseline"])

for domain in ("literacy", "numeracy"):
    frame = panel(domain)
    fit = smf.ols("gain ~ feeding_programme", data=frame).fit(
        cov_type="cluster", cov_kwds={"groups": frame["school_id"]})
    print(f"{domain}: {fit.params['feeding_programme[T.True]']:+.4f}")


# A placebo outcome: it should show nothing, and here it does.
```

La numératie donne +0,39 point, IC à 95 % −1,16 à +1,94 — le même protocole appliqué à un résultat que le programme n'était pas censé déplacer en premier, et il ne montre rien non plus. C'est une preuve faible et c'est le genre de preuve disponible.

Testez une période placebo s'il en existe une. Deux passages antérieurs au programme devraient donner des différences de différences nulles. Il n'y en a aucun ici, et le dire vaut mieux que de laisser croire le contraire.

Montrez que le résultat n'est pas porté par une seule unité. Retirez chaque école à tour de rôle et réajustez ; si l'estimation oscille, le protocole repose sur la tendance d'une école.

```
for school in sorted(d["school_id"].unique()):
    subset = d[d["school_id"] != school]
    est = smf.ols("gain ~ feeding_programme", data=subset).fit()
    print(f"without {school}: {est.params['feeding_programme[T.True]']:+.4f}")


# 24 refits, one line each. Cheap, and a reviewer will ask.
```

Retirer n'importe quelle école déplace l'estimation entre −1,61 et −0,37 point. Aucune école ne la porte à elle seule, ce qui vaut une ligne en annexe et constitue la vérification

qu'un relecteur effectue en premier sur 24 unités.

3.6 Là où les différences de différences échouent

Un calendrier différent. Si le programme a démarré à des mois différents selon les écoles, une seule césure avant-après classe mal une partie de la période traitée. Le cas échelonné exige un autre estimateur, et le modèle naïf à doubles effets fixes est aujourd'hui connu pour y être biaisé.

Un changement de composition. Les 585 enfants sont ceux qui ont les deux passages. Si des enfants sont partis de façon inégale entre les groupes, le panel n'est pas deux fois la même population — c'est le problème d'attrition que l'exercice du cours de statistiques a trouvé dans ce fichier même.

L'anticipation. Si les écoles ont changé de comportement en apprenant que le programme arrivait, le point de départ est déjà contaminé et l'estimation est trop petite.

Un groupe témoin qui est affecté. Des enfants qui changent d'école, des enseignants qui sont mutés, ou une école voisine qui modifie ses pratiques en réaction : chacun casse l'hypothèse que le groupe de comparaison montre ce qui se serait passé.

3.7 Rapportez-le en entier

```
School feeding and literacy, difference-in-differences

  585 students with both assessment rounds, in 24 schools.

                       Baseline   Endline   Change
    Feeding (15 sch)     57.75%    64.41%    +6.66
    No programme (9)     54.67%    62.29%    +7.62
    Difference           +3.09     +2.12     -0.96

  Estimate -0.96 points, 95% CI -3.48 to +1.56, clustered on 24 schools.
  No effect on literacy is detected.

  The estimate assumes the two groups would have changed by the same amount
  without the programme. Only two assessment rounds exist, so this cannot be
  examined and is argued rather than tested: assignment appears to follow
  district and school size rather than a trend in results, and the same
  design applied to numeracy also shows nothing.

  The interval excludes effects larger than 1.6 points in either direction
  but not smaller ones. The design's minimum detectable effect was 4.9
  points, so this is a null result about large effects only.
```

Le dernier paragraphe convertit un résultat nul en une affirmation bornée, ce que le cours de statistiques demandait et ce que la leçon sur la puissance rend précis.

3.8 La suite

Les différences de différences emploient le point de départ en le soustrayant. La leçon suivante l'emploie autrement — pour choisir quelles écoles de comparaison garder — et constate qu'améliorer l'équilibre revient à écarter six des quinze écoles du programme.

CHAPITRE 4

Un meilleur équilibre, six écoles en moins

Leçon 4 sur 8 · 180 min

L'appariement ramène le déséquilibre initial de 3,35 points à 1,41 et donne une estimation de −0,71 au lieu de −0,96. Il y parvient en écartant six des quinze écoles du programme — les six aux scores initiaux les plus élevés — si bien que la réponse porte désormais sur d'autres écoles que la question.

4.1 L'appariement, en un paragraphe

Trouvez, pour chaque unité traitée, une unité non traitée qui lui ressemble ; ne comparez que les paires. Tout le reste de la méthode est un choix sur « qui lui ressemble » et sur ce qu'il faut faire quand aucune telle unité n'existe.

L'attrait est de rendre la comparaison explicite. Le coût est que la seconde moitié de cette phrase — *que les paires* — change la population dont parle la réponse, et elle le fait discrètement.

4.2 Appariez les écoles

```python
import pandas as pd
import numpy as np

treated = schools[schools["feeding_programme"]].sort_values("baseline")
control = schools[~schools["feeding_programme"]].sort_values("baseline")

used, pairs = set(), []
for _, c in control.iterrows():
    available = treated[~treated["school_id"].isin(used)]
    nearest = (available["baseline"] - c["baseline"]).abs().idxmin()
    used.add(treated.loc[nearest, "school_id"])
    pairs.append({"control": c["school_id"],
                  "treated": treated.loc[nearest, "school_id"],
                  "base_c": c["baseline"], "base_t": treated.loc[nearest, "baseline"],
                  "gain_c": c["gain"], "gain_t": treated.loc[nearest, "gain"]})

matched = pd.DataFrame(pairs)
print(f"{len(matched)} pairs from {len(treated)} treated and {len(control)} control")
```

```r
library(MatchIt)

m <- matchit(feeding_programme ~ baseline, data = schools,
             method = "nearest", ratio = 1)
summary(m)
```

Neuf paires à partir de quinze écoles traitées et neuf témoins. L'appariement un pour un sans remise ne peut produire au plus qu'autant de paires que le plus petit groupe a d'unités, si bien que **six écoles traitées n'ont pas de partenaire et quittent l'analyse.**

	Avant appariement	Après
Écoles traitées	15	**9**
Écoles témoins	9	9
Écart initial	+3,35 pts	**+1,41 pt**
Estimation	−0,96 pt	**−0,71 pt**
Erreur-type	1,29	1,22

L'équilibre s'est amélioré et l'estimation a à peine bougé, ce qui est le résultat qu'on espère : cela dit que les différences de différences n'étaient pas portées par l'écart initial.

4.3 Quelles six écoles sont parties, et pourquoi cela compte

```
dropped = treated[~treated["school_id"].isin(used)]
print(dropped[["school_id", "baseline", "gain"]].round(4))
print(f"dropped mean baseline: {dropped['baseline'].mean():.4f}")
print(f"kept mean baseline:    {matched['base_t'].mean():.4f}")
```

```
# Which units matching threw away is the first table to print, not the last.
```

Les six écoles écartées sont les six aux plus hauts scores initiaux de littératie — de 53,9 % à 65,5 %, contre une moyenne de 55,8 % pour les écoles traitées appariées. Elles ont été écartées parce que le groupe témoin ne contient aucune école ayant obtenu un score aussi élevé : il n'y a rien à quoi les comparer.

Ce n'est pas un défaut de l'appariement ; c'est l'appariement qui vous dit quelque chose de vrai. Les neuf écoles témoins ne peuvent pas parler pour les écoles du programme les plus performantes, parce qu'aucune école témoin de ce niveau n'existe.

Mais cela change l'affirmation. L'estimation appariée répond à « parmi les écoles dont la littératie initiale est inférieure à environ 60 %, qu'a fait le programme ? » — et c'est une question plus étroite que celle que le rapport se proposait de traiter.

4.4 Ce que « lui ressemble » devrait signifier

Apparier sur une seule variable est une simplification pédagogique. En pratique le choix se fait entre trois approches.

Approche	Apparier sur	À
Exacte	Des valeurs identiques de quelques variables catégorielles	P
Plus proche voisin sur un score de propension	La probabilité prédite d'être traité	B
Exacte grossière	Des valeurs regroupées en classes	V

```
import statsmodels.formula.api as smf

ps = smf.logit("feeding_programme ~ baseline + size + district",
               data=schools).fit(disp=0)
schools["pscore"] = ps.predict(schools)
print(schools.groupby("feeding_programme")["pscore"].describe()[
    ["min", "mean", "max"]].round(3))
```

```
glm(feeding_programme ~ baseline + size + district, data = schools,
    family = binomial())
```

Un score de propension est un nombre unique résumant de nombreuses covariables, et apparier dessus équivaut à apparier sur toutes à la fois — ce qui le rend utile quand il y a plus de covariables qu'un petit échantillon ne peut en apparier exactement.

Ce n'est pas un moyen d'obtenir plus de données. Le score réduit la dimensionnalité ; il ne crée pas de comparateurs là où il n'y en a pas.

4.5 Vérifiez le recouvrement avant d'apparier, pas après

C'est la vérification que l'exercice du cours de régression a découverte à ses dépens sur les deux programmes PCIMA, et elle s'applique ici sans changement.

```
t = schools[schools["feeding_programme"]]["pscore"]
c = schools[~schools["feeding_programme"]]["pscore"]
print(f"treated range {t.min():.3f}–{t.max():.3f}")
print(f"control range {c.min():.3f}–{c.max():.3f}")
print(f"treated units above the control maximum: {(t > c.max()).sum()}")
```

```
# Plot the two propensity score distributions. The tails are the whole story.
```

Les unités hors de la plage de l'autre groupe n'ont pas de contrefactuel dans les données. L'appariement les écarte, ce qui est honnête ; l'ajustement par régression les garde et extrapole, ce qui ne l'est pas. **C'est la vraie différence entre les deux méthodes**, et c'est une meilleure raison de préférer l'appariement que n'importe quelle affirmation sur le biais.

4.6 Quatre choses à rapporter, toujours

Combien d'unités ont été écartées, et lesquelles. Le premier tableau, non une note de bas de page.

L'équilibre avant et après. Les différences standardisées sur chaque covariable, les deux colonnes, pour qu'un lecteur voie ce que l'appariement a acheté.

Quelle population l'estimation décrit désormais. Une phrase, dans les mots qu'emploie un responsable de programme — « les écoles dont la littératie initiale est sous 60 % », non « la région de support commun ».

L'estimation non appariée également. Si l'appariement a sensiblement changé la réponse, c'est en soi un constat sur le degré auquel la comparaison dépendait des unités qu'il a écartées.

4.7 Ce que l'appariement ne peut pas faire

Il ne peut pas corriger un facteur de confusion non mesuré. Deux écoles aux scores initiaux identiques peuvent encore différer par la raison pour laquelle l'une a été sélectionnée. L'appariement équilibre ce sur quoi vous avez apparié et rien d'autre — ce qui est exactement la limite de l'ajustement par régression, atteinte par un autre chemin.

Il ne peut pas créer un groupe témoin. Si les unités non traitées sont systématiquement différentes, l'appariement rapportera un excellent équilibre sur une poignée de paires et gardera le silence sur tous les autres.

Il ne supprime pas le besoin d'un protocole. Appariement plus point de départ font des différences de différences sur un sous-ensemble ; l'appariement seul est une coupe transversale mieux élevée. **L'estimation ci-dessus est le premier cas**, et combiner les deux est généralement plus solide que l'un ou l'autre.

4.8 Rapportez-le en entier

```
School feeding and literacy, matched difference-in-differences

  Nearest-neighbour matching on baseline literacy, 1:1 without replacement.
  9 matched pairs from 15 treated and 9 control schools.

  6 of 15 programme schools were discarded: SCH05, SCH12, SCH14, SCH15,
  SCH19, SCH21. Their mean baseline literacy is 60.7% against 55.8% for the
  matched treated schools; no control school scored high enough to match
  them.

  Baseline imbalance   +3.35 points before matching, +1.41 after.
  Estimate             -0.71 points, SE 1.22, on 9 pairs.
  Unmatched estimate   -0.96 points, 95% CI -3.48 to +1.56.

  The matched estimate describes schools with baseline literacy below about
  60%. It does not describe the six highest-performing programme schools,
  which have no comparator in this data.

  Matching balances the covariates it was given. It does not address why
  these schools were selected for the programme.
```

La liste nommée des écoles écartées est ce qui rend ceci rapportable. Un lecteur peut vérifier si les six parties sont les six qui intéressent le plus le programme, et aucune statistique de résumé ne le lui dira.

4.9 La suite

L'appariement et les différences de différences bâtissent tous deux un groupe de comparaison à partir d'unités qui n'ont pas été affectées par une règle. La leçon suivante prend le cas inverse — un programme où une règle a tout décidé, nettement, à un nombre — et constate que deux des trois choses qu'exige un tel protocole sont déjà dans les données.

Troisième partie

Des règles, et des limites

CHAPITRE 5

Une discontinuité parfaite avec rien derrière

Leçon 5 sur 8 · 180 min

Tous les enfants à 124 mm ont été orientés et 2 % de ceux à 125 mm l'ont été. L'affectation est aussi nette qu'une règle peut l'être et la densité ne montre aucune manipulation, mais le protocole ne peut pas tourner, faute d'un résultat mesuré chez les enfants non orientés.

5.1 Une règle est un protocole

Le protocole de dépistage du PB oriente un enfant vers l'alimentation supplémentaire en dessous de 125 mm et ne fait rien à ce seuil ou au-dessus. Ce seul nombre accomplit ce qu'une évaluation doit d'ordinaire payer : **il affecte le traitement sur une variable, à un point connu, sans latitude.**

```python
import pandas as pd
import numpy as np

screening = pd.read_csv("muac-screening-artibonite-2024.v1.csv")
screening["referred"] = screening["outcome"] != "no-action"

window = screening[screening["muac_mm"].between(118, 132)]
print(window.groupby("muac_mm")["referred"].agg(["mean", "size"]).round(3))
```

```r
library(dplyr)

screening |>
  filter(between(muac_mm, 118, 132)) |>
  summarise(referred = mean(outcome != "no-action"), n = n(), .by = muac_mm)
```

PB (mm)	Orientés	n
122	100,0 %	44
123	100,0 %	47
124	**100,0 %**	**45**
125	**2,0 %**	**51**
126	0,0 %	78
127	0,0 %	70

Un millimètre fait passer la probabilité d'orientation de 1,00 à 0,02. Les enfants à 124 et à 125 mm sont, à tout ce qui compte près, les mêmes enfants — l'erreur de mesure d'un ruban PB est plus grande que l'écart entre eux — et un groupe a été traité tandis que l'autre ne l'a pas été.

C'est ce qu'une donnée de routine produit de plus proche d'une expérience randomisée, et cela arrive gratuitement, parce que le programme allait de toute façon appliquer la règle.

5.2 Ce qu'exige une régression sur discontinuité

Trois conditions. Deux sont vérifiables dans ce fichier et la troisième ne l'est pas.

Un : l'affectation doit être réellement nette au seuil. Vérifié ci-dessus. Si la règle était appliquée avec souplesse — disons 60 % d'orientations juste en dessous et 20 % juste au-dessus —, le protocole fonctionne encore mais devient *flou*, et l'estimation doit être divisée par le saut de probabilité de traitement plutôt que lue directement.

Deux : personne ne doit manipuler la variable de position. Si un dépisteur souhaitant qu'un enfant soit traité enregistre 124 au lieu de 126, les enfants juste en dessous du seuil ne sont plus comparables à ceux juste au-dessus — ce sont ceux que quelqu'un a décidé d'aider.

```python
counts = screening["muac_mm"].value_counts().sort_index()
print(counts.loc[120:130])

below = screening["muac_mm"].between(120, 124).sum()
above = screening["muac_mm"].between(125, 129).sum()
print(f"120-124: {below}   125-129: {above}")

screening["last_digit"] = screening["muac_mm"] % 10
print((screening["last_digit"].value_counts(normalize=True).sort_index() * 100).round(1))
```

```r
table(screening$muac_mm)[as.character(118:132)]
table(screening$muac_mm %% 10)
```

Les effectifs montent régulièrement à travers le seuil : 45 à 124, 51 à 125, 78 à 126. Il n'y a pas d'accumulation juste sous 125, et la densité du côté traité est *plus faible* que du côté non traité, ce qui est l'allure d'une distribution de PB réellement croissante.

Les chiffres terminaux sont plats, de 9,6 % à 10,8 % sur les dix. Pas d'arrondi aux cinq, pas d'évitement du seuil. Le test de préférence de chiffres du cours sur la qualité des données, employé ici à une autre fin : **une vérification de manipulation est un test de préférence de chiffres posé à un nombre précis.**

Trois : un résultat doit être mesuré des deux côtés. Il ne l'est pas, et cela met fin à l'analyse.

5.3 La condition à laquelle ce registre échoue

```python
print(screening.columns.tolist())
print(f"children screened once: {(screening['child_id'].value_counts() == 1).sum()}")
print(f"children screened twice: {(screening['child_id'].value_counts() == 2).sum()}")
```

```r
names(screening)
table(table(screening$child_id))
```

Le fichier enregistre une mesure, une orientation, et rien d'autre. 4 194 des 4 206 enfants n'apparaissent qu'une fois. Il n'y a pas de PB de suivi, pas de statut de guérison, pas de second passage de dépistage — donc pour les enfants à 125 mm qui n'ont pas été orientés, aucun résultat n'existe.

Le registre de traitement ne comble pas le vide. `cmam-admissions-2024` contient des résultats pour les seuls enfants traités, et ses identifiants ne se rattachent pas au fichier de

dépistage. Même s'ils le faisaient, il ne porterait que le côté orienté, c'est-à-dire précisément la moitié qu'un protocole de discontinuité possède déjà.

Une estimation d'impact a besoin du côté non traité. Sans lui, la règle d'affectation la plus nette du secteur ne produit rien.

5.4 Ce qu'il aurait fallu

C'est la partie utile de la leçon, parce que c'est une spécification de collecte sur laquelle quelqu'un peut agir.

Une mesure de suivi chez les enfants juste au-dessus du seuil. Pas tous — les enfants entre 125 et 130 mm, remesurés à huit semaines. Quelques centaines d'enfants, une visite de plus, et le protocole devient estimable.

Le même résultat des deux côtés. La guérison est définie différemment pour les enfants traités et non traités, et une comparaison a besoin d'une seule définition appliquée aux deux : le PB à huit semaines, mesuré de la même façon quoi qu'il se soit passé entre-temps.

Assez d'enfants près du seuil. L'estimation emploie une fenêtre, et tout ce qui est en dehors est écarté.

```
for bandwidth in (2, 3, 5, 10):
    band = screening[
        screening["muac_mm"].between(125 - bandwidth, 124 + bandwidth)]
    print(f"±{bandwidth} mm: {len(band)} children"
          f" ({(band['muac_mm'] < 125).sum()} below, "
          f"{(band['muac_mm'] >= 125).sum()} above)")
```

```
# The bandwidth trade-off, in four lines.
```

Fenêtre	Enfants	Sous le seuil	Au-dessus
±2 mm	221	92	129
±3 mm	335	136	199
±5 mm	556	187	369
±10 mm	1 210	265	945

Une fenêtre étroite est plus crédible et moins précise. Les enfants à moins de 2 mm du seuil sont presque identiques entre eux ; il y en a 221. Élargissez à 10 mm et vous avez six fois l'échantillon et comparez des enfants qui diffèrent réellement.

Choisissez la fenêtre avant de voir le résultat, et rapportez l'estimation à deux ou trois autres. Un résultat qui n'apparaît qu'à une seule fenêtre est un résultat sur la fenêtre.

5.5 Où se cachent les seuils dans les données de programme

Dès qu'on cherche des règles d'éligibilité, ce secteur en regorge.

Seuil	Programme	Variable de position
PB < 125 mm	Alimentation supplémentaire	Périmètre brachial
PB < 115 mm	Alimentation thérapeutique	Périmètre brachial
Score de pauvreté sous un seuil	Transfert monétaire	Test de moyens indirect
Phase 3 de l'IPC	Réponse d'urgence	Classification de zone
Sous une cible de couverture	Appui aux formations sanitaires	Couverture rapportée

Chacun est une expérience naturelle que personne n'a conçue, et chacun exige les mêmes trois vérifications. La troisième — un résultat des deux côtés — est celle qui échoue le plus souvent, parce que les programmes mesurent ce qu'ils traitent.

C'est une décision de collecte, non d'analyse. Décider de mesurer un échantillon d'unités inéligibles coûte de l'argent une année où personne ne le demande, et achète trois ans plus tard une évaluation qui serait autrement impossible.

5.6 Rapportez-le en entier

```
Supplementary feeding eligibility: assessment of a regression discontinuity

  Assignment: referral to TSFP below MUAC 125 mm.

  Sharpness    100% referred at 124 mm (n=45), 2.0% at 125 mm (n=51).
               The rule is applied without exception.

  Manipulation No pile-up below the cut-off: counts are 45, 51, 78 at 124,
               125 and 126 mm. Terminal digits are uniform (9.6-10.8%).
               No evidence the running variable was adjusted.

  Outcome      Not available. The register records screening and referral
               only; 4,194 of 4,206 children appear once, and no follow-up
               measurement exists on either side of the cut-off.

  Conclusion   The design is available and cannot be estimated. To make it
               estimable, re-measure MUAC at 8 weeks on a sample of children
               screened between 125 and 130 mm, using the same definition
               applied to referred children. 556 children sit within
               +/-5 mm of the cut-off, of whom 369 were not referred.
```

Une évaluation de protocole qui conclut « non estimable » est un livrable, et celle-ci est plus utile qu'une estimation forcée : elle nomme la mesure manquante, sa taille et son coût.

5.7 La suite

Le protocole à seuil a échoué sur des données qui n'existent pas. La leçon suivante porte sur un protocole qui a tourné sur des données qui existent, et pose la question qui aurait dû venir en premier — quelle ampleur d'effet aurait-il jamais pu trouver.

CHAPITRE 6

La réponse était fixée avant la mesure du premier enfant

Leçon 6 sur 8 · 180 min

Vingt-quatre écoles de cinquante enfants chacune peuvent détecter 4,9 points. Les mêmes 1 200 enfants randomisés individuellement pourraient en détecter 2,3. Le protocole a décidé de ce que l'évaluation était capable de trouver, des années avant que quiconque lise un résultat.

6.1 Calculez ce que le protocole peut trouver

```
import numpy as np

SD = 0.1429        # within-school SD of attendance, from the register
ICC = 0.065        # intra-cluster correlation, from the empty mixed model
M = 50             # children per school
Z_ALPHA, Z_BETA = 1.96, 0.84       # 5% two-sided, 80% power

deff = 1 + (M - 1) * ICC
print(f"design effect {deff:.2f}")

def mde(k_treated, k_control):
    n1, n0 = k_treated * M, k_control * M
    return (Z_ALPHA + Z_BETA) * SD * np.sqrt(deff * (1 / n1 + 1 / n0))

print(f"15 vs 9 schools: {mde(15, 9) * 100:.2f} points")
```

```
# clusterPower::cpa.normal(), or the formula above. Either way, before the study.
```

Effet de plan 4,19. Effet minimal détectable 4,88 points.

Cette évaluation n'aurait jamais pu trouver un effet inférieur à environ cinq points de présence, quoi qu'ait fait l'analyse par la suite. C'est une propriété de 24 écoles, cinquante enfants chacune, et une corrélation intra-classe de 0,065 — les trois connues avant qu'un enfant soit mesuré.

6.2 D'où vient le nombre

Quatre entrées, et deux seulement sont des décisions.

Entrée	Ici	Décision ou fait ?
Écart-type du résultat	0,143	Fait, tiré d'un pilote ou d'un passage antérieur
Corrélation intra-classe	0,065	Fait, et celui que personne n'a d'ordinaire
Grappes par bras	15 et 9	**Décision**
Enfants par grappe	50	**Décision, et le levier le plus faible**

```
for k in (12, 24, 40, 60, 100):
    print(f"{k:3} schools: MDE {mde(k // 2, k // 2) * 100:.2f} points")
```

```
# The whole planning conversation in five rows.
```

Écoles	Enfants	EMD
12	600	6,68 pts
24	**1 200**	**4,73 pts**
40	2 000	3,66 pts
60	3 000	**2,99 pts**
100	5 000	2,32 pts

Diviser par deux l'effet détectable coûte environ quatre fois plus d'écoles. La précision s'améliore comme la racine carrée, si bien qu'un protocole visant 3 points au lieu de 4,7 a besoin de 60 écoles et non de 24.

6.3 Des grappes, pas des enfants

C'est le nombre qui surprend les responsables de programme et c'est toute la raison pour laquelle les protocoles en grappes coûtent cher.

```
individual = (Z_ALPHA + Z_BETA) * SD * np.sqrt(4 / 1200)
print(f"1,200 children, individually randomised: {individual * 100:.2f} points")
print(f"1,200 children in 24 schools:            {mde(12, 12) * 100:.2f} points")
print(f"ratio: {np.sqrt(deff):.2f}")
```

```
# Same children, same outcome, same budget. Twice the detectable effect.
```

2,31 points individuellement, 4,73 points en grappes — un facteur 2,05, qui est la racine carrée de l'effet de plan.

Ajouter des enfants aux écoles existantes n'aide presque pas. Avec une CCI de 0,065, passer de 50 à 100 enfants par école porte l'effet de plan de 4,19 à 7,44 et l'EMD de 4,73 à 4,45 points — 6 % d'amélioration pour le double du travail de terrain.

Ajouter des écoles aide proportionnellement. C'est la règle de planification : **quand le programme est attribué par grappe, achetez des grappes.**

6.4 La corrélation intra-classe est l'entrée que personne n'a

C'est la plus difficile des quatre à obtenir et celle qui déplace le plus la réponse.

```
for icc in (0.01, 0.03, 0.065, 0.10, 0.20):
    d = 1 + (M - 1) * icc
    m = (Z_ALPHA + Z_BETA) * SD * np.sqrt(d * (1 / 600 + 1 / 600))
    print(f"ICC {icc:.3f}: design effect {d:5.2f}, MDE {m * 100:.2f} points")
```

```
# Vary the ICC across the plausible range and report the MDE for each.
```

CCI	Effet de plan	EMD, 24 écoles
0,01	1,49	2,82 pts
0,03	2,47	3,63 pts
0,065	**4,19**	**4,73 pts**
0,10	5,90	5,61 pts
0,20	10,80	7,59 pts

L'EMD va de 2,8 à 7,6 points sur la plage plausible des CCI. Un calcul de puissance citant une seule CCI sans intervalle n'est pas un calcul, c'est une supposition assortie d'arithmétique.

Où en trouver une. Un passage antérieur de la même enquête, une étude publiée dans le même secteur et le même pays, ou un pilote. À défaut des trois, publiez le tableau ci-dessus et planifiez sur la ligne pessimiste — **et dites dans le protocole quelle ligne vous avez retenue.**

6.5 Ce qu'une étude sous-dimensionnée a le droit de conclure

C'est la partie qui change ce qui s'écrit, et il y a exactement trois affirmations permises.

« Aucun effet supérieur à X n'a été détecté. » Avec un EMD de 4,9 points et des différences de différences observées de −0,96 assorties d'un intervalle de −3,5 à +1,6, la phrase honnête est que les effets supérieurs à environ 3,5 points dans un sens ou l'autre sont écartés et les plus petits non.

« L'étude n'était pas conçue pour détecter un effet de l'ampleur qui compte. » Si le programme serait jugé valable à 2 points et que le protocole en détecte 4,9, c'est un constat sur l'évaluation, et il aurait dû être fait avant qu'elle tourne.

« Il faut plus d'unités, et en voici le nombre. » Le tableau ci-dessus, avec la ligne que devrait viser une évaluation future.

Trois affirmations qui ne sont pas permises. « Le programme n'a eu aucun effet » — les données ne peuvent pas l'étayer. « L'effet n'était pas statistiquement significatif, il faut donc arrêter le programme » — une décision reposant sur une absence de preuve. Et le silence sur la puissance, qui laisse le lecteur fournir lui-même les deux premières.

6.6 La puissance a posteriori n'existe pas

```
observed = -0.0096
se = 0.0129
print(f"observed effect {observed * 100:+.2f} points, SE {se * 100:.2f}")
```

```
# Do not compute power from the observed effect. It is a rearrangement of p.
```

La puissance calculée à partir de l'effet observé est une fonction monotone de votre valeur p — elle n'ajoute aucune information et crée le faux sentiment d'avoir vérifié quelque chose. Un résultat non significatif avec une « faible puissance observée » est exactement un résultat non significatif.

La quantité utile est l'intervalle de confiance, comme l'a établi le cours de statistiques, et l'EMD calculé depuis le *protocole* plutôt que depuis le résultat. Les deux sont disponibles ici ; la puissance a posteriori n'en fait pas partie.

6.7 Rapportez-le en entier

```
Statistical power, school feeding evaluation

  Design    15 treated and 9 control schools, ~50 children each.
  Inputs    outcome SD 0.143, ICC 0.065 (estimated from this study's own
            empty mixed model), 5% two-sided, 80% power.
```

```
Design effect 4.19. Minimum detectable effect 4.88 percentage points.

The observed difference-in-differences is -0.96 points (95% CI -3.5 to
+1.6). Effects larger than about 3.5 points in either direction are ruled
out; smaller effects are not, and the design could not have detected them.

This is reported as a null result about large effects. It is not evidence
that the programme has no effect.

Sensitivity: across ICCs from 0.01 to 0.20 the minimum detectable effect
ranges from 2.8 to 7.6 points. The ICC used is estimated from 24 clusters
and is itself imprecise.

A future evaluation targeting a 3-point effect requires about 60 schools.
Adding children to the existing 24 does not help: doubling to 100 per
school moves the minimum detectable effect from 4.73 to 4.45 points.
```

Le dernier paragraphe est celui sur lequel un programme peut agir, et c'est la raison de calculer la puissance après un résultat nul autant qu'avant une étude : il convertit « nous n'avons rien trouvé » en « voici ce que coûterait de trouver quelque chose ».

6.8 La suite

Tous les protocoles de ce cours ont été reconstitués après coup. La leçon suivante le fait dans le bon ordre — écrire le protocole, le résultat, l'analyse et le seuil d'intérêt avant l'arrivée des données, dans un document assez court pour que quelqu'un en produise réellement un.

Quatrième partie

Ce que l'évaluation affirme

CHAPITRE 7

Une page, écrite avant l'arrivée des données

Leçon 7 sur 8 · 180 min

Chaque protocole de ce cours a été reconstitué après coup, et chaque reconstitution contenait un choix que le résultat aurait pu influencer. Cette leçon écrit les choix d'abord — le protocole, le résultat, le seuil et l'analyse — sur une page qu'un programme produira réellement.

7.1 Six choix que le résultat aurait pu faire à votre place

Chaque leçon de ce cours en contenait un, et à chaque fois l'analyste a choisi après avoir vu quelque chose.

Leçon	Le choix	Ce qu'il aurait pu être choisi pour proc
1	Avant-après, coupe transversale ou différences de différences	+6,97, +2,12 ou −0,96 point
2	Quelles covariables comptent pour l'équilibre	Un tableau d'apparence acceptable
3	Erreurs-types groupées ou naïves	Un intervalle excluant zéro
4	Apparier ou non	−0,71 ou −0,96, et quelles écoles sont dedans
5	Quelle fenêtre	Celle qui est significative
6	Quelle CCI	Un EMD sous l'effet observé

Six choix binaires produisent soixante-quatre analyses. Aucun n'est malhonnête et chacun se défend isolément, ce qui est précisément pourquoi la défense doit être déposée avant que le résultat soit connu.

C'est le problème des comparaisons multiples du cours de statistiques sous un autre habit. Là c'étaient vingt-quatre tests ; ici c'est un test choisi parmi soixante-quatre analyses possibles, et la correction n'est pas arithmétique — c'est d'écrire le choix d'avance.

7.2 La page

Ni un protocole, ni un rapport enregistré. **Une page, produite la semaine où l'évaluation est commandée**, parce que l'alternative réaliste est rien du tout.

```
Evaluation plan: school feeding and learning outcomes
Written 2024-01-15, before any endline data exists.

1. QUESTION
   Does the school feeding programme raise literacy scores?

2. DESIGN
   Difference-in-differences. 15 programme schools, 9 comparison schools,
   baseline and endline literacy assessment on the same children.
   Assignment was not randomised; the balance table will be reported.

3. OUTCOME
   Percent correct on the literacy assessment. Baseline is out of 40 items
   and endline out of 50, so raw scores are not comparable and percent
   correct is the primary outcome. Numeracy is secondary.
```

```
4. THRESHOLD OF INTEREST
   3 percentage points. Below that the programme would not be expanded on
   these grounds, so an effect smaller than 3 points is reported as
   "no programme-relevant effect" regardless of its p-value.

5. ANALYSIS
   OLS of the individual gain on a programme indicator, standard errors
   clustered on school. Adjusted for baseline score and district.
   Reported with a 95% confidence interval, not a p-value alone.

6. POWER
   ICC assumed 0.065 from the attendance data. Minimum detectable effect
   4.9 points, which is above the 3-point threshold. The evaluation is
   therefore underpowered for the effect size that matters, and this is
   stated in the report whatever the result.

7. WHAT WOULD CHANGE THE CONCLUSION
   Attrition above 20%, or differential attrition between arms above 5
   points, would make the panel unrepresentative and the primary analysis
   would be reported alongside a bounds analysis.

8. WHAT WE WILL NOT CLAIM
   Causality beyond difference-in-differences. Effects on attendance,
   enrolment or nutrition, which are not measured here.
```

Le point 6 est ce qui rend ce document digne d'être écrit. Il est connaissable en janvier, il dit que l'étude ne peut pas répondre à la question pour laquelle elle a été commandée, et l'apprendre en janvier ne coûte rien tandis que l'apprendre en décembre coûte une année.

7.3 Pourquoi chaque section est là

La question, en une phrase. S'il en faut plus d'une, c'est plus d'une évaluation.

Le protocole, nommé. « Nous comparerons les écoles avec et sans programme » n'est pas un protocole ; ce sont trois protocoles qui donnent trois réponses.

Le résultat, défini à l'item près. L'exercice du cours de statistiques a trouvé une demande comparant une épreuve de 40 items à une de 50. Spécifier « pourcentage de réussite » en janvier est ce qui l'empêche.

Le seuil d'intérêt, en unités de programme. C'est la section que personne n'écrit et c'est celle qui empêche un rien significatif d'être rapporté comme un constat. Demandez au responsable de programme : *quelle ampleur faudrait-il pour que vous étendiez le programme ?*

L'analyse, spécifiée. Y compris les erreurs-types, parce que le choix entre naïves et groupées a déplacé l'intervalle dans chaque leçon du cours de régression.

La puissance, calculée. Avant, non après.

Ce qui changerait la conclusion. Nommer les modes d'échec d'avance signifie que le rapport n'aura pas à débattre de leur anticipation.

Ce que vous n'affirmerez pas. La section la plus courte et celle qui protège la crédibilité de l'évaluation quand quelqu'un d'autre la surlit.

7.4 S'écarter du plan

Les plans se trompent, les données arrivent abîmées, et la réponse n'est pas de faire comme si de rien n'était.

Les écarts sont permis et doivent être listés. Un court tableau dans le rapport — ce qui était prévu, ce qui a été fait, pourquoi — coûte quatre lignes et convertit une incohérence apparente en décision documentée.

```
Deviations from the evaluation plan

  Planned: adjust for baseline score and district.
  Done:    as planned.

  Planned: primary analysis on all enrolled students.
  Done:    restricted to the 585 students with both assessment rounds.
  Why:     156 students have no endline. Attrition analysis added; the
           students who left scored 39.3% at baseline against 56.8% for
           those who stayed, so the panel is not representative and this
           is reported as a limitation.
```

Un écart listé est une force. Un écart non listé, découvert par un relecteur, est la fin de la crédibilité de l'évaluation, et ce qui les sépare fait quatre lignes écrites sur le moment.

7.5 Préspécifier ne veut pas dire ne pas explorer

Deux sections, clairement étiquetées. L'analyse préspécifiée répond à la question posée. Tout le reste est exploratoire, est étiqueté exploratoire, et engendre des hypothèses pour le passage suivant plutôt que des conclusions pour celui-ci.

```
PRIMARY = "gain ~ feeding_programme + baseline + district"   # pre-specified
EXPLORATORY = [                                              # labelled as such
    "gain ~ feeding_programme * sex",
    "gain ~ feeding_programme * baseline_quartile",
    "gain ~ feeding_programme + C(school_id)",
]
print(f"pre-specified: 1 model.  exploratory: {len(EXPLORATORY)} models.")


# Two objects, two headings in the report. That is the whole discipline.
```

Un constat de sous-groupe est exploratoire à moins que le sous-groupe ait été nommé d'avance. Le cours de statistiques a trouvé deux écoles significatives sur vingt-quatre pour un effet exactement nul ; une analyse de sous-groupe choisie après avoir vu les données est la même machine.

7.6 Quand il n'y a pas de plan et que les données sont arrivées

Ce qui est le cas courant, et il n'est pas désespéré.

Écrivez le plan quand même, daté, avant de lancer l'analyse. Vous avez vu les données ; vous n'avez pas encore vu le résultat. Cela vaut mieux que rien et il est honnête de le dire.

Préspécifiez l'analyse principale et rapportez tout le reste comme exploratoire. La distinction garde du sens même tracée tardivement.

Dites combien d'analyses vous avez lancées. Le bloc de rapport du cours de statistiques se termine par « comparaisons lancées au total », et il a sa place ici pour la même raison.

7.7 Rapportez-le en entier

```
Analysis plan and deviations

  The evaluation plan was written on 2024-01-15, before endline data
  collection, and is reproduced in Annex A.

  Primary analysis as pre-specified: difference-in-differences on percent
  correct, clustered on school, adjusted for baseline and district.

  Deviations: one, listed in Annex A. The panel was restricted to students
  with both rounds; an attrition analysis was added.

  Exploratory analyses: 3, reported in Annex B and labelled as exploratory.
  None is presented as a finding.

  Threshold of interest: 3 percentage points, agreed with the programme team
  in January. The observed estimate is -0.96 points (95% CI -3.5 to +1.6)
  and is below the threshold in magnitude.
```

La date de la première ligne est l'élément porteur de tout le document. Tout le reste est une affirmation d'intention ; la date est ce qui la rend vérifiable.

7.8 La suite

Le plan dit ce que l'évaluation affirmera. La dernière leçon écrit le rapport qui en résulte — un rapport dont le constat central est que la question posée ne peut pas être tranchée avec les données existantes, et qui est plus utile que l'alternative.

CHAPITRE 8

L'évaluation qui décline la question

Leçon 8 sur 8 · 180 min

La commande demandait si la cantine scolaire améliore les apprentissages. La réponse honnête est que ce protocole n'aurait pas pu détecter un effet de l'ampleur qui intéresse le programme, que ce qu'il a détecté est nul, et que les deux affirmations doivent figurer dans le résumé et non en annexe.

8.1 Rassemblez ce que le cours a produit

Sept leçons, un programme, un fichier. Voici tout ce que l'analyse a établi.

Question	Réponse	D'où elle vie
Les scores ont-ils monté ?	+6,97 pts, IC à 95 % +6,1 à +7,9	Avant-après, le
Est-ce le programme ?	Non — les écoles témoins ont monté de 7,62	Leçon 1
Les groupes sont-ils comparables ?	Non — six caractéristiques sur six déséquilibrées	Tableau d'équi
Quel est l'effet ?	−0,96 pt, IC à 95 % −3,5 à +1,6	Différences de
L'appariement change-t-il cela ?	−0,71 pt, sur 9 des 15 écoles traitées	Leçon 4
Un meilleur protocole pourrait-il tourner ?	Pas sur ces données — aucun résultat au-delà du seuil	Leçon 5
Qu'aurait-il pu détecter ?	Rien en dessous de 4,9 points	Puissance, leço

Deux de ces sept lignes sont le rapport et les cinq autres sont l'annexe. Le constat est la quatrième ligne et la septième, et c'est la septième qui doit mener.

8.2 Le résumé honnête

```
School feeding and learning outcomes: evaluation summary

  FINDING

  This evaluation cannot answer whether the school feeding programme
  improves literacy, and the reason is in its design rather than its
  results.

  The programme is delivered in 15 schools and compared against 9. With
  fifty children per school and the observed clustering, the smallest
  effect this comparison could have detected is 4.9 percentage points. The
  programme team's threshold for expansion is 3 points. The evaluation was
  therefore incapable of answering the question it was commissioned to
  answer, and this was knowable before it began.

  What it does establish: literacy rose 7.0 points over the school year in
  both programme and comparison schools, with no difference between them
  (-0.96 points, 95% CI -3.5 to +1.6). Effects larger than 3.5 points in
  either direction are ruled out. Effects between 0 and 3 points -- the
  range the programme cares about -- are not.

  RECOMMENDATION
```

> Do not discontinue the programme on the basis of this evaluation. Do not expand it on the basis of the 7-point gain, which is the school year and is present in schools without the programme.
>
> To answer the question, an evaluation needs about 60 schools rather than 24. If the programme is expanding to new schools, randomising the order of the roll-out costs nothing and produces a comparison group that does not have to be argued for.

Le premier paragraphe refuse la question et le dernier dit ce qui y répondrait. Entre les deux il y a un constat, un intervalle et une borne, et aucun nombre n'est présenté comme un effet alors qu'il n'en est pas un.

8.3 Quatre phrases qu'une évaluation doit pouvoir écrire

Chacune est une vérification du rapport, et un rapport qui ne peut pas produire les quatre a une lacune.

« Le contrefactuel est X. » Ici : les écoles sans programme, supposées avoir suivi la même trajectoire.

« L'estimation vaut Y, avec l'intervalle Z, dans les unités qu'emploie le programme. » Ici : −0,96 point de pourcentage de littératie, −3,5 à +1,6, où 3 points est le seuil d'intérêt.

« Le plus petit effet que ce protocole pouvait détecter vaut W. » Ici : 4,9 points. **C'est la phrase qui manque à la plupart des évaluations**, et son absence est ce qui permet de lire un résultat nul comme une réfutation.

« Cette analyse ne peut pas écarter V. » Ici : tout effet entre zéro et environ trois points, c'est-à-dire toute la plage sur laquelle le programme discute.

8.4 Ce que vaut « aucun effet détecté »

Cela vaut quelque chose, et être précis sur combien fait la différence entre un résultat nul utile et un résultat nul dommageable.

Lecture	Étayée ?
« Le programme ne marche pas »	**Non** — le protocole ne pouvait pas détecter l'effet per
« Le programme n'a pas de grand effet sur la littératie »	**Oui** — au-delà de 3,5 points est écarté
« Le gain de 7 points n'est pas attribuable au programme »	**Oui** — les écoles témoins ont monté davantage
« Il faut réorienter le financement »	**Non** — cela exige une estimation d'effet que cette étu
« La prochaine évaluation a besoin de 60 écoles »	**Oui** — calculé, et actionnable

Les deux lignes « oui » du milieu sont de vrais constats et ce sont des constats importuns : ils retirent un chiffre phare qu'un programme employait. Le livrer est le travail, et le faire aux côtés de la recommandation de la dernière ligne est ce qui le rend recevable.

8.5 Écrire pour le lecteur qui citera une seule phrase

Quelqu'un extraira une ligne de ce rapport et la mettra sur une diapositive. Décidez maintenant laquelle.

Pas : « Aucun effet statistiquement significatif de la cantine scolaire sur la littératie n'a été trouvé. » Vrai, et cela sera lu comme « le programme ne marche pas ».

Pas : « La littératie a monté de 7 points dans les écoles du programme. » Vrai, et cela sera lu comme l'effet.

Ceci : « La littératie a monté de 7 points dans les écoles avec et sans programme ; cette évaluation n'était pas assez grande pour détecter la différence de 3 points qui intéresse le programme. »

Mettez cette phrase dans le résumé, dans la conclusion et dans le courriel d'accompagnement. La phrase citable est choisie par vous ou elle est choisie pour vous.

8.6 Ce dont ce cours a traité

Huit leçons et une idée : **une affirmation d'impact est une comparaison dont une moitié est absente, et le protocole est l'argumentation sur ce qui la comble.**

Chaque méthode présentée — randomisation, différences de différences, appariement, discontinuité — est une argumentation différente pour la même quantité manquante, et chacune est plus ou moins solide selon des faits concernant le déploiement du programme plutôt que selon quoi que ce soit dans l'analyse.

Ce qui signifie que les décisions les plus lourdes de conséquences ont été prises avant l'existence de la moindre donnée : qui a reçu le programme et dans quel ordre, ce qui a été mesuré et sur qui, et combien d'unités ont été affectées. Un analyste qui arrive après choisit parmi les argumentations que le déploiement a laissées disponibles.

La chose utile que ce cours demande n'est donc pas une technique. C'est d'être dans la pièce en janvier, avec une page, à demander comment les écoles seront choisies et si l'ordre peut être randomisé — une question qui ne coûte rien à poser et devient sans réponse un an plus tard.

8.7 Rapportez-le en entier

```
Evaluation of the school feeding programme: methods and limitations

  Design            Difference-in-differences, 15 programme and 9
                    comparison schools, baseline and endline literacy on
                    585 students with both rounds.

  Counterfactual    Comparison schools, assumed to have been on the same
                    trajectory. Two rounds exist, so parallel trends is
                    argued (numeracy shows the same null; no single school
                    drives the result) and not tested.

  Estimate          -0.96 percentage points, 95% CI -3.5 to +1.6,
                    clustered on 24 schools.

  Power             Minimum detectable effect 4.9 points at 80% power,
                    against a 3-point threshold of interest. Underpowered
                    for the relevant effect size.

  Balance           All six baseline characteristics imbalanced beyond 0.25
                    standardised. Adjusted for baseline and district;
                    unmeasured selection is not addressed.

  Attrition         585 of 741 baseline students have an endline. Those who
                    left scored 39.3% at baseline against 56.8% for those
                    who stayed.

  Not claimed       Any effect on attendance, enrolment or nutrition. Any
```

```
                    causal effect beyond the difference-in-differences
                    assumption. Any conclusion about effects below 3 points.

Plan                Written 2024-01-15, one deviation listed in Annex A,
                    three exploratory analyses in Annex B.
```

Huit lignes, et un relecteur peut reconstituer chaque décision à partir d'elles. C'est le livrable vers lequel ce cours a construit : non pas un nombre, mais un nombre entouré de tout ce dont un lecteur a besoin pour savoir ce qu'il est.

8.8 La suite

Le module 5 est complet : l'incertitude, puis les modèles, puis les protocoles. Le module 6 porte sur la restitution — transformer ce que vous avez établi en quelque chose qu'un programme peut voir, employer et reconstruire, c'est-à-dire la dernière chose qui sépare une analyse d'une décision.

À propos de cette édition

Ce PDF est produit à partir de la même source que la version web de la formation ; les deux ne peuvent donc pas diverger. L'édition en ligne comporte le code exécutable, les jeux de données téléchargeables et les corrections publiées depuis la production de ce fichier.

> Tous les jeux de données cités dans cette formation sont synthétiques ou entièrement anonymisés. Aucun enregistrement ne renvoie à une personne réelle.

Cette formation est publiée en anglais et en français. L'édition française est une adaptation professionnelle reprenant la terminologie employée par l'UNICEF, l'OMS, le PAM, OCHA et les ministères nationaux — et non une traduction littérale.

Consulter et exécuter la formation en ligne sur data-analysis.cassion.dev/fr/cours/impact-evaluation-methods

Généré le 29 juillet 2026.