CASSION · FORMATION

Analyse de programmes d'éducation

Inscriptions, présence, rétention et résultats d'apprentissage, avec le calendrier scolaire et la distinction inscription/présence qui met en défaut tous les tableaux de bord.

Formateur	Pierre Robentz Cassion
Niveau	Intermédiaire
Heures apprenant	14 h
Enseignée en	Python · R
Mise à jour	28 juillet 2026
Secteurs	Éducation
Normes et méthodologies	Objectifs de développement durable (ODD) · Définitions d'indicateurs de l'UNICEF · Enquête par grappes à indicateurs multiples (MICS)

Consulter et exécuter la formation en ligne sur data-analysis.cassion.dev/fr/cours/education-programme-analysis

Ce que vous saurez faire

— Calculer les taux bruts et nets de scolarisation et expliquer pourquoi l'un dépasse 100 % sans qu'aucun enfant ne soit compté deux fois

— Distinguer l'assiduité moyenne quotidienne de la proportion d'élèves assidus, et dire sur laquelle un programme est jugé

— Suivre une cohorte à travers passage, redoublement et abandon, et calculer la survie jusqu'à la dernière année

— Lire une évaluation des apprentissages sur deux passations aux instruments différents, et dire ce qui est comparable

— Détecter un sous-échantillon d'évaluation non aléatoire et encadrer le biais qu'il introduit

— Désagréger par sexe, handicap et déplacement, et énoncer l'échantillon qu'exige chaque ventilation

Prérequis

Aucun. Cette formation ne suppose aucune expérience préalable de la programmation.

Sommaire

À propos de cette formation

L'éducation est le secteur où l'indicateur le plus cité est le plus souvent calculé sur le mauvais dénominateur, et où la distinction dont dépend un programme — inscrit contre présent contre apprenant — est ramenée sur presque tous les tableaux de bord à un nombre unique appelé *couverture*.

Le cours s'appuie sur quatre fichiers. Le **registre de présence** et sa liste d'élèves sont apparus dans trois cours antérieurs comme problème de nettoyage et de jointure ; ils sont ici analysés comme de la présence. Trois sont nouveaux : un **registre d'inscriptions sur deux ans** avec âge, sexe, handicap, déplacement et une issue de fin d'année ; la **population d'âge scolaire** qui en est le dénominateur ; et une **évaluation des apprentissages en deux passations**.

Trois résultats structurent le cours, tous calculés à partir de ces fichiers.

Le taux brut de scolarisation vaut 109,2 % et le taux net 97,4 %. Aucun n'est faux et aucun n'est l'autre. 36,4 % des effectifs du primaire dépassent l'âge officiel de leur année et un enfant sur neuf a douze ans ou plus tout en étant encore au primaire — c'est ce qui porte le taux brut au-dessus de 100 % sans qu'un seul enfant soit compté deux fois.

L'assiduité moyenne quotidienne vaut 88,4 % et 62,1 % des élèves sont présents au moins 90 % des jours où ils ont été pointés. Un programme qui rapporte le premier décrit un système ; un programme qui agit sur le second décrit des enfants.

L'abandon atteint 19,5 % chez les élèves en retard d'âge contre 4,8 % chez les autres. Être en retard est le prédicteur le plus fort de ce registre, et il se compose : 94,5 % des enfants ayant redoublé en 2023 sont en retard d'âge en 2024, et un enfant en retard d'âge a quatre fois plus de chances de partir.

Python et R tout du long. Vous devriez avoir suivi le module 3 — ce cours suppose que vous savez défendre un dénominateur, lire un registre censuré et poser un intervalle sur une proportion.

Première partie

L'inscription et son dénominateur

CHAPITRE 1

109 % scolarisés, 97 % scolarisés

Leçon 1 sur 8 · 105 min

Deux taux issus d'un même registre. Le brut vaut 109,2 % et le net 97,4 %, les deux sont exacts, et les douze points d'écart sont des enfants scolarisés au mauvais âge plutôt que des enfants absents.

1.1 Deux taux, un registre

```python
import pandas as pd

enrolment = pd.read_csv("school-enrolment-2024.v1.csv")
population = pd.read_csv("school-age-population-2024.v1.csv")

current = enrolment[enrolment["school_year"] == 2024]
primary = current[current["grade"].between(1, 6)]

denominator = population.loc[
    (population["reference_year"] == 2024)
    & population["age_years"].between(6, 11),
    "projected_population",
].sum()

gross = len(primary) / denominator
net = len(primary[primary["age_years"].between(6, 11)]) / denominator

print(f"gross enrolment ratio: {gross:.1%}")
print(f"net enrolment ratio:   {net:.1%}")
```

```r
library(dplyr)

primary <- enrolment |> filter(school_year == 2024, between(grade, 1, 6))
denominator <- population |>
  filter(reference_year == 2024, between(age_years, 6, 11)) |>
  summarise(n = sum(projected_population)) |> pull(n)

tibble(
  gross = nrow(primary) / denominator,
  net = sum(between(primary$age_years, 6, 11)) / denominator
)
```

Taux	Numérateur	Dénominateur	Valeur
Brut	Tous les inscrits au primaire, tout âge	Enfants de 6 à 11 ans	**109,2 %**
Net	Inscrits au primaire de 6 à 11 ans	Enfants de 6 à 11 ans	**97,4 %**

Les numérateurs diffèrent ; le dénominateur est identique. C'est toute la définition, et c'est pourquoi les deux taux ne sont pas deux estimations d'une même chose.

1.2 Pourquoi un taux peut dépasser 100 %

Un taux brut supérieur à 100 % n'est pas une erreur. Trois choses le produisent et une seule est un défaut.

Le retard et l'avance d'âge. Les enfants hors de la tranche d'âge officielle sont au numérateur et pas au dénominateur. C'est la cause dominante ici et c'est un trait réel du système, non un problème de données.

Un dénominateur trop petit. Le fichier de population est une projection et non un dénombrement, et une projection qui sous-estime gonfle tout taux bâti dessus.

Le double comptage. Un enfant inscrit dans deux écoles figure deux fois. Celui-là *est* un défaut, et ce registre l'a.

```
duplicates = current.duplicated(subset=["student_id", "school_year"], keep=False)
print(f"student-years appearing more than once: {duplicates.sum()}")

unique = current.drop_duplicates(subset=["student_id", "school_year"])
unique_primary = unique[unique["grade"].between(1, 6)]
print(f"gross ratio after deduplication: {len(unique_primary) / denominator:.1%}")
```

```
enrolment |> filter(school_year == 2024) |>
  count(student_id) |> filter(n > 1) |> nrow()
```

Douze élèves figurent deux fois, une fois sous chaque école, parce qu'un transfert a été saisi comme une nouvelle inscription plutôt que comme un déplacement. **Dédupliquez sur élève et année avant l'un ou l'autre taux**, et notez que l'effet est ici faible — l'enjeu n'est pas sa taille, c'est qu'un registre qui compte double se trompe sur qui existe.

1.3 Le dénominateur est une projection

```
CENSUS_YEAR, GROWTH = 2015, 0.021
factor = (1 + GROWTH) ** (2024 - CENSUS_YEAR)
print(f"nine years compounded at {GROWTH:.1%}: factor {factor:.3f}")
print(f"so about {1 - 1 / factor:.0%} of the denominator is an assumption")
```

```
(1 + 0.021)^(2024 - 2015)
```

Un facteur de 1,21, donc environ un sixième du dénominateur n'a jamais été compté. C'est le dénominateur vaccinal du cours de santé publique, dans un autre secteur : une base de recensement, une hypothèse de croissance, et neuf ans de composition.

La conséquence est précise. **Si la projection est trop basse de 5 %, le taux brut passe de 109,2 % à environ 104 %** et le net de 97,4 % à environ 93 %. Aucune conclusion ne change, mais un rapport affirmant que la scolarisation a gagné deux points d'une année à l'autre rapporte peut-être le taux de croissance et non les écoles.

```
for error in (-0.05, 0.0, 0.05):
    adjusted = denominator * (1 + error)
    print(f"projection {error:+.0%}: gross {len(primary) / adjusted:.1%}, "
          f"net {len(primary[primary['age_years'].between(6, 11)]) / adjusted:.1%}")
```

```
# Vary the denominator and see which conclusions survive.
```

Montrez la sensibilité plutôt que l'estimation ponctuelle là où le dénominateur est projeté. Cela coûte trois lignes et c'est la différence entre un taux et un taux défendable.

1.4 Lequel rapporter

Aucun, seul.

Le taux net répond à « les enfants d'âge scolaire sont-ils scolarisés ». C'est le cadrage de l'ODD 4.1 et le bon taux pour une question de couverture. Il ne peut pas dépasser 100 %, ce qui en fait le nombre le plus sûr à publier.

Le taux brut répond à « quelle quantité de scolarité primaire est délivrée ». C'est le bon taux pour une question de capacité — enseignants, salles, manuels — car un enfant en retard d'âge a besoin d'un pupitre exactement autant qu'un enfant à l'âge attendu.

L'écart entre les deux est le constat, et c'est la raison de rapporter les deux : douze points de retard d'âge sont un énoncé sur le redoublement et l'entrée tardive qu'aucun des deux taux ne fait seul.

```
Primary enrolment, 2024, four districts

  Gross enrolment ratio       109.2%   1,052 enrolees / 963 children aged 6-11
  Net enrolment ratio          97.4%     938 in-age enrolees / same denominator
  Over-age share of enrolment  36.4%   above the official age for their grade

  Denominator is a projection from the 2015 census at 2.1% a year. A 5%
  error in it moves the gross ratio by about five points and changes no
  conclusion.
  12 student-years were recorded twice after transfers and are deduplicated.
```

1.5 Ce que les taux ne peuvent pas dire

Aucun n'est la présence. Un enfant inscrit et jamais présent figure aux deux numérateurs. L'unité suivante porte sur cette distinction, et elle vaut plus que l'un ou l'autre taux.

Aucun n'est l'achèvement. L'inscription est un stock à un moment de l'année ; savoir si ces enfants terminent est une question de cohorte et la troisième unité.

Aucun n'est l'apprentissage. Un système peut scolariser chaque enfant d'âge scolaire et n'en instruire aucun, et la dernière unité porte sur l'instrument qui s'en apercevrait.

1.6 La suite

Douze points de l'écart entre les deux taux sont du retard d'âge. La leçon suivante porte sur qui sont ces enfants, pourquoi le retard est le prédicteur le plus fort de ce registre, et comment le redoublement le fait se composer.

CHAPITRE 2

En retard, et de plus en plus

Leçon 2 sur 8 · 105 min

Le retard d'âge vaut 17,0 % en première année et 50,8 % en sixième. Il monte parce que le redoublement le crée — 94,5 % des enfants ayant redoublé en 2023 sont en retard d'âge en 2024 — et parce qu'être en retard quadruple le risque de partir.

2.1 La définition, et le contrôle d'étendue qui doit venir avant

Un élève est en retard d'âge lorsque son âge dépasse l'âge officiel, qui vaut ici le numéro de l'année plus cinq.

```python
import pandas as pd

enrolment = pd.read_csv("school-enrolment-2024.v1.csv")
print(enrolment["age_years"].describe().round(1))
print(f"ages above 20: {(enrolment['age_years'] > 20).sum()}")
```

```r
library(dplyr)
enrolment |> summarise(min = min(age_years), max = max(age_years),
                       impossible = sum(age_years > 20))
```

Vingt et une lignes portent un âge supérieur à 20 dans un registre du primaire. SCH04 enregistre une partie de ses effectifs en mois, si bien qu'un enfant de neuf ans apparaît à 112. Une moyenne les absorberait ; un contrôle d'étendue les trouve, et la correction est une division et non une suppression.

```python
clean = enrolment[enrolment["age_years"] <= 20].copy()
clean["official_age"] = clean["grade"] + 5
clean["over_age"] = clean["age_years"] > clean["official_age"]
```

```r
enrolment |> filter(age_years <= 20) |>
  mutate(official_age = grade + 5, over_age = age_years > official_age)
```

Faites le contrôle d'étendue avant l'indicateur, non après. Un âge de 112 est en retard pour toutes les années, si bien qu'un indicateur calculé sur la colonne brute est techniquement exact pour vingt et un enfants et faux sur ce qu'il signifie.

2.2 Il se compose avec l'année d'études

```python
primary = clean[(clean["school_year"] == 2024) & clean["grade"].between(1, 6)]
by_grade = primary.groupby("grade").agg(
    students=("student_id", "size"),
    over_age=("over_age", "mean"),
)
print((by_grade * [1, 100]).round(1))
```

```
primary |> summarise(n = n(), over_age = mean(over_age), .by = grade)
```

Année	Âge officiel	Élèves	En retard d'âge
1	6	53	**17,0 %**
2	7	153	25,5 %
3	8	247	29,6 %
4	9	258	36,8 %
5	10	209	47,8 %
6	11	132	**50,8 %**

Le retard d'âge triple entre la première et la sixième année. Deux mécanismes produisent cette forme et ce ne sont pas le même problème.

L'entrée tardive place un enfant au-dessus de l'âge dès la première année et l'y maintient. Elle apparaît comme le plancher de 17,0 %, et l'intervention est l'inscription précoce.

Le redoublement crée le retard d'âge *pendant* la scolarité. Il apparaît comme la montée, et l'intervention est entièrement différente.

```
previous = clean[clean["school_year"] == 2023].set_index("student_id")
current = clean[clean["school_year"] == 2024].copy()
current["last_year"] = current["student_id"].map(previous["end_of_year_status"])

print(current.groupby("last_year")["over_age"].agg(["mean", "size"]).round(3))
```

```
enrolment |>
  filter(age_years <= 20) |>
  select(student_id, school_year, grade, age_years, end_of_year_status) |>
  tidyr::pivot_wider(names_from = school_year,
                     values_from = c(grade, age_years, end_of_year_status))
```

94,5 % des élèves ayant redoublé en 2023 sont en retard d'âge en 2024, contre 30,6 % de ceux qui ont été promus. Le redoublement ne corrèle pas avec le retard d'âge ; il le *cause*, avec un an de décalage, et le registre permet d'observer le mécanisme plutôt que de l'inférer.

Notez que la comparaison doit se faire entre années. En 2023 un redoublant n'est pas encore en retard d'âge — il est dans son année à l'âge normal, et l'année supplémentaire apparaît en septembre suivant.

2.3 Pourquoi cela compte : le lien avec l'abandon

```
year23 = clean[clean["school_year"] == 2023]
dropout = year23.groupby("over_age")["end_of_year_status"].apply(
    lambda s: (s == "dropped-out").mean()
)
counts = year23.groupby("over_age").size()
print(pd.DataFrame({"dropout": (dropout * 100).round(1), "n": counts}))
```

```
enrolment |> filter(school_year == 2023) |>
  summarise(dropout = mean(end_of_year_status == "dropped-out"),
            n = n(), .by = over_age)
```

	Élèves	Ont abandonné
En retard d'âge	478	**19,5 %**
À l'âge attendu	814	4,8 %

Être en retard quadruple le risque de partir, et c'est ce qui boucle la chaîne : un enfant redouble, prend du retard d'âge, et devient alors bien plus susceptible d'abandonner que ce que le redoublement visait à prévenir.

Le redoublement n'est donc pas une intervention neutre. Il est présenté comme une seconde chance et il augmente de façon mesurable la probabilité de partir tout à fait. Ce constat s'obtient à partir de deux colonnes d'un registre, et c'est l'argument le plus fort que ce jeu de données soutient.

2.4 La désagrégation qui ne donne rien

Le retard d'âge est inhabituel parmi les indicateurs éducatifs en ce que les ventilations évidentes montrent très peu.

```python
for column in ("sex", "disability_reported", "displacement_status"):
    print(primary.groupby(column, dropna=False)["over_age"].agg(
        ["mean", "size"]).round(3), "\n")
```

```r
primary |> summarise(over_age = mean(over_age), n = n(), .by = sex)
```

Ventilation	En retard d'âge	n
Garçons	39,2 %	510
Filles	33,8 %	542
Handicap signalé	33,7 %	89
Sans handicap signalé	36,7 %	963

Cinq points entre garçons et filles, sur 510 et 542 élèves. C'est à la limite de ce que produit la variation d'échantillonnage, donc calculez l'intervalle avant d'écrire la phrase — la machinerie du cours sur les enquêtes, sur exactement le type d'écart qui se publie comme un constat et disparaît au passage suivant.

Le handicap montre trois points en sens inverse sur 89 élèves, ce qui est du bruit sur un dénominateur aussi petit.

Une désagrégation qui ne montre presque rien est un résultat. Le retard d'âge est ici un phénomène de système porté par le redoublement et l'entrée tardive, non une inéquité de groupe — et le dire est plus utile que de chasser un sous-groupe jusqu'à en trouver un.

2.5 Rapportez le profil, non le chiffre principal

```
Over-age enrolment, primary, 2024

  Overall                    36.4%   383 of 1,052 primary enrolees
  Grade 1                    17.0%   the late-entry floor
  Grade 6                    50.8%   after five years of repetition
  Repeated in 2023           94.5%   over-age in 2024, against 30.6% promoted

  Dropout, over-age          19.5%   against 4.8% in-age (2023 cohort)
```

```
21 ages recorded in months at SCH04, corrected by division.
Sex difference is 39.2% (boys) against 33.8% (girls); report with an
interval or not at all.
```

Le profil par année est le livrable, car 36,4 % au total pourrait signifier une entrée tardive uniforme ou un redoublement qui s'accumule, et ces deux cas appellent des programmes différents.

2.6 La suite

Tous les enfants de cette leçon sont inscrits. Savoir s'ils sont en classe relève d'un autre registre et d'un autre nombre — et la leçon suivante en trouve deux, distants de vingt-six points, dans le même fichier.

Deuxième partie

Inscrit n'est pas présent

CHAPITRE 3

88 % de présence, 62 % des élèves

Leçon 3 sur 8 · 105 min

L'assiduité moyenne quotidienne vaut 88,4 %. La part des élèves présents au moins 90 % de leurs jours pointés vaut 62,1 %. Les deux viennent des mêmes 70 245 lignes, et les vingt-six points d'écart sont les enfants qu'une moyenne ne voit pas.

3.1 Un registre, deux questions

```python
import pandas as pd

attendance = pd.read_csv("school-attendance-2024.v1.csv")
print(f"{len(attendance):,} rows, {attendance['student_id'].nunique():,} students")
print(attendance["present"].value_counts(dropna=False))
```

```r
library(dplyr)
attendance |> count(present)
```

La colonne `present` a cinq valeurs, non deux. SCH09 a saisi `Y` et `N` au lieu de `true` et `false` sur 873 de ses lignes, et 271 jours n'ont jamais été pointés dans un sens ni dans l'autre. Une conversion booléenne transforme ces 873 observations réelles en valeurs manquantes et retire discrètement une école du dénominateur.

```python
MARKS = {"true": True, "Y": True, "false": False, "N": False}
marked = attendance[attendance["present"].isin(MARKS)].copy()
marked["attended"] = marked["present"].map(MARKS)
print(f"usable marks: {len(marked):,} of {len(attendance):,}")
```

```r
attendance |>
  mutate(attended = case_when(present %in% c("true", "Y") ~ TRUE,
                              present %in% c("false", "N") ~ FALSE)) |>
  filter(!is.na(attended))
```

Le même fichier répond maintenant à deux questions différentes.

3.2 L'assiduité moyenne quotidienne

```python
ada = marked["attended"].mean()
print(f"average daily attendance: {ada:.1%}")
```

```r
marked |> summarise(ada = mean(attended))
```

88,4 %. Chaque pointage compte une fois, si bien qu'un élève présent 60 jours sur 60 et un élève présent 30 sur 60 contribuent à proportion de leur nombre de pointages.

C'est le nombre qu'un ministère rapporte et ce qu'un indicateur de système doit être. Il répond à *à quel point la classe est remplie un jour moyen*, ce qui est la bonne question pour les effectifs, la cantine et les manuels.

3.3 La proportion d'élèves assidus

```python
by_student = marked.groupby("student_id")["attended"].agg(["mean", "size"])
eligible = by_student[by_student["size"] >= 20]

for threshold in (0.80, 0.85, 0.90):
    share = (eligible["mean"] >= threshold).mean()
    print(f"attending at least {threshold:.0%} of marked days: {share:.1%}")
```

```r
marked |>
  summarise(rate = mean(attended), days = n(), .by = student_id) |>
  filter(days >= 20) |>
  summarise(across(everything(), ~ mean(rate >= 0.9)))
```

Seuil	Élèves l'atteignant
Au moins 80 % des jours	87,3 %
Au moins 85 %	78,9 %
Au moins 90 %	**62,1 %**

62,1 % au seuil de 90 %, contre une assiduité moyenne quotidienne de 88,4 %. Les deux nombres sont écartés de vingt-six points et aucun n'est faux.

Une moyenne sur des pointages décrit le système ; une proportion sur des élèves décrit des enfants. Une école où chaque enfant manque un jour sur huit et une école où sept enfants sur huit sont présents parfaitement tandis qu'un ne vient jamais ont la même assiduité moyenne quotidienne et des problèmes entièrement différents.

3.4 Sur lequel un programme est jugé

Rapportez les deux, et mettez en tête celui qui correspond à la décision.

La décision	Le nombre
Combien de repas, de pupitres, de manuels	Assiduité moyenne quotidienne
Quels enfants ont besoin d'une visite de suivi	Proportion sous le seuil
Si une intervention a marché	Les deux, car ils peuvent bouger en sens opposés

Cette dernière ligne est celle qu'il faut intérioriser. **Une intervention qui fait passer les plus mauvais assidus de 40 % à 60 % ne déplace presque pas l'assiduité moyenne quotidienne et ne déplace pas du tout la proportion au-dessus de 90 %** — et serait enregistrée comme un échec par l'un ou l'autre nombre pris seul.

```python
struggling = eligible[eligible["mean"] < 0.75]
print(f"students below 75%: {len(struggling)} "
      f"({len(struggling) / len(eligible):.1%})")
print(f"their marks as a share of all marks: "
      f"{struggling['size'].sum() / eligible['size'].sum():.1%}")
```

```
# The students furthest behind are a small share of the rows and the whole
# of the problem.
```

3.5 Énoncez le seuil, et d'où il vient

Le seuil de 90 % est une convention, non un standard, et il fait l'essentiel du travail dans ces 62,1 %. Portez-le à 85 % et le chiffre devient 78,9 %.

```
sensitivity = {f"{t:.0%}": f"{(eligible['mean'] >= t).mean():.1%}"
               for t in (0.75, 0.80, 0.85, 0.90, 0.95)}
print(sensitivity)
```

```
# Print the curve, not the point.
```

À 95 % il vaut 39,8 % et à 75 % il vaut 90,5 %. **Publiez le seuil à côté du nombre, à chaque fois.** Deux rapports citant « la proportion d'élèves assidus » à des seuils différents produisent des chiffres incomparables qui ont tous deux l'air officiel — le même défaut que les deux jeux de seuils du score de consommation alimentaire, dans un autre secteur.

3.6 La décision de dénominateur que personne ne prend explicitement

`by_student` ci-dessus a été filtré aux élèves ayant au moins vingt jours pointés. C'est un choix et il doit être déclaré.

Sans lui, un élève inscrit la dernière semaine du trimestre et présent trois jours sur trois apparaît comme assidu à 100 %, et un élève pointé deux fois apparaît dans la queue de distribution.

Avec lui, vous avez exclu précisément les élèves arrivés tard et partis tôt, qui sont ceux qu'un programme d'assiduité veut le plus voir.

```
print(f"students with any marks:      {len(by_student)}")
print(f"students with 20 or more:     {len(eligible)}")
```

```
# Two denominators, both defensible, and the report says which.
```

Ici les deux sont presque identiques, si bien que le choix ne déplace pas la réponse. **Dans un registre couvrant une année entière il la déplacerait beaucoup**, et l'habitude de le déclarer est ce qui rend le chiffre transportable.

3.7 Rapportez la paire

```
Attendance, February to April 2024

  Average daily attendance         88.4%   69,974 usable marks
  Students attending 90%+ of days  62.1%   1,200 students with 20+ marks
  Students attending 85%+          78.9%
  Students below 75%                9.5%   114 students, the follow-up list

  SCH09 recorded 873 marks as Y/N rather than true/false; recoded, not
  dropped. 271 days were never marked and are excluded from both figures.
```

3.8 La suite

Les deux nombres de cette leçon divisent par des jours qui ont été pointés. La leçon suivante porte sur les jours qui ne sont pas du tout dans le fichier — quinze, dans deux écoles, qui transforment une grève en urgence d'abandon si on les laisse faire.

CHAPITRE 4

Une grève qui ressemble à une urgence

Leçon 4 sur 8 · 105 min

SCH07 est à 88,2 % de présence et SCH18 à 85,2 %. Construisez le registre comme un calendrier complet et comptez chaque jour manquant comme une absence, et ils affichent 66,1 % et 63,9 % — les pires écoles du district, sur quinze jours où personne n'était censé être là.

4.1 Les lignes qui n'y sont pas

Les lignes de présence n'existent que pour les jours où une école était ouverte. **Une date sans ligne est une fermeture, non une absence**, et rien dans le fichier ne dit laquelle.

```python
import pandas as pd

attendance = pd.read_csv("school-attendance-2024.v1.csv")
roster = pd.read_csv("school-roster-2024.v1.csv")

joined = attendance.merge(roster[["student_id", "school_id"]], on="student_id")
calendar = sorted(attendance["attendance_date"].unique())

days_per_school = joined.groupby("school_id")["attendance_date"].nunique()
print(f"school days in the calendar: {len(calendar)}")
print(days_per_school.sort_values().head())
```

```r
library(dplyr)

attendance |>
  left_join(roster, by = "student_id") |>
  summarise(days = n_distinct(attendance_date), .by = school_id) |>
  arrange(days)
```

Vingt-deux écoles ont les soixante jours. SCH07 et SCH18 en ont quarante-cinq.

```javascript
missing = {
    school: sorted(set(calendar) - set(group["attendance_date"]))
    for school, group in joined.groupby("school_id")
}
for school, dates in missing.items():
    if dates:
        print(f"{school}: {len(dates)} days, {dates[0]} to {dates[-1]}")
```

```sql
# Which dates, not just how many. The pattern is the diagnosis.
```

Aux deux il manque exactement du 11 au 29 mars, et les mêmes quinze jours. Ce ne sont pas deux écoles avec un problème d'assiduité ; c'est un seul événement.

Des jours manquants consécutifs dans plusieurs écoles sont une fermeture jusqu'à preuve du contraire. Des jours manquants épars dans une seule école sont un problème d'enregistrement. La forme dit laquelle, et la regarder prend une ligne.

4.2 Ce qui arrive si on ne regarde pas

La façon naturelle de bâtir un tableau de présence est de construire la grille complète élèves × jours d'école et d'y reporter les pointages. C'est aussi la façon de transformer une grève en crise.

```
def attendance_rate(school, missing_counts_as_absent):
    students = roster.loc[roster["school_id"] == school, "student_id"]
    marks = attendance[attendance["student_id"].isin(students)]
    present = marks["present"].isin(["true", "Y"]).sum()
    if missing_counts_as_absent:
        denominator = len(students) * len(calendar)
    else:
        denominator = marks["present"].isin(["true", "Y", "false", "N"]).sum()
    return present / denominator

for school in ("SCH07", "SCH18", "SCH01"):
    observed = attendance_rate(school, False)
    filled = attendance_rate(school, True)
    print(f"{school}: observed {observed:.1%}, grid-filled {filled:.1%}")
```

```
# Two denominators: marks made, and student-days in the calendar.
```

École	Sur les pointages faits	Sur une grille complète
SCH07	88,2 %	**66,1 %**
SCH18	85,2 %	**63,9 %**
SCH01	91,3 %	91,3 %

Vingt-deux points, inventés. Sur les chiffres de la grille complète, SCH07 et SCH18 sont de loin les deux pires écoles du district, et un programme lisant ce tableau enverrait une réponse d'urgence contre l'abandon à deux écoles dont les enfants sont venus normalement chaque jour où on le leur a demandé.

L'école non touchée est inchangée, ce qui rend l'erreur si difficile à attraper : le tableau a l'air correct, l'essentiel *l'est*, et les deux lignes fausses sont les deux sur lesquelles on agit.

4.3 La règle

Construisez le dénominateur à partir des jours d'ouverture de l'école, non du calendrier.

```
open_days = joined.groupby("school_id")["attendance_date"].nunique()
enrolled = roster.groupby("school_id")["student_id"].nunique()
expected = (open_days * enrolled).rename("student_days_expected")

actual = joined.groupby("school_id").size().rename("marks_made")
coverage = (actual / expected).rename("mark_coverage")
print(pd.concat([expected, actual, coverage.round(3)], axis=1).head())
```

```
# Expected student-days uses each school's own open days.
```

Cela donne un second nombre utile : la **couverture de pointage**, la part des jours-élèves attendus qui portent effectivement un pointage. Une école à 100 % de présence et 60 % de couverture de pointage n'est pas une école à bonne assiduité.

4.4 Quand une fermeture est le constat

Quinze jours d'école font un quart du trimestre. **La fermeture pèse plus que tout chiffre de présence de ce fichier**, et une analyse qui exclut correctement ces jours puis n'en dit rien a retiré la plus grande chose qui se soit produite.

```
Attendance, February to April 2024

  Average daily attendance, all schools    88.4%   on marks made
  SCH07                                    88.2%
  SCH18                                    85.2%

  SCH07 and SCH18 were closed for the fifteen school days from 11 to 29
  March, a quarter of the term. Their attendance figures are computed on the
  45 days they were open and are comparable with other schools; their
  instructional time is not.

  Counting the closure days as absences would report these two schools at
  66.1% and 63.9% and rank them worst in the district.
```

Rapportez la fermeture en jours d'enseignement perdus, séparément de la présence. Les deux répondent à des questions différentes : la présence porte sur la venue des enfants, les jours d'enseignement sur le fait que l'école a eu lieu.

4.5 La forme générale

C'est la troisième fois que cette plateforme rencontre le même défaut dans un fichier différent.

Cours	La chose absente	Ce à quoi cela ressemblait
Données de routine et DHIS2	Une formation qui n'a pas rapporté	Un district à couverture en baisse
Analyse EAH	Une tournée que personne n'a conduite	Un district à fonctionnalité en hausse
Éducation	**Un jour où l'école était fermée**	**Deux écoles en urgence d'abandon**

Dans les trois cas, l'absence n'est pas une valeur et rien dans le fichier ne l'annonce. L'habitude qui attrape les trois est la même : avant de calculer un taux, construisez ce que le dénominateur *devrait* être à partir d'autre chose que des lignes dont vous disposez, et comparez.

4.6 La suite

La présence décrit un trimestre. Savoir si un enfant est encore scolarisé l'année suivante est une autre question, et l'unité suivante suit la cohorte 2023 à travers passage, redoublement et abandon pour y répondre.

Troisième partie

La cohorte

CHAPITRE 5

Trois issues qui doivent sommer à un

Leçon 5 sur 8 · 105 min

70,4 % de passages, 9,8 % de redoublements, 10,2 % d'abandons, 5,7 % d'achèvements et 2,9 % de transferts. Chacune exige le même dénominateur, et les transferts sont la raison pour laquelle ce n'est pas l'effectif inscrit.

5.1 L'année, comme un ensemble de sorties

```python
import pandas as pd

enrolment = pd.read_csv("school-enrolment-2024.v1.csv")
clean = enrolment[enrolment["age_years"] <= 20]
cohort = clean[clean["school_year"] == 2023]

outcomes = cohort["end_of_year_status"].value_counts(normalize=True)
print((outcomes * 100).round(1))
print(f"\ncohort: {len(cohort)} students")
```

```r
library(dplyr)

enrolment |>
  filter(age_years <= 20, school_year == 2023) |>
  count(end_of_year_status) |>
  mutate(share = n / sum(n))
```

Issue	Élèves	Part
Passés	910	70,4 %
Abandonné	132	**10,2 %**
Redoublé	127	**9,8 %**
Achevé la dernière année	74	5,7 %
Transféré	37	2,9 %
Encore inscrits à la clôture	12	0,9 %

Les trois taux principaux sont le passage, le redoublement et l'abandon, et ils ne signifient quelque chose que s'ils partagent un dénominateur. Cela paraît évident et c'est l'étape la plus souvent sautée.

5.2 Le dénominateur n'est pas l'effectif inscrit

Deux catégories le compliquent, et dans des sens opposés.

Les transferts ne sont pas des abandons. Un enfant parti dans une autre école est toujours scolarisé. Les compter comme abandons surestime l'abandon de près de trois points et — pire — attribue une réussite du système à un échec d'école.

Les transferts ne sont pas non plus des réussites. Ils ne peuvent pas être comptés comme passages, car ce registre ne sait pas s'ils se sont réinscrits.

```python
def rates(frame, exclude_transfers):
    base = frame[frame["end_of_year_status"] != "transferred-out"] \
        if exclude_transfers else frame
    return {
        "n": len(base),
        "promotion": (base["end_of_year_status"]
                      .isin(["promoted", "completed-final-grade"]).mean()),
        "repetition": base["end_of_year_status"].eq("repeated").mean(),
        "dropout": base["end_of_year_status"].eq("dropped-out").mean(),
    }

for exclude in (False, True):
    result = rates(cohort, exclude)
    label = "excluding transfers" if exclude else "all students"
    print(f"{label:20} n={result['n']}  "
          f"promotion {result['promotion']:.1%}  "
          f"repetition {result['repetition']:.1%}  "
          f"dropout {result['dropout']:.1%}")
```

```r
cohort |>
  filter(end_of_year_status != "transferred-out") |>
  summarise(n = n(),
            promotion = mean(end_of_year_status %in% c("promoted", "completed-final-grade"
    )),
            repetition = mean(end_of_year_status == "repeated"),
            dropout = mean(end_of_year_status == "dropped-out"))
```

C'est le dénominateur du taux de guérison PCIMA rencontré plus tôt dans ce module, dans un autre secteur. La convention y était d'exclure les transferts parce que le programme ne peut pas revendiquer leur issue ; la convention est ici la même, pour la même raison. **Dites laquelle vous avez employée, et les trois taux sommeront à quelque chose de défendable.**

5.3 Le redoublement est une décision, non un événement

```python
by_district = clean.groupby("admin2")["repeating"].agg(["mean", "size"])
print((by_district * [100, 1]).round(1))
```

```r
enrolment |> filter(age_years <= 20) |>
  summarise(repeating = mean(repeating), n = n(), .by = admin2)
```

District	Taux de redoublement	n
Sud	12,3 %	583
Artibonite	12,1 %	618
Nord-Ouest	12,0 %	625
Centre	**7,4 %**	624

Centre rapporte un redoublement inférieur de près de cinq points à tous les autres districts. **C'est soit le district le plus performant de la région, soit une habitude d'enregistrement, et le registre seul ne peut pas trancher.**

C'est une habitude d'enregistrement : Centre signale une partie de ses vrais redoublants comme nouveaux entrants. La façon de le soupçonner sans qu'on vous le dise est que le

redoublement devrait suivre le retard d'âge, et à Centre il ne le suit pas — **8,7 % de redoublement contre 11,8 % ailleurs, alors que la part de retard d'âge de Centre vaut 40,6 % contre 35,7 %.** Le district qui rapporte le moins de redoublement a le *plus* d'enfants en retard d'âge, ce qui est à l'envers pour toute explication autre que le codage.

```
centre = clean[(clean["admin2"] == "Centre") & (clean["school_year"] == 2023)]
elsewhere = clean[(clean["admin2"] != "Centre") & (clean["school_year"] == 2023)]
for name, frame in [("Centre", centre), ("elsewhere", elsewhere)]:
    over = frame["age_years"] > frame["grade"] + 5
    print(f"{name}: repetition {frame['repeating'].mean():.1%}, "
          f"over-age {over.mean():.1%}")
```

```
# Repetition should predict over-age. Where it does not, suspect the coding.
```

Un taux en décalage avec l'indicateur qu'il cause mécaniquement est une question sur le bureau avant d'être un constat sur les écoles. C'est le même raisonnement que le fourre-tout des motifs de clôture du cours de protection, et il mérite d'être un réflexe.

5.4 L'abandon par année, et le pic de sixième

```
by_grade = cohort[cohort["grade"].between(1, 6)].groupby("grade").agg(
    n=("student_id", "size"),
    dropout=("end_of_year_status", lambda s: s.eq("dropped-out").mean()),
    repetition=("end_of_year_status", lambda s: s.eq("repeated").mean()),
)
print((by_grade * [1, 100, 100]).round(1))
```

```
cohort |> filter(between(grade, 1, 6)) |>
  summarise(n = n(),
            dropout = mean(end_of_year_status == "dropped-out"),
            repetition = mean(end_of_year_status == "repeated"), .by = grade)
```

Année	n	Abandon	Redoublement
1	149	8,1 %	2,0 %
2	271	8,5 %	7,0 %
3	332	13,0 %	6,6 %
4	231	8,7 %	10,4 %
5	165	8,5 %	19,4 %
6	110	**18,2 %**	12,7 %

Deux choses de ce tableau demandent une lecture attentive.

Le redoublement monte avec l'année et l'abandon non. La cinquième retient 19,4 % de ses élèves alors que son abandon vaut 8,5 % ; la sixième fait l'inverse. Les deux sont des sorties alternatives d'un même point de décision, et lire l'une seule donne une forme fausse du problème.

L'abandon en sixième vaut 18,2 % sur 110 élèves. C'est le taux le plus élevé du tableau et le plus petit dénominateur, donc avant d'en faire la priorité, posez-lui un intervalle — la machinerie du cours sur les enquêtes, sur une proportion qui bougerait de plusieurs points si quatre enfants avaient fait autre chose.

5.5 Rapportez les sorties, puis les taux

```
End of school year 2023, 1,292 students

  Promoted                     70.4%   910
  Dropped out                  10.2%   132
  Repeated                      9.8%   127
  Completed the final grade     5.7%    74
  Transferred out               2.9%    37   excluded from the rates below
  Still enrolled at cut-off     0.9%    12

  Rates on 1,255 students, transfers excluded:
    Promotion (incl. completion)  78.4%
    Repetition                    10.1%
    Dropout                       10.5%

  Centre reports repetition at 7.4% against 12.0-12.3% elsewhere while
  holding the highest over-age share of the four districts. Treat as a
  recording difference pending verification, not as performance.
```

5.6 La suite

Ce sont les sorties d'une année. Les transformer en « combien d'enfants achèvent le primaire » exige une cohorte, et la leçon suivante en construit une — avec la raison pour laquelle le nombre qu'elle produit est presque certainement faux.

CHAPITRE 6

Un taux de survie qu'il ne faut pas publier

Leçon 6 sur 8 · 105 min

Chaînez les taux de passage d'une année et 19,3 % des enfants atteignent la dernière année. Comptez les redoublants comme retenus et cela fait 40,3 %. Les deux viennent du même fichier, et aucun n'est un taux d'achèvement.

6.1 Le calcul vers lequel tout le monde se tourne

Vous avez les taux de passage par année pour une année scolaire. Chaînez-les et vous obtenez la survie jusqu'à la dernière année — la mesure d'achèvement de type ODD — sans attendre six ans.

```python
import pandas as pd

enrolment = pd.read_csv("school-enrolment-2024.v1.csv")
clean = enrolment[enrolment["age_years"] <= 20]
cohort = clean[clean["school_year"] == 2023]

survival, cumulative = {}, 1.0
for grade in range(1, 7):
    grade_rows = cohort[cohort["grade"] == grade]
    promoted = grade_rows["end_of_year_status"].isin(
        ["promoted", "completed-final-grade"]).mean()
    survival[grade] = cumulative
    cumulative *= promoted
    print(f"grade {grade}: promotion {promoted:.1%}, "
          f"survival to here {survival[grade]:.1%}")

print(f"\nsurvival to grade 6 entry: {cumulative:.1%}")
```

```r
library(dplyr)

cohort |>
  filter(between(grade, 1, 6)) |>
  summarise(promotion = mean(end_of_year_status %in%
              c("promoted", "completed-final-grade")), .by = grade) |>
  arrange(grade) |>
  mutate(survival = cumprod(lag(promotion, default = 1)))
```

Année	Passage	Survie cumulée
1	89,3 %	100,0 %
2	83,0 %	89,3 %
3	71,7 %	74,1 %
4	76,2 %	53,1 %
5	70,9 %	40,5 %
6	67,3 %	28,7 %
—	—	**19,3 %**

19,3 % des enfants entrant en première année atteignent la fin du primaire. C'est un nombre frappant, c'est ce que la méthode produit, et il ne doit pas quitter votre écran.

6.2 Pourquoi il est faux

C'est une **cohorte reconstruite** : elle suppose que le taux de passage de cinquième année observé cette année est celui que les enfants aujourd'hui en première année connaîtront dans quatre ans. Quatre choses cassent cette hypothèse, et ici elles la cassent toutes dans le même sens.

Les redoublants sont comptés comme des échecs. Un enfant qui redouble la deuxième année n'est pas déscolarisé — il est de nouveau en deuxième année. La chaîne traite le non-passage comme une sortie, si bien que 10,1 % de redoublement se composent six fois en une perte apparente énorme.

```
promotion_or_repeat = cohort[cohort["grade"].between(1, 6)].groupby("grade")[
    "end_of_year_status"].apply(
    lambda s: s.isin(["promoted", "completed-final-grade", "repeated"]).mean())

still_enrolled = 1.0
for grade, rate in promotion_or_repeat.items():
    still_enrolled *= rate
print(f"survival counting repeaters as retained: {still_enrolled:.1%}")
```

```
# Retention, not promotion, is what a survival rate needs.
```

Les transferts sont eux aussi comptés comme des échecs. 2,9 % sont partis dans une autre école et la chaîne les lit comme quittant l'éducation.

Une seule année est prise pour six. Le taux de passage de sixième observé en 2023 appartient à des enfants entrés en 2018, dans d'autres conditions.

Les effectifs par année ne forment pas une cohorte. La première année compte 149 élèves et la sixième 110, et une partie de cet écart est une attrition réelle tandis qu'une autre est une population qui change — une cohorte de naissances plus large arrivant en première année fait ressembler la pyramide à de l'abandon.

6.3 Que calculer à la place

La rétention, non le passage, s'il faut employer une seule année. Comptez comme retenus les enfants encore scolarisés — passés ou redoublants.

```
def chained(statuses):
    rate, product = {}, 1.0
    for grade in range(1, 7):
        rows = cohort[cohort["grade"] == grade]["end_of_year_status"]
        product *= rows.isin(statuses).mean()
        rate[grade] = product
    return product

promotion_only = chained(["promoted", "completed-final-grade"])
retained = chained(["promoted", "completed-final-grade", "repeated"])
print(f"chained promotion: {promotion_only:.1%}")
print(f"chained retention: {retained:.1%}")
```

```
# Same chain, different numerator. The gap is repetition compounding.
```

19,3 % contre 40,3 %. Plus de vingt points de la perte apparente étaient du redoublement lu comme une sortie, et le chiffre de rétention est le plus proche de quelque chose de signifiant — sans être pour autant un taux d'achèvement.

Suivez de vrais élèves quand vous le pouvez. Ce registre a deux années et les mêmes identifiants, si bien qu'une transition d'un an est directement observable plutôt que supposée.

```
years = clean.pivot_table(index="student_id", columns="school_year",
                          values="grade", aggfunc="first")
both = years.dropna()
print(f"students observed in both years: {len(both)}")
print(f"  advanced a grade: {(both[2024] > both[2023]).mean():.1%}")
print(f"  same grade:       {(both[2024] == both[2023]).mean():.1%}")
```

1 103 élèves figurent dans les deux années : 88,5 % ont avancé d'une année et 11,5 % sont deux fois dans la même. C'est une transition mesurée et non supposée, et c'est le nombre à mettre dans un rapport.

```
enrolment |>
  select(student_id, school_year, grade) |>
  tidyr::pivot_wider(names_from = school_year, values_from = grade) |>
  filter(!is.na(`2023`), !is.na(`2024`))
```

Une transition d'un an observée sur de vrais élèves vaut mieux qu'une chaîne de six ans assemblée à partir des taux d'une seule année, même si elle répond à une question plus petite. La question plus petite est défendable.

6.4 Qui part

La désagrégation qui, elle, donne quelque chose, contrairement aux ventilations du retard d'âge de la leçon 2.

```
year23 = cohort.assign(
    over_age=cohort["age_years"] > cohort["grade"] + 5,
    dropped=cohort["end_of_year_status"].eq("dropped-out"),
)
for column in ("over_age", "disability_reported", "displacement_status", "sex"):
    table = year23.groupby(column, dropna=False)["dropped"].agg(["mean", "size"])
    print((table * [100, 1]).round(1), "\n")
```

```
cohort |>
  mutate(dropped = end_of_year_status == "dropped-out") |>
  summarise(dropout = mean(dropped), n = n(), .by = displacement_status)
```

Ventilation	Abandon	n
En retard d'âge	**19,5 %**	478
À l'âge attendu	4,8 %	814
Retournés	19,7 %	71
Handicap signalé	17,9 %	112
Déplacés internes	14,7 %	163
Communauté d'accueil	11,6 %	95
Résidents	8,6 %	940
Garçons	10,4 %	634
Filles	10,0 %	658

Le retard d'âge est le prédicteur le plus fort et le groupe le plus grand. Le handicap et le déplacement doublent à peu près le taux sur des dénominateurs plus petits.

Le sexe ne montre rien — 10,4 % contre 10,0 % sur 634 et 658 élèves. C'est un résultat qui mérite d'être énoncé, car un écart de genre dans l'abandon est le constat qu'un programme éducatif s'attend le plus à voir et ce registre n'en contient pas.

6.5 Les données manquantes ne le sont pas au hasard

```python
print(year23["displacement_status"].isna().sum(), "students with no status")
print(year23.loc[year23["displacement_status"].isna(), "dropped"].mean())
```

```r
cohort |> filter(is.na(displacement_status)) |> nrow()
```

Le statut de déplacement est vide pour 42 élèves du registre, dont 23 dans la cohorte 2023. Les vides proviennent entièrement de ménages déplacés et retournés — c'est le champ qu'un enquêteur saute quand une famille vient d'arriver.

Soyez précis sur ce que cela biaise et ne biaise pas. Les vides sont retirés à peu près au hasard *à l'intérieur* de ces deux catégories, donc :

— le **taux d'abandon** des déplacés internes et des retournés n'est pas biaisé — 14,7 % et 19,7 % sont les bons nombres pour ces groupes ;

— leurs **effectifs et leur part de population** sont sous-estimés d'environ un dixième. 234 élèves sont enregistrés comme déplacés ou retournés en 2023 et le vrai chiffre approche 260.

Donc « les enfants déplacés abandonnent à 14,7 % » est défendable et « les enfants déplacés représentent 18 % des effectifs » ne l'est pas. **Un motif de valeurs manquantes peut biaiser un effectif sans biaiser un taux**, et savoir sur lequel repose votre phrase décide si le motif compte.

L'abandon observé parmi les vides eux-mêmes vaut 8,7 % sur 23 élèves, ce qui est trop peu pour en lire quoi que ce soit et ne doit pas être rapporté comme une catégorie.

6.6 Rapportez-le comme un profil de risque

```
Dropout, 2023 cohort, 1,292 students

  Overall                      10.2%

  Over-age for grade           19.5%   478 students
  In-age                        4.8%   814
```

```
Returnee                    19.7%    71 students
Disability reported         17.9%   112
Internally displaced        14.7%   163
Resident                     8.6%   940

No meaningful difference by sex: 10.4% boys, 10.0% girls.

42 students have no displacement status, all drawn from displaced and
returnee households. The rates above are unbiased for those groups; their
counts are understated by about a tenth.

Not reported: survival to the final grade. A reconstructed cohort from a
single year gives 19.3% on promotion and 40.3% on retention, neither of
which is a completion rate. A true cohort needs six years of the register
or a household survey. The measured one-year transition is 88.5%.
```

« Non rapporté, et pourquoi » est la ligne la plus utile de ce bloc. Quelqu'un demandera un taux d'achèvement, et la réponse en est une vraie : pas à partir de ceci, à partir d'une enquête ménage, et voici ce que ceci peut dire à la place.

6.7 La suite

L'inscription, la présence et la rétention disent si les enfants sont à l'école. Aucune ne dit s'ils apprennent, et la dernière unité porte sur l'instrument qui le dit — où deux passations se révèlent non comparables.

Quatrième partie

Les apprentissages

CHAPITRE 7

Huit points de progrès, dont deux que personne n'a gagnés

Leçon 7 sur 8 · 105 min

La lecture passe de 53,1 % à 61,8 % des items entre les deux passations. Sur les 585 élèves ayant passé les deux, elle passe de 56,8 % à 63,7 %. L'écart tient à qui s'est présenté, et les scores bruts n'auraient pu donner ni l'un ni l'autre.

7.1 La comparaison qui ne peut pas être faite

```python
import pandas as pd

assessment = pd.read_csv("learning-assessment-2024.v1.csv")
literacy = assessment[assessment["domain"] == "literacy"]

print(literacy.groupby("assessment_round").agg(
    students=("student_id", "nunique"),
    items=("items", "first"),
    mean_raw=("raw_score", "mean"),
).round(1))
```

```r
library(dplyr)

assessment |>
  filter(domain == "literacy") |>
  summarise(students = n_distinct(student_id), items = first(items),
            mean_raw = mean(raw_score), .by = assessment_round)
```

Passation	Élèves	Items	Score brut moyen
Initiale	741	**40**	21,2
Finale	658	**50**	30,9

Un score brut est passé de 21,2 à 30,9 et cette comparaison ne signifie rien, car les instruments n'ont pas la même longueur. Neuf points de la hausse sont dix questions de plus.

C'est l'erreur la plus courante du rapportage des évaluations éducatives, et elle est invisible tant qu'on ne regarde pas la colonne `items` — que la plupart des extraits ne portent pas et que la plupart des analystes ne demandent pas.

7.2 Trois choses comparables, et une qui ne l'est pas

Non comparable : le score brut. Dénominateurs différents.

Comparable avec prudence : la proportion d'items réussis. Elle met les deux passations sur une échelle de 0 à 1, ce qui est nécessaire et non suffisant — les items de la passation finale peuvent être plus ou moins difficiles, et une proportion suppose qu'ils sont équivalents.

```
literacy = literacy.assign(proportion=literacy["raw_score"] / literacy["items"])
print(literacy.groupby("assessment_round")["proportion"].mean().round(3))
```

```
assessment |> filter(domain == "literacy") |>
  summarise(proportion = mean(raw_score / items), .by = assessment_round)
```

53,1 % au départ et 61,8 % à la fin. Mieux, et reposant encore sur une hypothèse que personne n'a vérifiée.

Comparable par construction : les bandes de compétence. Elles ont été équatées lors de la conception des instruments, ce qui est l'objet même de l'équatage — une procédure psychométrique qui place deux tests différents sur une seule échelle, et c'est la raison pour laquelle les bandes existent comme colonne distincte plutôt que dérivées du score.

```
bands = pd.crosstab(assessment["assessment_round"],
                    assessment["proficiency_band"], normalize="index")
order = ["below-minimum", "minimum", "proficient", "advanced"]
print((bands[order] * 100).round(1))
```

```
assessment |> count(assessment_round, proficiency_band) |>
  mutate(share = n / sum(n), .by = assessment_round)
```

Bande	Initiale	Finale
Sous le minimum	19,3 %	**11,1 %**
Minimum	30,6 %	23,9 %
Compétent	38,2 %	38,9 %
Avancé	11,9 %	**26,1 %**

La bande sous le minimum passe de 19,3 % à 11,1 % et la bande avancée de 11,9 % à 26,1 %. C'est la comparaison à rapporter, car c'est la seule dont l'échelle a été conçue pour survivre à un changement d'instrument.

7.3 Qui a passé l'épreuve

```
sat = assessment.groupby("assessment_round")["student_id"].apply(set)
baseline, endline = sat["baseline"], sat["endline"]

print(f"baseline only: {len(baseline - endline)}")
print(f"endline only:  {len(endline - baseline)}")
print(f"both rounds:   {len(baseline & endline)}")
```

```
# Three groups, and only one of them supports a change estimate.
```

741 ont passé l'initiale, 658 la finale, et 585 les deux. 156 élèves ayant passé le premier test n'ont pas passé le second.

Ce ne sont pas 156 élèves au hasard. Le jour de l'évaluation attrape qui se trouve en classe, et les enfants qui n'y sont pas sont ceux qui fréquentent le moins — qui sont aussi, d'après la leçon 3, ceux qui obtiennent les scores les plus bas.

La moyenne finale est donc calculée sur un groupe dont les plus faibles ont été retirés de façon disproportionnée, et elle montera pour cette seule raison.

7.4 Mesurez la sélection au lieu de l'écarter par hypothèse

Le registre permet de faire ce qui tranche : calculer l'évolution sur les élèves présents aux deux passations.

```python
panel = literacy.pivot_table(index="student_id", columns="assessment_round",
                             values="proportion")
panel = panel.dropna()

print(f"panel of {len(panel)} students")
print(f"  baseline {panel['baseline'].mean():.1%} → endline {panel['endline'].mean():.1%}
    ")

all_comers = literacy.groupby("assessment_round")["proportion"].mean()
print(f"all comers")
print(f"  baseline {all_comers['baseline']:.1%} → endline {all_comers['endline']:.1%}")
```

```r
assessment |> filter(domain == "literacy") |>
  mutate(proportion = raw_score / items) |>
  select(student_id, assessment_round, proportion) |>
  tidyr::pivot_wider(names_from = assessment_round, values_from = proportion) |>
  filter(!is.na(baseline), !is.na(endline)) |>
  summarise(n = n(), baseline = mean(baseline), endline = mean(endline))
```

	Initiale	Finale	Évolution
Tous présents	53,1 %	61,8 %	**+8,7 points**
Panel de 585	56,8 %	63,7 %	**+6,9 points**

Deux des 8,7 points tiennent à qui s'est présenté. La valeur initiale du panel dépasse de 3,7 points celle de tous les présents, ce qui est la mesure directe de la sélection : les élèves revenus étaient déjà plus forts.

Rapportez l'estimation du panel et montrez le chiffre tous-présents à côté. Le panel est le nombre défendable ; l'écart entre les deux est la preuve que la sélection était réelle et son ampleur.

7.5 Ce que coûte le panel

Il n'est pas gratuit, et l'arbitrage doit être énoncé.

Le panel n'est pas la population. 585 élèves sur 741 à l'initiale, et ce sont les plus assidus. Les +6,9 points sont donc une estimation non biaisée de l'évolution *pour les enfants restés scolarisés et présents deux fois*, ce qui est une affirmation plus étroite que celle que l'on attend.

Il ne dit rien des 156 partis. Leurs apprentissages ont pu aller n'importe où, et le rapport honnête dit que l'estimation ne les couvre pas.

```python
left = literacy[literacy["student_id"].isin(baseline - endline)
                & (literacy["assessment_round"] == "baseline")]
print(f"the 156 who did not return scored "
      f"{left['proportion'].mean():.1%} at baseline")
print(f"the 585 who did scored "
      f"{panel['baseline'].mean():.1%}")
```

```
# Quantify the difference rather than asserting it.
```

Les 156 qui ne sont pas revenus obtenaient 39,3 % à l'initiale contre 56,8 % pour le panel — dix-sept points en dessous. Ce n'est pas un effet de sélection subtil ; c'est le sixième le plus faible de la cohorte qui sort de la moyenne finale.

Chiffrer la sélection est le livrable, non l'éliminer. Une analyse qui dit « l'échantillon final était plus assidu, et voici de combien » a fait son travail ; une analyse qui rapporte +8,7 points ne l'a pas fait.

7.6 Rapportez les trois nombres

```
Literacy, baseline to endline

  Proficiency bands, all who sat each round
    Below minimum        19.3% → 11.1%
    Advanced             11.9% → 26.1%

  Proportion of items correct
    All comers           53.1% → 61.8%   +8.7 points   n=741, n=658
    Panel of both rounds 56.8% → 63.7%   +6.9 points   n=585

  Instruments differed: 40 items at baseline, 50 at endline. Raw scores are
  not comparable and are not reported. Proficiency bands were equated at
  design; the proportion of items assumes equivalent difficulty.

  156 baseline students did not sit the endline. They scored 39.3% at
  baseline against 56.8% for those who returned, so the all-comers change
  overstates learning by about 1.8 points. The panel estimate covers
  children who sat both rounds and does not cover those 156.
```

7.7 La suite

Vous disposez désormais de tous les indicateurs éducatifs que ce cours peut produire — inscription, présence, rétention et apprentissages — chacun avec son dénominateur et sa réserve. La dernière leçon porte sur le rapport où ils vont, et sur les quatre nombres sur lesquels un directeur d'école peut réellement agir.

CHAPITRE 8

Quatre nombres sur lesquels un directeur d'école peut agir

Leçon 8 sur 8 · 105 min

Ce cours a produit quatorze indicateurs sur six dénominateurs. Un rapport de district les exige tous ; un directeur d'école en a besoin de quatre, et choisir lesquels est le travail analytique et non la mise en forme.

8.1 Comptez d'abord les dénominateurs

L'habitude que ce module construit depuis le début, une dernière fois.

Indicateur	Dénominateur	n
Taux brut de scolarisation	Enfants de 6 à 11 ans (projetés)	963
Taux net de scolarisation	Enfants de 6 à 11 ans (projetés)	963
Part en retard d'âge	Inscrits du primaire	1 052
Assiduité moyenne quotidienne	**Pointages faits**	69 974
Élèves présents 90 %+	**Élèves avec 20 pointages ou plus**	1 200
Passage, redoublement, abandon	Élèves 2023 hors transferts	1 255
Transition d'un an	Élèves présents les deux années	1 103
Apprentissages, tous présents	**Élèves ayant passé chaque épreuve**	741 / 658
Apprentissages, panel	**Élèves ayant passé les deux**	585

Six dénominateurs différents et aucun n'est « les élèves ». Deux sont des populations projetées, deux sont des pointages, trois sont des effectifs d'élèves sous des règles différentes, et un est un sous-échantillon d'un sous-échantillon.

```
import pandas as pd

denominators = pd.DataFrame([
    ("gross enrolment", "children aged 6-11, projected", 963),
    ("attendance, ADA", "marks made", 69974),
    ("attendance, regular", "students with 20+ marks", 1200),
    ("dropout", "2023 students, transfers excluded", 1255),
    ("learning change", "students who sat both rounds", 585),
], columns=["indicator", "denominator", "n"])
print(denominators)
```

```
# Build it in code, print it above the results, every time.
```

8.2 Le rapport de district

```
Education, four districts, school year 2023-24

ENROLMENT                                    denominator: 963 projected 6-11
  Gross enrolment ratio           109.2%     1,052 primary enrolees
```

```
  Net enrolment ratio              97.4%      938 aged 6-11
  Over-age for grade               36.4%    grade 1 17.0%, grade 6 50.8%

  Denominator is a 2015 census projected at 2.1% a year; about a sixth of it
  is an assumption. 12 student-years were recorded twice and deduplicated.

ATTENDANCE                                  February to April 2024
  Average daily attendance         88.4%    69,974 marks
  Students attending 90%+ of days  62.1%    1,200 students with 20+ marks
  Students below 75%                9.5%    114 students

  SCH07 and SCH18 were closed 11-29 March, fifteen school days. Their rates
  are computed on the 45 days they opened. Counting the closure as absence
  would report them at 66.1% and 63.9% and rank them worst.

RETENTION                                   2023 cohort, transfers excluded
  Promotion (incl. completion)     78.4%    n=1,255
  Repetition                       10.1%
  Dropout                          10.5%
  One-year transition observed     88.5%    n=1,103 in both years

  Dropout is 19.5% among over-age students against 4.8% in-age.
  Centre reports repetition at 7.4% against 12.0-12.3% elsewhere while
  holding the highest over-age share; treat as a recording difference.

LEARNING                                    literacy
  Below minimum band        19.3% -> 11.1%
  Advanced band             11.9% -> 26.1%
  Items correct, panel      56.8% -> 63.7%   +6.9 points, n=585
  Items correct, all comers 53.1% -> 61.8%   +8.7 points

  Instruments were 40 and 50 items; raw scores are not comparable and are not
  reported. 156 baseline students did not sit the endline and scored 39.3%
  against 56.8% for those who did.

NOT REPORTED
  Survival to the final grade. A reconstructed cohort gives 19.3% on
  promotion and 40.3% on retention; neither is a completion rate. A true
  cohort needs six years of the register or a household survey.
  Safely managed anything. This is an enrolment register, not a survey of
  children out of school -- every denominator above counts children the
  system already has.
```

8.3 Les quatre nombres pour un directeur d'école

Un rapport de district n'est pas un rapport d'école, et remettre à un directeur le tableau ci-dessus est une façon de garantir qu'il ne se passera rien.

Quatre nombres, chacun rattaché à quelque chose qu'une école peut changer.

Nombre	Pourquoi celui-là	L'action
Élèves sous 75 % de présence	Une liste, non un taux	Visite de suivi ce trimestr
Part en retard d'âge en première année	L'entrée tardive se corrige à l'inscription	Campagne d'inscription
Taux de redoublement de cette école	Une décision que l'école prend	Revoir les critères
Part sous la bande minimum	Les enfants qui n'apprennent pas à lire	Soutien ciblé

```
follow_up = (attendance_rates < 0.75).sum()
print(f"students to visit this term: {follow_up}")
```

```
# One number, one list, one action.
```

Remarquez ce qui ne figure pas sur cette liste. Le taux brut de scolarisation est un nombre de planification de district et un directeur ne peut pas le déplacer. L'assiduité moyenne quotidienne est un indicateur de système et elle masque précisément les enfants que l'école devrait visiter. Les deux ont leur place dans le rapport de district et aucun n'a sa place au mur d'une école.

Faites correspondre l'indicateur au périmètre réel de décision de son destinataire, ce qui est la règle qu'appliquait le cours EAH à une équipe mobile et le cours de protection à une charge de dossiers — et la raison pour laquelle un tableau de bord unique pour tous les publics n'en sert aucun.

8.4 La distinction pour laquelle tout ce cours existait

Inscrit, présent et apprenant sont trois populations différentes, et ce cours a produit un nombre pour chacune sur les mêmes enfants.

```
funnel = pd.DataFrame([
    ("enrolled in primary", 1052),
    ("attending 90%+ of days", int(1200 * 0.621)),
    ("proficient or advanced in literacy", int(658 * 0.65)),
], columns=["stage", "students"])
print(funnel)
```

```
# Three stages, three denominators. Do not draw them as one funnel without
# saying so.
```

Ces trois-là ne peuvent pas être enchaînés en entonnoir, car ils sont mesurés sur des dénominateurs différents et sur des périodes différentes — et les dessiner comme un seul est l'erreur la plus courante des tableaux de bord éducatifs après le taux de scolarisation.

Ce qu'ils permettent en revanche, c'est la phrase que ce cours existe pour rendre disponible : **un système peut scolariser presque chaque enfant d'âge scolaire, avoir la plupart d'entre eux présents la plupart des jours, et laisser malgré tout un enfant sur neuf incapable de lire au niveau minimum** — et chacun de ces trois faits exige son propre instrument, son propre dénominateur et sa propre réserve.

8.5 Le module 4, en un paragraphe

Six secteurs, six vocabulaires, un seul jeu de questions.

Qu'ai-je compté, sur quoi ? Des personnes-temps en épidémiologie, des visites en EAH, des cas consentants en protection, des pointages en éducation. Aucun de ces dénominateurs n'était « la population » et aucun n'allait de soi.

Qu'est-ce qui pourrait aussi l'expliquer ? Une structure par âge dans une épidémie de choléra, une route coupée dans un district d'eau, l'ouverture d'un service dans une charge de dossiers de protection, une grève dans un registre scolaire. Dans chaque cas la chose absente n'était pas une valeur et rien dans le fichier ne l'annonçait.

Que devrait-on faire différemment ? La question qui décide si les deux premières

valaient la peine d'être posées, et celle qui décide quels quatre nombres parviennent à la personne capable d'agir.

8.6 La suite

Le module 4 est achevé. Le module 5 passe de la description des données de programme à leur modélisation — régression, prévision et les tableaux de bord qui en portent le résultat — et chacun de ses cours s'appuie sur les fichiers que ce module a construits.

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/education-programme-analysis

Généré le 28 juillet 2026.