CASSION · FORMATION

Santé publique et épidémiologie appliquée

Incidence, prévalence, cascades et couverture pour les programmes VIH, tuberculose, paludisme et vaccination, ainsi que la courbe épidémique qu'il faudra peut-être tracer dans l'urgence.

Formateur	Pierre Robentz Cassion
Niveau	Intermédiaire
Heures apprenant	22 h
Enseignée en	Python · R
Mise à jour	28 juillet 2026
Secteurs	Santé publique
Normes et méthodologies	Définitions d'indicateurs de l'UNICEF · Objectifs de développement durable (ODD) · Standards Sphère

Consulter et exécuter la formation en ligne sur data-analysis.cassion.dev/fr/cours/public-health-epidemiology

Ce que vous saurez faire

— Distinguer un taux, un rapport et une proportion, et dire ce que les personnes-temps apportent qu'un dénominateur en effectifs ne peut pas

— Construire une cascade comme une suite de dénominateurs enchaînés et localiser l'étape où la plus grande part se perd

— Tracer une courbe épidémique à partir d'une liste linéaire, et dire ce que les dates de début manquantes font à son pic

— Calculer des taux d'attaque et une létalité avec leurs dénominateurs énoncés, au regard des seuils sur lesquels la réponse est jugée

— Standardiser deux districts sur l'âge et dire quelle part de leur différence apparente était de la structure

— Nommer les facteurs de confusion d'une comparaison sur données de routine, et énoncer franchement ce qu'un constat observationnel peut et ne peut pas affirmer

Prérequis

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

Sommaire

À propos de cette formation

L'épidémiologie est la discipline des dénominateurs, et ce cours est le point où la consigne permanente de la plateforme — toujours énoncer comment un indicateur est calculé — rencontre des mesures qui ont des noms et des seuils publiés.

Il s'exécute sur deux registres. La **liste linéaire du choléra** compte 975 cas enregistrés un par un sur seize semaines, livrée avec la population dont ils proviennent, et c'est le seul jeu de données d'ici qui soutienne à la fois une courbe épidémique, un taux d'attaque et une létalité. L'**extraction vaccinale** a servi dans six cours antérieurs et celui-ci l'achève — la couverture de trois façons, et l'abandon entre doses qui dit si les enfants qui commencent un calendrier le terminent.

Trois résultats donnent sa forme au cours, tous calculés sur ces fichiers.

La structure par âge fait la moitié de l'écart entre le pire et le meilleur district. Les taux d'attaque bruts du choléra valent 7,94 et 5,51 pour 1 000. Standardisés sur une population commune ils valent 7,25 et 5,95, si bien que l'écart passe de 2,43 à 1,30. La moitié de la différence apparente tenait à qui habite là, non à ce qui leur est arrivé.

La létalité vaut 3,90 % contre une cible de 1 %, et 6,30 % dans un district contre 2,00 % dans un autre. Le prédicteur dans ces données est le délai entre début des symptômes et admission, ce qui est une question de service et non de clinique.

La courbe épidémique n'est pas l'épidémie. Quarante-sept cas n'ont pas de date de début, et le registre d'un district est rempli après coup depuis le cahier d'admission, si bien que sa courbe est en réalité une courbe d'admissions décalée vers la droite.

Python et R tout du long. Vous devriez avoir suivi le module 3 — ce cours suppose que vous savez défendre un dénominateur, mettre un intervalle sur une proportion, et lire un chiffre de routine à travers son taux de rapportage.

Première partie

Le dénominateur, c'est l'épidémiologie

CHAPITRE 1

Taux, rapport, proportion — et personnes-temps

Leçon 1 sur 8 · 135 min

Trois formes que l'on appelle toutes « taux » et une seule l'est, plus le dénominateur qui donne son sens à l'incidence quand les gens entrent et sortent d'une population à des moments différents.

1.1 Trois formes, un seul mot

Le cours sur la conception d'indicateurs a établi quatre familles de mesure. L'épidémiologie en emploie trois constamment et les appelle toutes « taux », et c'est de là que part l'essentiel des contresens.

Forme	Structure	Plage	Exemple
Proportion	Partie sur tout, mêmes unités	0 à 1	Létalité — décès parmi les cas
Rapport	Deux quantités non emboîtées	quelconque	Sex-ratio des cas ; rapport de cotes
Taux	Événements sur population-**temps**	positif	Incidence pour 1 000 personnes-années

Seule la troisième est un taux, et la différence est le temps au dénominateur. Une proportion demande « quelle part » ; un taux demande « à quelle vitesse ».

1.2 Prévalence et incidence

Le couple autour duquel tourne tout ce cours.

— **La prévalence** — combien de personnes ont l'affection *maintenant*. Une proportion, issue d'une enquête, un instantané.

— **L'incidence** — combien de *nouveaux* cas surviennent par unité de population-temps. Un taux, issu de la surveillance, un flux.

Ils répondent à des questions différentes et bougent pour des raisons différentes. La prévalence monte si l'incidence monte **ou si les gens survivent plus longtemps avec l'affection**, ce qui explique qu'un programme de traitement réussi puisse faire monter la prévalence du VIH tout en faisant baisser son incidence — le couple de nombres le plus mal lu de ce secteur.

```
import pandas as pd

cases = pd.read_csv("cholera-line-list-2024.v1.csv", parse_dates=["onset_date"])
population = pd.read_csv("district-population-2024.v1.csv")

total_population = population["population"].sum()
print(f"{len(cases)} cases in a population of {total_population:,}")
print(f"attack rate over the outbreak: {1000 * len(cases) / total_population:.2f} per
    1,000")
```

```r
library(dplyr)

cases <- readr::read_csv("cholera-line-list-2024.v1.csv")
population <- readr::read_csv("district-population-2024.v1.csv")

c(cases = nrow(cases),
  population = sum(population$population),
  per_1000 = 1000 * nrow(cases) / sum(population$population))
```

975 cas dans 145 000 personnes — **6,7 pour 1 000 sur l'épidémie**. À strictement parler c'est un *taux d'attaque*, qui est une proportion malgré son nom : les cas sur la population à risque, sur une période d'épidémie définie. La leçon 4 le prend au sérieux.

1.3 Ce qu'apportent les personnes-temps

Un dénominateur en effectifs suppose que tout le monde était présent et à risque pendant toute la période. Dans ce secteur, ce n'est régulièrement pas le cas — les gens arrivent, partent, naissent, meurent, ou entrent dans un programme en cours d'année.

Les personnes-temps comptent chaque personne aussi longtemps qu'elle était réellement à risque.

```python
# One person followed for 6 months contributes 0.5 person-years
follow_up = pd.DataFrame({
    "person_id": ["A", "B", "C"],
    "entered": pd.to_datetime(["2024-01-01", "2024-01-01", "2024-07-01"]),
    "exited": pd.to_datetime(["2024-12-31", "2024-06-30", "2024-12-31"]),
})
follow_up["person_years"] = (
    (follow_up["exited"] - follow_up["entered"]).dt.days / 365.25
)

events = 1
print(f"{events / follow_up['person_years'].sum():.3f} events per person-year")
print(f"naive headcount rate: {events / len(follow_up):.3f} per person")
```

```r
follow_up <- tibble::tibble(
  person_id = c("A", "B", "C"),
  entered = as.Date(c("2024-01-01", "2024-01-01", "2024-07-01")),
  exited  = as.Date(c("2024-12-31", "2024-06-30", "2024-12-31"))
) |>
  mutate(person_years = as.numeric(exited - entered) / 365.25)

c(per_person_year = 1 / sum(follow_up$person_years),
  naive = 1 / nrow(follow_up))
```

Trois personnes, dont deux présentes la moitié de l'année. Le dénominateur en effectifs vaut 3 ; le dénominateur en personnes-temps vaut 2,0 personnes-années. **Les deux réponses diffèrent d'un tiers**, et la seconde est celle qui se compare entre périodes et lieux à durées de suivi différentes.

Employez les personnes-temps dès que :

— **le suivi varie** — une cohorte de traitement où l'on s'inscrit tout au long de l'année ;

— **la population change** — déplacements, arrivées en camp, bassin qui grandit ;

— **vous comparez des périodes de longueurs différentes** — une réponse de neuf mois face à une de douze.

Employez un dénominateur en effectifs quand tout le monde était présent tout du long et que la période est fixée, ce qui rend légitime un taux d'attaque d'épidémie.

1.4 Mettez le dénominateur et le multiplicateur dans le nom

```
indicators = {
    "cholera_attack_rate_per_1000_outbreak": 1000 * len(cases) / total_population,
    "case_fatality_percent_of_cases_with_outcome": None,
    "incidence_per_1000_person_years": None,
}
```

```
# The same discipline: the name carries the denominator and the multiplier
```

Deux choses ont leur place dans le nom et en sont d'ordinaire absentes.

Le multiplicateur. Pour 100, pour 1 000, pour 100 000. Les taux d'attaque du choléra se donnent conventionnellement pour 1 000 ; la mortalité maternelle pour 100 000 ; la couverture vaccinale pour 100. Se tromper de convention représente un facteur cent, et cela arrive.

La population du dénominateur. « Pour 1 000 habitants » et « pour 1 000 enfants de moins de cinq ans » sont deux indicateurs différents. Le cours sur les indicateurs le disait en général ; ici les conventions sont publiées et s'en écarter en silence est pire que d'en inventer une.

1.5 La population à risque n'est pas toujours toute la population

Le dénominateur d'un taux est la population **à risque de l'événement**.

— **Le ratio de mortalité maternelle** — décès pour 100 000 *naissances vivantes*, non pour la population, car seules les grossesses sont à risque.

— **La mortalité néonatale** — décès dans les 28 premiers jours pour 1 000 *naissances vivantes*.

— **Le taux d'attaque du choléra** — cas pour 1 000 habitants, car tout le monde est à risque.

— **Le taux d'attaque de la rougeole** — sans doute pour 1 000 personnes *réceptives*, et la population réceptive est précisément ce qu'un programme de vaccination modifie.

```
under_five = population.loc[population["age_band"] == "0-4", "population"].sum()
u5_cases = (cases["age_band"] == "0-4").sum()

print(f"under-five attack rate: {1000 * u5_cases / under_five:.2f} per 1,000")
print(f"all-age attack rate:    {1000 * len(cases) / total_population:.2f} per 1,000")
```

```
under5 <- population |> filter(age_band == "0-4") |> summarise(sum(population)) |> pull()
u5_cases <- sum(cases$age_band == "0-4")

c(under_five = 1000 * u5_cases / under5,
  all_age = 1000 * nrow(cases) / sum(population$population))
```

Les moins de cinq ans ont ici environ le double du taux d'attaque tous âges. **Cet écart est toute la raison d'être de la leçon 6** — deux districts aux structures par âge différentes différeront sur le taux tous âges même si chaque taux par âge est identique.

1.6 Le contrôle à faire sur tout taux

Quatre questions, et un taux incapable d'y répondre n'est pas utilisable.

— **Que compte le numérateur** — des événements, des personnes, des épisodes ?

— **Qui est au dénominateur**, et tous étaient-ils à risque ?

— **Sur quelle période**, et tous y étaient-ils présents ?

— **Quel est le multiplicateur**, et est-ce la convention de cet indicateur ?

```python
def describe_rate(numerator, denominator, period, multiplier, at_risk_note):
    return {
        "value": multiplier * numerator / denominator,
        "numerator": numerator, "denominator": denominator,
        "period": period, "multiplier": multiplier, "at_risk": at_risk_note,
    }
```

```r
describe_rate <- function(numerator, denominator, period, multiplier, at_risk) {
  list(value = multiplier * numerator / denominator, numerator = numerator,
       denominator = denominator, period = period, at_risk = at_risk)
}
```

C'est la fiche de référence d'indicateur du module 3, réduite aux cinq champs dont une mesure épidémiologique ne peut pas se passer.

1.7 La suite

Un dénominateur est simple. La leçon suivante en enchaîne plusieurs — les cascades de soins du VIH et de la tuberculose, où le dénominateur de chaque étape est le numérateur de la précédente, et où le nombre intéressant est le maillon qui perd le plus.

CHAPITRE 2

Des dénominateurs enchaînés

Leçon 2 sur 8 · 135 min

Une cascade est une suite où le dénominateur de chaque étape est le numérateur de la précédente. 1 850 cas de protection deviennent 757 atteignant un service, et le nombre utile n'est pas les 41 % — c'est le maillon qui a le plus perdu.

2.1 La forme

La cascade de traitement du VIH est l'exemple le plus connu et la structure est générale.

```
people living with HIV
  -> diagnosed
    -> on treatment
      -> virally suppressed
```

Chaque flèche est une proportion, et **le dénominateur de chaque étape est l'étape précédente.** C'est ce qui en fait une cascade plutôt que quatre indicateurs séparés, et ce qui rend l'arithmétique digne d'attention.

La cascade de soins de la tuberculose a la même forme — cas incidents estimés, notifiés, mis sous traitement, traités avec succès — et c'est aussi celle de tout circuit de référencement, de tout calendrier vaccinal, et de tout programme dont on peut sortir entre deux étapes.

2.2 Deux façons d'exprimer chaque étape

```python
import pandas as pd

referrals = pd.read_csv("protection-referrals-2024.v1.csv")

steps = [
    ("cases", len(referrals)),
    ("consented", (referrals["consent_to_refer"] == True).sum()),
    ("referral made", ((referrals["consent_to_refer"] == True)
                       & (referrals["referral_made"] == True)).sum()),
    ("reached a service", ((referrals["consent_to_refer"] == True)
                           & (referrals["referral_made"] == True)
                           & (referrals["referral_accepted"] == True)).sum()),
]

first = steps[0][1]
previous = None
for name, count in steps:
    line = f"{name:20} {count:5}  {count / first:6.1%} of all"
    if previous:
        line += f"   {count / previous:6.1%} of previous"
    print(line)
    previous = count
```

```r
library(dplyr)
```

```r
referrals |>
  summarise(
    cases = n(),
    consented = sum(consent_to_refer),
    referred = sum(consent_to_refer & referral_made),
    reached = sum(consent_to_refer & referral_made & referral_accepted)
  )
```

Étape	n	Sur le total	Sur la précédente
Cas	1 850	100 %	—
Ont consenti au référencement	1 638	88,5 %	88,5 %
Référencement émis	1 142	61,7 %	**69,7 %**
Ont atteint un service	757	40,9 %	66,3 %

Les deux colonnes sont nécessaires et elles disent des choses différentes.

La colonne *sur le total* est la cascade telle qu'on la dessine d'ordinaire — un escalier descendant de 100 % à 41 %. Elle dit à un bailleur quelle part de la charge de cas aboutit.

La colonne *sur la précédente* est celle qui identifie le problème. La plus grosse perte unique se situe à « référencement émis » : 30,3 % des personnes ayant consenti n'ont jamais vu de référencement émis. C'est une étape sous le contrôle de l'organisation, et elle est invisible dans la première colonne, où la chute de 88,5 % à 61,7 % ressemble à celle de 61,7 % à 40,9 %.

Classez par perte conditionnelle, agissez sur la plus grande.

```python
losses = pd.DataFrame(steps, columns=["step", "n"])
losses["lost"] = losses["n"].shift(1) - losses["n"]
losses["conditional_loss"] = losses["lost"] / losses["n"].shift(1)
print(losses.dropna().sort_values("conditional_loss", ascending=False))
```

```r
tibble::tibble(step = c("consented", "referred", "reached"),
               n = c(1638, 1142, 757), previous = c(1850, 1638, 1142)) |>
  mutate(conditional_loss = (previous - n) / previous) |>
  arrange(desc(conditional_loss))
```

2.3 Le premier dénominateur est le difficile

Toute cascade a une étape qui ne se compte pas, seulement s'estime, et c'est toujours la première.

— **VIH** — les personnes vivant avec le VIH. Personne ne les compte ; le chiffre vient d'un modèle, et tout « % diagnostiqués » hérite de l'incertitude de ce modèle.

— **Tuberculose** — les cas incidents estimés. De même, et l'écart entre estimés et notifiés est le constat phare de la plupart des rapports nationaux.

— **Protection** — les personnes ayant subi un incident. Inconnaissable, et la raison pour laquelle la cascade de référencement démarre ici aux *cas connus du système* plutôt qu'au besoin.

```python
print("cascade base: cases known to the system, not people in need")
```

```
# The base of this cascade is cases the system saw. It is not incidence.
```

Dites quelle base vous avez employée. Une cascade partant d'un besoin estimé et une partant des cas déjà dans le système se ressemblent et veulent dire des choses complètement différentes — la première mesure si le système atteint les gens, la seconde seulement s'il les traite.

2.4 La cascade vaccinale

L'extraction vaccinale en porte une, et elle n'exige aucun dénominateur modélisé puisque chaque étape est comptée.

```
vax = pd.read_csv("vaccination-coverage-2024.v1.csv")
reported = vax[vax["report_submitted"] == True]
doses = reported.groupby("antigen")["doses_administered"].sum()

for first_dose, last_dose in [("penta1", "penta3"), ("mcv1", "mcv2")]:
    dropout = (doses[first_dose] - doses[last_dose]) / doses[first_dose]
    print(f"{first_dose} -> {last_dose}: {doses[first_dose]:,} -> "
          f"{doses[last_dose]:,}, dropout {dropout:.1%}")
```

```
doses <- vax |> filter(report_submitted) |>
  summarise(doses = sum(doses_administered), .by = antigen)
```

Penta1 vers penta3 : 13,8 % d'abandon. Première vers seconde dose de rougeole : 22,9 %.

Deux choses que cela rend possible et qu'un taux de couverture ne permet pas.

C'est robuste au dénominateur. Numérateur et dénominateur viennent des mêmes formations aux mêmes mois, si bien qu'une projection de recensement vieille de neuf ans ne peut pas le déplacer. Le cours sur la conception d'indicateurs avançait cet argument ; ici c'est la raison pour laquelle l'abandon est souvent le meilleur indicateur à rapporter.

Cela mesure le mécanisme. La couverture monte si davantage d'enfants commencent ; l'abandon ne baisse que si davantage d'enfants qui commencent terminent aussi, ce que change réellement une intervention de rappel ou de recherche des perdus de vue.

2.5 Surveillez le sens de l'abandon

```
if doses["penta3"] > doses["penta1"]:
    print("negative dropout: check the denominator and the period")
```

```
if (doses$doses[doses$antigen == "penta3"] > doses$doses[doses$antigen == "penta1"])
  warning("negative dropout")
```

Un abandon négatif — plus de troisièmes doses que de premières — est arithmétiquement possible et signifie toujours quelque chose de précis : une campagne de rattrapage, un problème de bornes de période, ou des enfants vaccinés ailleurs pour la dose antérieure. Ce n'est pas une erreur à ramener à zéro, c'est un constat à expliquer.

2.6 Dessinez-la, mais rapportez le tableau

Un graphique en cascade est l'une des rares visualisations par défaut réellement bonnes de ce secteur — les barres descendantes rendent les pertes immédiates. Mais le graphique ne montre que la colonne *sur le total*, alors publiez le tableau à côté avec les deux.

```python
cascade = pd.DataFrame({
    "step": ["Cases", "Consented", "Referral made", "Reached service"],
    "n": [1850, 1638, 1142, 757],
})
cascade["of_all"] = cascade["n"] / cascade["n"].iloc[0]
cascade["of_previous"] = cascade["n"] / cascade["n"].shift(1)
cascade["base"] = "cases known to the system"
```

```r
tibble::tribble(
  ~step,             ~n,
  "Cases",         1850,
  "Consented",     1638,
  "Referral made", 1142,
  "Reached",        757
) |> mutate(of_all = n / first(n), of_previous = n / lag(n))
```

Quatre colonnes, et la dernière — la base — est ce qui empêche un lecteur de comparer votre cascade à celle de quelqu'un d'autre partie d'ailleurs.

2.7 La suite

Une cascade est un ensemble de décomptes déjà agrégés. L'unité suivante revient aux enregistrements individuels — une liste linéaire d'épidémie, où l'analyse commence par mettre 975 cas dans l'ordre du temps et demander quelle forme ils dessinent.

Deuxième partie

Une épidémie, un cas à la fois

CHAPITRE 3

La courbe n’est pas l’épidémie

Leçon 3 sur 8 · 135 min

975 cas, 47 sans date de début, et un district dont le registre est rempli après coup depuis le cahier d’admission — si bien que son délai médian d’admission affiche zéro alors que sa létalité est la plus élevée des trois.

3.1 À quoi sert la courbe

Une courbe épidémique, ce sont les cas par date de début des symptômes. C’est la première chose que l’on trace dans toute épidémie et elle répond à trois questions qu’aucun tableau ne règle aussi vite — **cela croît-il, où est le pic, et la forme est-elle compatible avec une source ponctuelle ou avec une transmission interhumaine ?**

C’est aussi l’analyse la plus facilement cassée par un champ manquant, ce qui est le véritable sujet de cette leçon.

3.2 Tracez-la

```python
import pandas as pd

cases = pd.read_csv("cholera-line-list-2024.v1.csv", parse_dates=["onset_date"])

with_onset = cases[cases["onset_date"].notna()]
print(f"{len(with_onset)} of {len(cases)} cases have an onset date")

curve = (
    with_onset.assign(week=with_onset["onset_date"].dt.to_period("W").dt.start_time)
    .groupby("week").size()
)
print(curve)
print(f"peak week {curve.idxmax().date()} with {curve.max()} cases")
```

```r
library(dplyr)

cases <- readr::read_csv("cholera-line-list-2024.v1.csv")

curve <- cases |>
  filter(!is.na(onset_date)) |>
  mutate(week = lubridate::floor_date(onset_date, "week")) |>
  count(week)
```

928 des 975 cas portent une date de début, et la courbe culmine à la septième semaine — celle qui s’achève le 23 juin — avec 111 cas. **Par date de début, non d’admission, non de saisie** — tout l’objet de la courbe est le moment où les gens sont tombés malades, et chacune des deux autres dates est décalée d’un montant différent et inconnu.

Employez des pas hebdomadaires pour une épidémie de cette longueur. Le journalier est trop bruité pour être lu ; le mensuel masque entièrement le pic.

3.3 Ce que font les dates manquantes

Quarante-sept cas n'ont pas de date de début, et ils ne sont pas dispersés au hasard.

```python
missing = cases[cases["onset_date"].isna()]
print(missing.groupby("district").size())
print(f"{len(missing) / len(cases):.1%} of all cases")
```

```r
cases |> filter(is.na(onset_date)) |> count(district)
```

Trois traitements possibles, et ils donnent trois courbes différentes.

Les exclure. Le défaut, et généralement le bon. La courbe porte alors sur 928 cas et le libellé doit le dire, car 975 apparaîtra ailleurs dans le même rapport.

Les imputer d'après la date d'admission. Tentant, et cela décale toute la courbe vers la droite du délai médian, ce qui aplatit la montée et déplace le pic apparent.

Les tracer en bande séparée. Honnête et rarement fait — une barre empilée avec « début inconnu » dans une teinte distincte, placée à la semaine d'admission, pour que le lecteur voie à la fois la courbe et son incertitude.

```python
imputed = cases.copy()
fallback = pd.to_datetime(imputed["admission_date"], errors="coerce")
imputed["onset_or_admission"] = imputed["onset_date"].fillna(fallback)

both = pd.DataFrame({
    "excluded": curve,
    "imputed": (imputed.assign(
        week=imputed["onset_or_admission"].dt.to_period("W").dt.start_time)
        .groupby("week").size()),
}).fillna(0).astype(int)
print(both)
```

```r
cases |>
  mutate(onset_or_admission = coalesce(onset_date, admission_date),
         week = lubridate::floor_date(onset_or_admission, "week")) |>
  count(week)
```

Quel que soit votre choix, énoncez le dénominateur sur le graphique. « n = 928 des 975 cas avec une date de début enregistrée » fait une ligne et c'est la différence entre une courbe et une affirmation.

3.4 Le district dont la courbe n'est pas une courbe

Vient maintenant le constat pour lequel ce jeu de données a été bâti.

```python
delay = (
    pd.to_datetime(cases["admission_date"], errors="coerce") - cases["onset_date"]
).dt.days

by_district = cases.assign(delay=delay).dropna(subset=["delay"]).groupby("district")
print(by_district["delay"].median())
print((by_district["delay"].apply(lambda s: (s == 0).mean()) * 100).round(0))
```

```r
cases |>
  filter(!is.na(onset_date), !is.na(admission_date)) |>
```

```
    mutate(delay = as.integer(admission_date - onset_date)) |>
    summarise(median_delay = median(delay), zero_day = mean(delay == 0), .by = district)
```

District	Délai médian	Même jour
Nord	**0 jour**	**80 %**
Centre	2 jours	13 %
Sud	2 jours	20 %

Les cas de Nord semblent atteindre le traitement le jour même où ils tombent malades, quatre fois plus souvent que partout ailleurs. Ce n'est pas un système de santé performant. C'est un registre rempli après coup depuis le cahier d'admission — quelqu'un saisit la date d'admission dans les deux champs parce que la date de début n'a jamais été demandée.

Deux conséquences, graves toutes deux.

La courbe épidémique de Nord est une courbe d'admissions. Elle est décalée vers la droite du temps que ses cas ont réellement mis à arriver, si bien que comparer le moment du pic entre districts compare deux quantités différentes.

La statistique de délai de Nord est inutilisable, et c'est celle que la leçon suivante emploie pour expliquer la létalité — où Nord est le pire à 6,30 % contre 2,00 % au Centre. Le district qui paraît avoir l'accès le plus rapide a la mortalité la plus élevée, et la résolution est que l'apparence est un artefact.

```
print("Nord: median delay 0 days, case fatality 6.30%")
print("Centre: median delay 2 days, case fatality 2.00%")
```

```
# The contradiction is the finding: the fastest-looking district dies most.
```

Une contradiction entre deux indicateurs d'un même registre est une question de données avant d'être une question d'épidémiologie. La règle du cours de nettoyage sur les colonnes qui doivent concorder, arrivant dans un décor où s'y tromper oriente mal une réponse d'épidémie.

3.5 Lire la forme

Une fois la courbe digne de confiance, trois lectures sont classiques.

— **Un pic unique et net** suggère une source ponctuelle — un point d'eau contaminé, un événement — avec des cas apparaissant à moins d'une période d'incubation les uns des autres.

— **Une série de pics espacés d'environ une période d'incubation** suggère une propagation interhumaine.

— **Un long plateau** suggère une exposition à source commune persistante, ce qui pour le choléra signifie en général que l'approvisionnement en eau n'a pas été réparé.

Cette épidémie monte sur six semaines et décroît sur dix, ce qui est une forme propagée plutôt qu'une source ponctuelle. **Dites quelle lecture vous faites et pourquoi**, et soyez prudent — la forme est aussi affectée par le moment où la réponse a démarré, et une intervention efficace tronque la courbe d'une manière qui ressemble à une décroissance naturelle.

3.6 Tracez des cas, non des taux

Un dernier point pratique. La courbe épidémique est un décompte dans le temps, non un taux. Diviser par la population rend les districts comparables et détruit ce à quoi sert la courbe — la charge absolue arrivant chaque semaine dans les centres de traitement.

Tracez des effectifs. Mettez les taux dans le tableau à côté, ce qui est la leçon suivante.

3.7 La suite

La courbe dit quand. La leçon suivante dit combien et à quel point c'est grave — taux d'attaque avec le dénominateur de population que ce jeu de données livre, et létalité au regard du 1 % sur lequel la réponse est jugée.

CHAPITRE 4

Taux d'attaque et le 1 % que personne n'atteint

Leçon 4 sur 8 · 135 min

Le taux d'attaque exige la population ; la létalité exige un dénominateur qui exclut le cas encore hospitalisé. 3,90 % contre une cible de 1 %, 6,30 % dans un district — et l'explication repose sur une statistique de délai que la leçon précédente a montrée cassée.

4.1 Le taux d'attaque

Le taux d'attaque, ce sont les cas sur la population à risque, sur la période de l'épidémie. Malgré son nom c'est une proportion, et il exige le fichier de population que la liste linéaire livre avec elle.

```python
import pandas as pd

cases = pd.read_csv("cholera-line-list-2024.v1.csv")
population = pd.read_csv("district-population-2024.v1.csv")

by_district = (
    cases.groupby("district").size().rename("cases")
    .to_frame()
    .join(population.groupby("district")["population"].sum())
)
by_district["per_1000"] = 1000 * by_district["cases"] / by_district["population"]
print(by_district.round(2))
```

```r
library(dplyr)

cases |> count(district, name = "cases") |>
  left_join(summarise(population, population = sum(population), .by = district),
            by = "district") |>
  mutate(per_1000 = 1000 * cases / population)
```

District	Cas	Population	Pour 1 000
Nord	381	48 000	**7,94**
Centre	401	62 000	6,47
Sud	193	35 000	**5,51**

Nord est le plus touché à 7,94 et Sud le moins à 5,51, soit un écart de 2,43 pour 1 000. Retenez ce nombre : la leçon 6 montre que la moitié n'est pas ce qu'elle paraît.

Remarquez aussi que les effectifs de cas ne se classent **pas** comme les taux — Centre a le plus de cas et n'est pas le plus touché. **Un effectif répond à « où envoyer les intrants » ; un taux répond à « où le risque est le plus élevé ».** Rapportez les deux, et ne laissez jamais un diagramme en barres d'effectifs se lire comme une carte du risque.

4.2 Taux d'attaque par âge

```python
by_band = (
    cases.groupby(["district", "age_band"]).size().rename("cases")
    .to_frame()
    .join(population.set_index(["district", "age_band"])["population"])
)
by_band["per_1000"] = 1000 * by_band["cases"] / by_band["population"]
print(by_band["per_1000"].unstack().round(2))
```

```r
cases |> count(district, age_band, name = "cases") |>
  left_join(population, by = c("district", "age_band")) |>
  mutate(per_1000 = 1000 * cases / population) |>
  tidyr::pivot_wider(id_cols = age_band, names_from = district, values_from = per_1000)
```

Tranche d'âge	Nord	Centre	Sud
0-4	14,77	12,69	11,69
5-14	8,26	8,06	6,29
15-44	3,93	4,03	3,51
45+	6,41	4,75	5,71

Les moins de cinq ans sont trois à quatre fois plus touchés que les adultes en âge de travailler, dans chaque district. Ce gradient est l'épidémiologie de fond, et c'est aussi la raison pour laquelle la comparaison brute ci-dessus induit en erreur — Nord compte deux fois plus de moins de cinq ans que Sud.

4.3 La létalité

La létalité, ce sont les décès parmi les cas. Deux décisions au dénominateur, et les deux ont déjà été prises dans ce cours.

```python
with_outcome = cases[cases["outcome"].notna()]

print(f"cases: {len(cases)}, with an outcome: {len(with_outcome)}")
print(f"CFR: {(with_outcome['outcome'] == 'died').mean():.2%}")
```

```r
cases |> filter(!is.na(outcome)) |>
  summarise(n = n(), cfr = mean(outcome == "died"))
```

3,90 % sur 974 cas à issue enregistrée.

Un cas était encore hospitalisé à la date d'arrêt et n'a pas d'issue. Ce n'est ni un décès ni une guérison, et le raisonnement est le même que pour les soixante et onze enfants encore en traitement PCIMA dans le cours nutrition — **une issue en attente n'est pas une issue**, et le dénominateur doit dire lequel il a employé. Ici cela ne déplace rien ; dans une épidémie encore en cours cela déplace beaucoup.

4.4 Le seuil

Sphere et l'OMS retiennent une **létalité sous 1 %** comme marque d'une réponse choléra bien conduite. Le choléra non traité tue une grande part des cas sévères ; traité rapidement par réhydratation orale il ne tue presque personne. Le seuil est donc un énoncé sur l'accès au traitement plutôt que sur l'agent pathogène.

```python
TARGET = 0.01
cfr = (with_outcome["outcome"] == "died").mean()
print(f"CFR {cfr:.2%} against a target of {TARGET:.0%}: "
      f"{'meets' if cfr < TARGET else 'does not meet'} the standard")

by_district_cfr = with_outcome.groupby("district")["outcome"].apply(
    lambda s: (s == "died").mean()
)
print((by_district_cfr * 100).round(2))
```

```r
cases |> filter(!is.na(outcome)) |>
  summarise(cfr = mean(outcome == "died"), n = n(), .by = district)
```

District	Létalité	n
Nord	**6,30 %**	381
Sud	3,11 %	193
Centre	2,00 %	400

Tous les districts dépassent 1 %, et Nord de plus de six fois. **C'est le constat de l'épidémie** — non le taux d'attaque, qui est un fait sur l'exposition, mais la létalité, qui est un fait sur la réponse.

4.5 L'explication, et pourquoi elle est indisponible

L'explication classique d'une létalité élevée du choléra est le délai d'accès au traitement, et la liste linéaire a les champs pour la tester.

```python
delay = (
    pd.to_datetime(cases["admission_date"], errors="coerce")
    - pd.to_datetime(cases["onset_date"], errors="coerce")
).dt.days

testable = cases.assign(delay=delay).dropna(subset=["delay", "outcome"])
banded = testable.assign(
    band=pd.cut(testable["delay"], [-1, 1, 3, 99], labels=["0-1", "2-3", "4+"])
)
print(banded.groupby("band")["outcome"].apply(lambda s: (s == "died").mean()).round(3))
```

```r
cases |>
  filter(!is.na(onset_date), !is.na(admission_date), !is.na(outcome)) |>
  mutate(delay = as.integer(admission_date - onset_date),
         band = cut(delay, c(-1, 1, 3, 99), labels = c("0-1", "2-3", "4+"))) |>
  summarise(cfr = mean(outcome == "died"), n = n(), .by = band)
```

Parmi les cas admis, le gradient existe mais reste modeste — environ 1,9 % de zéro à trois jours et 3,3 % à quatre jours ou plus. Il est plus faible que l'écart entre districts, et il y a deux raisons à cela, l'une de fond et l'autre non.

La raison de fond — l'admission elle-même est la protection. Les cas jamais admis portent l'essentiel de la mortalité, et la variable de délai n'existe que pour ceux qui ont été admis. Comparer des bandes de délai à l'intérieur des cas admis conditionne sur ce qui compte le plus.

Celle qui n'en est pas — le délai de Nord est fabriqué par son registre. La leçon précédente a trouvé 80 % de ses cas avec début égal à admission. Le district ayant la

plus forte létalité rapporte donc le délai le plus court, et la comparaison qui expliquerait sa mortalité est précisément celle que ses données ne peuvent pas soutenir.

```
print("Nord CFR 6.30% with a median recorded delay of 0 days.")
print("The delay is a register artefact; the CFR is not.")
```

```
# One of these two numbers is real. Say which, in the report.
```

Consignez cela comme une limite, non comme un résultat. « La létalité est la plus élevée à Nord ; le délai entre début et admission qui permettrait de tester l'explication habituelle n'y est pas fiable, car 80 % de ses cas enregistrent début et admission le même jour » est la phrase honnête, et elle engendre aussi l'action corrective — réparer le registre — qu'une explication fabriquée ne produirait pas.

4.6 Le bloc de rapportage

```
Cholera outbreak, weeks 1-16

  Cases                      975      attack rate 6.7 per 1,000 (145,000 population)
  Attack rate by district    7.94 / 6.47 / 5.51 per 1,000 (Nord / Centre / Sud)
                             crude; see standardised rates before comparing

  Deaths                      38      case fatality 3.90%
  Denominator                974      one case still admitted at cut-off, excluded
  Against Sphere target       1%      not met in any district

  Highest CFR: Nord at 6.30% (n=381). Onset-to-admission delay is not usable
  in Nord (80% of cases record onset = admission), so the usual explanation
  cannot be tested there from this register.
```

Dix lignes, et ce sont les trois dernières qui en font un constat plutôt qu'un tableau.

4.7 La suite

Les taux d'attaque par district ressemblent à une comparaison et n'en sont pas encore une, car les districts ne contiennent pas les mêmes gens. L'unité suivante y remédie, après avoir d'abord réglé une question de couverture ouverte depuis le module 2.

Troisième partie

Comparer des lieux

CHAPITRE 5

Trois couvertures qui divergent

Leçon 5 sur 8 · 135 min

Couverture administrative, couverture par enquête, et abandon entre doses. La première dépend d'une projection, la deuxième d'un échantillon, et seule la troisième échappe aux deux.

5.1 Trois façons de mesurer la même chose

La couverture vaccinale est l'indicateur sanitaire le plus rapporté de ce secteur et il se produit de trois façons différentes, qui divergent régulièrement.

Méthode	Numérateur	Dénominateur	Échoue quand
Administrative	Doses enregistrées par les formations	Projection de population	La projection est fausse, ou le
Par enquête	Enfants avec carte ou dose rappelée	Enfants échantillonnés	Les cartes sont perdues ; le ra
Abandon	Première dose moins dernière dose	Première dose	Rarement — et c'est tout l'int

Le cours DHIS2 a calculé la première. Le cours sur les enquêtes vous a donné la machinerie de la deuxième. Cette leçon dit pourquoi elles diffèrent et laquelle employer.

5.2 La couverture administrative et ses deux dépendances

```python
import pandas as pd

vax = pd.read_csv("vaccination-coverage-2024.v1.csv", parse_dates=["period"])
reported = vax[vax["report_submitted"] == True]

penta3 = reported[reported["antigen"] == "penta3"]
annual_target = (
    vax[vax["antigen"] == "penta3"].groupby("facility_id")["target_population"].first()
)
months = penta3.groupby("facility_id").size()
prorated = (annual_target * months.reindex(annual_target.index).fillna(0) / 12).sum()

print(f"doses {penta3['doses_administered'].sum():,}")
print(f"administrative coverage: "
      f"{penta3['doses_administered'].sum() / prorated:.1%}")
```

```r
library(dplyr)
# same shape: doses over a denominator pro-rated to months reported
```

77,5 %, parmi les couples formation-mois ayant rapporté. Ce chiffre porte deux dépendances, et le cours DHIS2 a établi les deux.

Le dénominateur est une projection. Les nourrissons survivants, issus d'une projection de recensement composée à un taux de croissance supposé. Neuf ans à 2,4 % font un facteur de 1,24, si bien qu'un quart du dénominateur est une hypothèse.

Le taux de rapportage valait 76,5 %. Les formations n'ayant pas rapporté n'apportent aucune dose, si bien que le chiffre est une borne inférieure à moins d'imputer leurs doses.

Une couverture administrative supérieure à 100 % est fréquente et signifie toujours que l'une de ces deux choses est fausse, plus une troisième possibilité — des enfants vaccinés hors de leur bassin, ce qui gonfle un numérateur de formation face à un dénominateur de bassin.

5.3 La couverture par enquête et ses deux dépendances

Une enquête de couverture échantillonne des enfants d'un âge donné et demande s'ils ont été vaccinés, en vérifiant sur la carte lorsqu'il y en a une.

Elle supprime le problème de projection — le dénominateur est l'échantillon — et en introduit deux autres.

La conservation des cartes. Là où les cartes sont perdues, la couverture repose sur le souvenir de l'accompagnant, qui surestime pour certains antigènes et sous-estime pour d'autres. Rapportez séparément la couverture vérifiée par carte et celle fondée sur le rappel ; l'écart entre les deux mesure à quel point l'estimation dépend de la mémoire.

La précision. La couverture par enquête arrive avec un intervalle de confiance, et le cours sur les enquêtes a montré qu'un plan en grappes l'élargit. Une enquête qui estime la couverture à plus ou moins six points ne peut pas trancher si un district a franchi une cible de 80 %.

```python
# From the survey course: a proportion with a design-adjusted interval
def coverage_interval(p, n, deff, t=1.99):
    se = (p * (1 - p) / n * deff) ** 0.5
    return p - t * se, p + t * se

print(coverage_interval(0.775, 900, 2.0))
```

```r
coverage_interval <- function(p, n, deff, t = 1.99) {
  se <- sqrt(p * (1 - p) / n * deff)
  c(p - t * se, p + t * se)
}
```

5.4 Pourquoi elles divergent

Cinq raisons, et savoir laquelle s'applique change ce que vous faites.

— **La projection est fausse.** La plus fréquente, et elle ne déplace que la couverture administrative. Un dénominateur trop petit pousse l'administrative au-dessus de l'enquête.

— **Le rapportage est incomplet.** Fait baisser l'administrative.

— **La traversée des bassins.** Déplace les chiffres administratifs par formation dans les deux sens et s'annule au niveau du district.

— **L'erreur de rappel.** Déplace la couverture par enquête, généralement vers le haut.

— **Des cohortes d'âge différentes.** La couverture administrative compte les doses données cette année ; une enquête compte les enfants de 12 à 23 mois vaccinés à n'importe quel moment. Ce ne sont pas les mêmes enfants.

La dernière est la plus fréquente et la moins soupçonnée. Deux chiffres de « couverture penta3 » portant sur des cohortes différentes ne sont pas deux estimations d'une quantité ; ce sont deux quantités.

5.5 L'abandon — le chiffre qui survit à tout cela

```python
doses = reported.groupby("antigen")["doses_administered"].sum()

for start, end in [("penta1", "penta3"), ("mcv1", "mcv2")]:
    dropout = (doses[start] - doses[end]) / doses[start]
    print(f"{start} -> {end}: {dropout:.1%}")
```

```r
vax |> filter(report_submitted) |>
  summarise(doses = sum(doses_administered), .by = antigen)
```

Penta1 vers penta3 : 13,8 %. Première vers seconde dose de rougeole : 22,9 %.

L'abandon est un rapport de deux numérateurs de même source, mêmes formations et mêmes mois. Il a donc :

— **aucun dénominateur de population**, donc aucune projection à se tromper ;

— **le même biais de rapportage dans les deux termes**, qui s'annule largement ;

— **aucune ambiguïté de cohorte**, puisque les deux doses sont comptées pareillement.

Le coût est qu'il ne dit rien du niveau. Un district pourrait avoir 13,8 % d'abandon et 30 % de couverture — bon à retenir les enfants qu'il commence, mauvais à les commencer. **Rapportez l'abandon à côté de la couverture, jamais à sa place.**

5.6 Ce que les deux abandons disent ensemble

L'abandon de la seconde dose de rougeole vaut 22,9 % contre 13,8 % pour le pentavalent, et la différence est instructive plutôt que du bruit.

Les doses de pentavalent se donnent rapprochées dans les premiers mois de vie, quand les accompagnants viennent déjà pour d'autres services. La seconde dose de rougeole se donne bien plus tard, souvent après la fin de la période de contacts intensifs. **Un abandon qui croît avec l'intervalle entre doses est un problème de rappel et de proximité**, non d'approvisionnement, et les actions correctives diffèrent.

```python
summary = pd.DataFrame({
    "series": ["Pentavalent 1->3", "Measles 1->2"],
    "dropout": [0.138, 0.229],
    "interval": ["weeks", "months"],
    "implication": ["retention within the early contact period",
                    "retention after the early contact period ends"],
})
```

```r
tibble::tribble(
  ~series,             ~dropout, ~implication,
  "Pentavalent 1->3",     0.138, "retention within early contacts",
  "Measles 1->2",         0.229, "retention after early contacts end"
)
```

5.7 Laquelle rapporter

```
Penta3 coverage, administrative     77.5%   among reporting facility-months
```

```
                                        (reporting rate 76.5%)
                                        denominator: MoH projection from 2015 census
Penta1 to penta3 dropout          13.8%  same source, both terms
Measles 1 to 2 dropout            22.9%
Survey coverage                   not available this year
```

Quatre lignes, et la dernière est un constat. **Là où aucune enquête n'existe, dites-le** — car le lecteur suppose par défaut qu'un chiffre de couverture a été validé face à une enquête, et ici il ne l'a pas été.

5.8 La suite

La couverture compare un programme à une population. La leçon suivante compare deux populations entre elles, et découvre que la comparaison de la leçon 4 était en partie un artefact de qui habite où.

CHAPITRE 6

La moitié de l'écart tenait à qui habite là

Leçon 6 sur 8 · 135 min

Le taux d'attaque brut de Nord vaut 7,94 pour 1 000 et celui de Sud 5,51. Standardisés sur une population commune ils valent 7,25 et 5,95, si bien que l'écart tombe de 2,43 à 1,30 — et la moitié disparue était la structure par âge.

6.1 Deux districts ne sont pas comparables tels quels

La leçon 4 a posé trois taux d'attaque bruts côte à côte et la comparaison paraissait simple. Elle ne l'est pas, et la raison est dans le fichier de population.

```python
import pandas as pd

population = pd.read_csv("district-population-2024.v1.csv")

structure = (
    population.pivot(index="district", columns="age_band", values="population")
)
shares = structure.div(structure.sum(axis=1), axis=0)
print((shares * 100).round(1))
```

```r
library(dplyr)

population |>
  mutate(share = population / sum(population), .by = district) |>
  tidyr::pivot_wider(id_cols = district, names_from = age_band, values_from = share)
```

District	0-4	5-14	15-44	45+
Nord	**22 %**	30 %	35 %	13 %
Centre	15 %	25 %	42 %	18 %
Sud	**11 %**	20 %	44 %	25 %

Nord compte deux fois plus de moins de cinq ans que Sud, et la leçon 4 a établi que les moins de cinq ans ont trois à quatre fois le taux d'attaque des adultes en âge de travailler.

Nord aurait donc un taux brut plus élevé que Sud **même si chaque taux par âge des deux districts était identique**. La comparaison brute mélange deux choses — à quel point chaque district est risqué, et qui y habite.

6.2 La standardisation directe

La méthode tient en une ligne d'arithmétique appliquée systématiquement — **calculez les taux par âge de chaque district, puis appliquez-les à une seule population commune.**

```python
cases = pd.read_csv("cholera-line-list-2024.v1.csv")

observed = (
    cases.groupby(["district", "age_band"]).size().rename("cases").to_frame()
    .join(population.set_index(["district", "age_band"])["population"])
)
observed["rate"] = observed["cases"] / observed["population"]

standard = population.groupby("age_band")["population"].sum()
standard_share = standard / standard.sum()

standardised = (
    observed["rate"].unstack()          # district x age_band
    .mul(standard_share, axis=1).sum(axis=1)
)
crude = (
    cases.groupby("district").size()
    / population.groupby("district")["population"].sum()
)

comparison = pd.DataFrame({
    "crude_per_1000": 1000 * crude,
    "standardised_per_1000": 1000 * standardised,
})
comparison["difference"] = (
    comparison["standardised_per_1000"] - comparison["crude_per_1000"]
)
print(comparison.round(2))
```

```r
observed <- cases |> count(district, age_band, name = "cases") |>
  left_join(population, by = c("district", "age_band")) |>
  mutate(rate = cases / population)

standard <- population |> summarise(pop = sum(population), .by = age_band) |>
  mutate(share = pop / sum(pop))

observed |>
  left_join(standard, by = "age_band") |>
  summarise(standardised = sum(rate * share), .by = district)
```

District	Brut	Standardisé	Écart
Nord	7,94	**7,25**	−0,69
Centre	6,47	6,60	+0,13
Sud	5,51	**5,95**	+0,44

Nord baisse, Sud monte, et l'écart entre eux se resserre de 2,43 à 1,30 pour 1 000. Environ la moitié de la différence apparente entre le pire et le meilleur district était une structure par âge et non un risque.

Les 1,30 restants sont réels. Nord est réellement plus touché — ses taux par âge sont plus élevés dans trois bandes sur quatre — et c'est la standardisation qui permet de le dire plutôt que de l'affirmer.

6.3 Ce qu'est la population standard, et pourquoi elle compte

Le standard est la population à laquelle vous appliquez les taux de chaque district. Trois choix courants.

— **La population combinée de l'étude**, comme ci-dessus. Simple, défendable, et interne — les taux standardisés sont comparables entre eux et à rien d'autre.

— **Une population nationale.** Permet la comparaison avec d'autres districts standardisés de la même façon.

— **Un standard mondial publié** — la population type de l'OMS ou de Segi. Permet la comparaison internationale, et produit des nombres qui ne ressemblent en rien aux taux bruts.

Énoncez le standard. Un taux standardisé ne s'interprète que face à d'autres standardisés sur la même population, et deux rapports employant des standards différents produisent des nombres incomparables qui ont tous deux l'air officiel.

```
print("Standard: combined population of the three districts, 145,000")


# Say it in the table caption, every time.
```

6.4 La standardisation indirecte, et quand elle est nécessaire

La standardisation directe exige des taux par âge pour chaque district, ce qui exige assez de cas dans chaque cellule. Là où un district compte quatre cas dans une bande, son taux par âge est instable et la méthode directe propage cette instabilité.

La méthode indirecte inverse le problème — appliquer un **jeu de taux standard** à la population propre de chaque district, et comparer les cas observés aux attendus.

```
standard_rates = observed.groupby("age_band").apply(
    lambda g: g["cases"].sum() / g["population"].sum()
)

expected = (
    population.assign(rate=population["age_band"].map(standard_rates))
    .assign(expected=lambda d: d["population"] * d["rate"])
    .groupby("district")["expected"].sum()
)
smr = cases.groupby("district").size() / expected
print(smr.round(3))
```

```
standard_rates <- observed |>
  summarise(rate = sum(cases) / sum(population), .by = age_band)

population |>
  left_join(standard_rates, by = "age_band") |>
  summarise(expected = sum(population * rate), .by = district) |>
  left_join(count(cases, district, name = "observed"), by = "district") |>
  mutate(smr = observed / expected)
```

Le résultat est un **rapport standardisé de morbidité** — observés sur attendus, où 1,0 signifie que le district a exactement les cas que sa structure par âge prédit. Au-dessus de 1 c'est pire qu'attendu, en dessous c'est mieux.

Employez la standardisation indirecte quand les cellules sont petites, et dites quelle méthode vous avez employée — les deux répondent à des questions légèrement différentes et leurs nombres ne sont pas interchangeables.

6.5 L'âge n'est pas le seul facteur de confusion

La standardisation retire la variable sur laquelle vous standardisez, et rien d'autre. Les taux d'attaque du choléra varient aussi avec la source d'eau, la densité de population, la distance à un centre de traitement et le statut de déplacement, et aucun de ces facteurs n'est dans le fichier de population.

Standardiser sur l'âge puis affirmer que la différence restante est la performance du programme est l'erreur que cette leçon rend possible, et la leçon suivante lui est entièrement consacrée.

6.6 Rapportez la paire

```
Cholera attack rate by district, weeks 1-16

  District   Crude    Age-standardised   Population
  Nord       7.94     7.25               48,000
  Centre     6.47     6.60               62,000
  Sud        5.51     5.95               35,000

  Standardised directly to the combined population of the three districts.
  About half the crude Nord-Sud difference is age structure: Nord has 22% of
  its population under five against Sud's 11%, and under-fives have three to
  four times the attack rate of adults aged 15-44.
```

Les deux colonnes, le standard nommé, et une phrase disant de combien cela a bougé et pourquoi. Un tableau qui n'a que la colonne standardisée cache qu'un ajustement a eu lieu ; un tableau qui n'a que la colonne brute invite à une comparaison à moitié démographique.

6.7 La suite

La standardisation a retiré une explication alternative. La leçon suivante liste les autres — et demande à quoi une différence entre deux districts a le droit d'être attribuée quand aucune d'elles ne peut être retirée.

Quatrième partie

Ce que la comparaison peut affirmer

CHAPITRE 7

Qu'est-ce qui pourrait aussi l'expliquer

Leçon 7 sur 8 · 135 min

Un facteur de confusion cause le résultat et diffère entre les groupes. Quatre candidats séparent les districts de cette épidémie — l'un a été retiré, l'un n'est pas mesuré, et l'un n'est pas un facteur de confusion du tout : c'est le mécanisme.

7.1 La définition, et pourquoi il vaut la peine d'être strict

Un **facteur de confusion** est une variable qui satisfait trois conditions à la fois :

— elle est **associée à l'exposition** — elle diffère entre les groupes comparés ;

— elle est une **cause du résultat**, indépendamment de l'exposition ;

— elle n'est **pas sur le chemin causal** entre l'exposition et le résultat.

Les trois comptent. Une variable qui diffère entre groupes sans causer le résultat est sans objet. Une variable qui cause le résultat mais est identique dans les deux groupes ne peut pas expliquer un écart. Et une variable sur le chemin causal n'est pas un facteur de confusion — c'est le mécanisme, et ajuster dessus détruit le constat.

7.2 Les quatre candidats de cette épidémie

Nord a un taux d'attaque standardisé de 7,25 pour 1 000 contre 5,95 pour Sud, et une létalité de 6,30 % contre 3,11 %. Qu'est-ce qui pourrait aussi expliquer ces écarts ?

Candidat	Diffère entre districts ?	Cause le résultat ?	Sur le chemin ?
Structure par âge	Oui, nettement	Oui	Non — facteur de confusion, déjà
Source d'eau et densité	Presque certainement	Oui	Non — facteur de confusion, **non**
Distance au traitement	Oui	Oui, pour la létalité	**Oui** — c'est le mécanisme
Qualité du registre	Oui	Non	Non — pas un facteur de confusio

Parcourez ce tableau et chaque ligne appelle une réponse différente.

La structure par âge a été traitée. La leçon 6 l'a retirée et a chiffré ce qu'elle valait — environ la moitié de l'écart brut.

La source d'eau, l'assainissement et la promiscuité sont les causes réelles de la transmission du choléra, elles diffèrent certainement entre un district de type camp et un district rural, et **elles ne sont pas dans ce jeu de données**. C'est la phrase la plus importante de cette leçon. Un facteur de confusion non mesuré ne peut pas être ajusté, et son existence doit être énoncée plutôt qu'ignorée.

```
print("Available for adjustment: age, sex, district")
print("Known to matter and unavailable: water source, sanitation, crowding, "
      "displacement status")
```

```
# The list of what you could not adjust for belongs in the limitations section.
```

La distance au traitement est l'intéressante et n'est pas du tout un facteur de confusion. La chaîne est : Nord est plus loin des centres de traitement → ses cas arrivent plus tard → davantage meurent. La distance cause la létalité *à travers* le délai, donc le délai est sur le chemin causal. **Ajuster sur le délai supprimerait précisément l'effet que vous cherchez à démontrer**, ce qui est l'erreur classique de sur-ajustement.

La qualité du registre n'est pas non plus un facteur de confusion. Les dates de début rétro-remplies de Nord ne causent pas de décès ; elles faussent la mesure de l'explication. C'est un biais d'information, et il se corrige en réparant le registre, non en ajustant.

7.3 Stratifiez pour le voir

L'ajustement le plus simple, et celui qui montre son travail.

```python
import pandas as pd

cases = pd.read_csv("cholera-line-list-2024.v1.csv")
with_outcome = cases[cases["outcome"].notna()]

stratified = (
    with_outcome.assign(died=with_outcome["outcome"] == "died")
    .groupby(["district", "age_band"])
    .agg(cases=("died", "size"), deaths=("died", "sum"))
)
stratified["cfr"] = stratified["deaths"] / stratified["cases"]
print((stratified["cfr"].unstack() * 100).round(1))
```

```r
library(dplyr)

cases |>
  filter(!is.na(outcome)) |>
  summarise(cases = n(), cfr = mean(outcome == "died"), .by = c(district, age_band)) |>
  tidyr::pivot_wider(id_cols = age_band, names_from = district, values_from = cfr)
```

Deux choses se lisent sur un tableau stratifié, et une seule est l'estimation ajustée.

L'écart persiste-t-il à l'intérieur des strates ? Si la létalité de Nord est plus élevée dans chaque tranche d'âge, l'âge n'est pas l'explication. Si elle s'inverse dans certaines tranches, il se passe quelque chose de plus intéressant.

Les strates sont-elles cohérentes ? Un écart grand dans une tranche et absent des autres est une **modification d'effet** — l'exposition agit réellement différemment selon les groupes — et il doit être rapporté par strate plutôt que moyenné en un seul chiffre ajusté.

La modification d'effet est un constat ; la confusion est une nuisance. Réduire la première à un nombre unique détruit le résultat ; ne pas retirer la seconde en fabrique un.

7.4 Quand la stratification s'épuise

Deux variables et les cellules se vident vite — la leçon sur la désagrégation du module 3, arrivant avec une autre conséquence. Trois districts par quatre tranches d'âge par deux sexes font vingt-quatre cellules sur 974 issues, et les décès par choléra sont rares.

```python
cells = with_outcome.groupby(["district", "age_band", "sex"]).size()
print(f"{len(cells)} cells, smallest {cells.min()}, {(cells < 30).sum()} below 30")
```

```r
cases |> filter(!is.na(outcome)) |> count(district, age_band, sex) |>
```

```
    summarise(cells = n(), smallest = min(n), under_30 = sum(n < 30))
```

C'est là que la régression prend le relais, et *Régression appliquée aux données de programme*, plus loin dans le programme, porte exactement là-dessus — ajuster sur plusieurs variables à la fois sans épuiser les cellules. **Son objet est le même que celui de la stratification**, et un modèle qui ajuste sur un mécanisme commet la même erreur qu'une stratification qui le fait.

7.5 Les trois questions à poser à toute comparaison

Avant d'attribuer un écart à quoi que ce soit.

1. Qu'est-ce qui diffère entre ces groupes en dehors de ce que j'étudie ? Listez-le. La liste est la section des limites, et ce que vous ne pouvez pas mesurer y figure aussi.

2. Lesquels causent le résultat ? Ceux-là seuls sont des facteurs de confusion. Un district ayant davantage d'écoles ne confond pas une comparaison sur le choléra à moins que les écoles ne causent le choléra.

3. Lesquels sont sur le chemin causal ? N'ajustez pas dessus. Le délai d'accès au traitement, dans cette épidémie, est la façon dont la distance tue.

```
adjustment_plan = pd.DataFrame({
    "variable": ["age", "sex", "water source", "delay to treatment", "register quality"],
    "differs": [True, False, True, True, True],
    "causes_outcome": [True, True, True, True, False],
    "on_causal_path": [False, False, False, True, False],
    "action": ["adjust", "no need", "unmeasured; state it", "do not adjust",
               "fix the register"],
})
print(adjustment_plan)
```

```
tibble::tribble(
  ~variable,        ~action,
  "age",            "adjust",
  "water source",   "unmeasured; state it",
  "delay",          "do not adjust; it is the mechanism",
  "register",       "fix it; this is bias, not confounding"
)
```

Écrivez ce tableau avant l'analyse, non après. Décider sur quoi ajuster une fois qu'on a vu quel ajustement donne la réponse la plus agréable est ce qui rend l'épidémiologie observationnelle peu digne de confiance, et c'est indiscernable d'un travail honnête après coup.

7.6 La suite

Vous avez nommé ce qui pourrait aussi expliquer un écart et retiré ce que vous pouviez. La dernière leçon porte sur ce qui reste dicible — la phrase à laquelle une comparaison observationnelle a droit, et les plusieurs auxquelles elle n'a pas droit.

CHAPITRE 8

La phrase à laquelle cette analyse a droit

Leçon 8 sur 8 · 135 min

La létalité de Nord vaut 6,30 % contre 2,00 % pour Centre, et le gradient du délai repose sur un décès parmi trente cas. Quatre affirmations pourraient en sortir, trois sont fausses, et l'écart entre elles est tout ce cours.

8.1 Quatre phrases, un seul jeu de nombres

Tout ce que ce cours a calculé pointe vers une comparaison, et il y a quatre façons de la rédiger.

1. *« La réponse de Nord échoue : la létalité y est le triple de celle de Centre. »*
2. *« La présentation tardive tue — les cas admis quatre jours ou plus après le début meurent à 3,3 % contre 1,9 % pour ceux admis plus tôt. »*
3. *« La létalité de Nord vaut 6,30 % (24 décès parmi 381 cas avec une issue enregistrée) contre 2,00 % pour Centre (8 sur 400). Les deux dépassent le seuil Sphère de 1 %. Le délai début-admission n'explique pas l'écart dans ces données : les dates de début de Nord sont rétro-remplies depuis le registre d'admission, et la bande de délai la plus élevée porte un décès parmi trente cas. »*
4. *« Les données ne permettent pas de comparer les districts. »*

La troisième est la seule défendable, et c'est aussi la seule assez longue pour être utile. Comprenez pourquoi chacune des autres échoue et vous avez le contenu de cette leçon.

8.2 Pourquoi les trois premières échouent

« La réponse de Nord échoue » attribue un écart à une cause que les données ne distinguent pas. Nord diffère de Centre par la structure par âge, la source d'eau, la distance au traitement, le statut de déplacement et la qualité du registre. La leçon 6 en a retiré une, la leçon 7 a nommé les autres et en a trouvé la moitié non mesurées. **Une comparaison dont un seul facteur de confusion a été retiré n'est pas une comparaison sans aucun.**

« La présentation tardive tue » est une affirmation plus forte que le plan d'étude ne le permet, et elle repose sur deux cellules incapables de la porter. Le délai début-admission se mesure sur un champ qu'un district rétro-remplit — **80 % des cas de Nord affichent un délai nul**, ce qui est le registre d'admission qui parle et non une présentation rapide. Et la bande de quatre jours ou plus, qui produit les 3,3 %, contient **un décès parmi trente cas** : déplacez ce seul cas et le gradient disparaît.

```
import pandas as pd

cases = pd.read_csv("cholera-line-list-2024.v1.csv", parse_dates=["onset_date"])
with_onset = cases[cases["onset_date"].notna()]

delayed = with_onset.assign(delay=lambda d: (
```

```
    pd.to_datetime(d["admission_date"]) - d["onset_date"]).dt.days)

print(delayed.groupby("district")["delay"]
      .agg(median="median", same_day=lambda s: (s == 0).mean()).round(3))

band = pd.cut(delayed["delay"], [-1, 1, 3, 99], labels=["0-1", "2-3", "4+"])
print(delayed.dropna(subset=["outcome"]).groupby(band).size())
```

```
library(dplyr)

cases |>
  filter(!is.na(onset_date)) |>
  mutate(delay = as.numeric(admission_date - onset_date)) |>
  summarise(median = median(delay), same_day = mean(delay == 0), .by = district)
```

Le délai médian de Nord vaut zéro jour avec 80 % de cas le jour même ; ceux de Centre et de Sud valent deux jours. Et les trois bandes contiennent 372, 214 et 30 cas avec une issue. **Imprimez toujours les effectifs sous un gradient**, car un pourcentage s'imprime de la même façon qu'il repose sur trois cents cas ou sur un seul.

« Les données ne permettent pas de comparer » est le mode d'échec de l'analyste prudent, et c'est le plus coûteux des quatre. Une réponse qui cesse de rapporter parce que ses données sont imparfaites laisse la décision se prendre sur rien, et il n'existe nulle part une épidémie aux données propres. **Le métier consiste à dire ce que les données permettent, limites attachées, non à refuser.**

8.3 Les quatre éléments d'une affirmation

Chacun manque à au moins une des trois premières phrases.

Le numérateur et le dénominateur, en effectifs. « 6,30 % » est invérifiable ; « 24 décès parmi 381 cas avec une issue enregistrée » se contrôle et s'additionne à celui d'un autre district. Cela déclare aussi l'exclusion — 974 des 975 cas ont une issue, et dire quel dénominateur vous avez employé est ce qui permet au lecteur de savoir si le cas encore hospitalisé a été compté.

La référence à laquelle elle se compare. Un taux sans point de comparaison est un nombre en quête d'opinion. Sphère place la létalité du choléra sous 1 % ; 3,90 % en vaut quatre fois plus et la phrase doit le dire.

L'ajustement, ou son absence. Si vous avez standardisé, nommez le standard. Si vous ne l'avez pas fait, dites que la comparaison est brute.

La limite susceptible de changer le sens. Non pas un paragraphe de précautions — la ou les deux choses qui, si elles allaient dans l'autre sens, changeraient ce que quelqu'un devrait faire. Les dates rétro-remplies de Nord en sont une. La source d'eau non mesurée est l'autre.

```
claim = {
    "numerator": 24,
    "denominator": 381,
    "denominator_definition": "cases with a recorded outcome",
    "value": "6.30%",
    "benchmark": "Sphere: below 1%",
    "adjustment": "none; crude",
    "limitation": "onset dates back-filled in this district",
}
```

```
# The same fields, in the footnote of the table.
```

8.4 Force de la preuve, et où se situe la donnée de routine

Tous les constats observationnels ne sont pas également faibles, et les différences ont des noms.

Plan d'étude	Ce qu'il permet d'affirmer	Où il apparaît ici
Donnée de routine, transversale	Association, ajustée sur ce qui a été mesuré	Tout ce cours
Cohorte	Association avec l'exposition précédant le résultat	Le registre PCIMA, suivi
Cas-témoins	Association, efficacement pour des résultats rares	Rechercher la source d'u
Essai randomisé	Causalité	Presque jamais en suivi-é

La donnée de programme vit sur la première ligne, et cette ligne peut porter beaucoup de poids si elle est honnête sur la ligne où elle se trouve. Ce qu'elle ne peut pas faire, c'est sauter à la dernière parce que l'association est grande ou le mécanisme plausible.

Deux des arguments les plus solides ouverts au travail observationnel restent disponibles, et cette épidémie en fournit un exemple de chaque.

Un constat qui survit à l'ajustement sur tout ce que vous avez pu mesurer est plus crédible qu'un constat qui s'évanouit. L'excès de Nord s'est resserré sous la standardisation par âge sans disparaître, et c'est le plus solide des deux résultats ici.

Un gradient dose-réponse — une létalité croissant régulièrement d'une bande de délai à l'autre plutôt que sautant à un seul seuil — est plus difficile à écarter qu'un contraste isolé. C'est l'argument qui n'était **pas** disponible : la bande supérieure comptait trente cas, et un gradient reposant sur un décès n'est pas un gradient.

Ni l'un ni l'autre, pris seul, ne rend un constat causal. Savoir lequel vous avez est ce qui dit avec quelle fermeté écrire.

8.5 Écrivez la limite comme une décision, non comme une excuse

Le paragraphe de limites que porte la plupart des rapports est une liste de regrets. Faites-en plutôt une liste de conséquences.

```
Cholera case fatality, weeks 1-16

  Overall     3.90%   38 deaths / 974 cases with an outcome
  Nord        6.30%   24 / 381
  Sud         3.11%    6 / 193
  Centre      2.00%    8 / 400

  Against the Sphere threshold of 1%, every district is above standard.

  What this supports: prioritising treatment capacity in Nord.
  What it does not: attributing the gap to clinical management or to late
  presentation. Nord's register back-fills onset dates, so its delays are not
  comparable, and the 4+ day band holds one death among 30 cases. Water
  source and crowding, the main determinants of cholera severity, are not
  recorded at all.
  What would settle it: two weeks of prospectively recorded onset dates in
  Nord, and the WASH assessment already scheduled for week 18.
```

« Ce qui trancherait » est la ligne qui rend la limite utile. Une limite sans issue se lit comme une excuse ; une limite avec une étape suivante nommée est un plan de travail, et c'est ce qui fait d'une analyse imparfaite la première moitié d'une meilleure.

8.6 L'habitude que ce cours visait

Trois questions, dans cet ordre, avant toute rédaction.

1. **Qu'ai-je compté, sur quoi ?** Numérateur, dénominateur, et qui a été exclu.
2. **Qu'est-ce qui pourrait aussi l'expliquer ?** Le tableau de la leçon 7, écrit avant l'analyse et non après.
3. **Que devrait-on faire différemment à cause de cela ?** Si rien, l'analyse ne valait pas le temps passé ; si quelque chose, l'affirmation doit être assez solide pour le porter.

Chaque mesure de ce cours — incidence et prévalence, la cascade, la courbe épidémique, les taux d'attaque, la létalité, la couverture de trois façons, le taux standardisé — répond à la première. Les deux dernières sont ce qui empêche la première d'être détournée, et c'est la part qu'aucun logiciel ne calcule à votre place.

8.7 La suite

Vous savez défendre un taux, ajuster une comparaison et énoncer ce qu'elle a le droit d'affirmer. Le reste du module 4 applique la même discipline à d'autres secteurs — niveaux de service EAH, assiduité scolaire, gestion de cas en protection — et *Régression appliquée aux données de programme* revient au problème de la leçon 7 avec un outil capable de tenir plusieurs facteurs de confusion à la fois.

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/public-health-epidemiology

Généré le 28 juillet 2026.