

La moitié de l'écart tenait à qui habite là

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.

Pierre Robentz Cassion

Leçon 6 sur 8

Santé publique et épidémiologie appliquée — Comparer des lieux

Ce que couvre cette leçon

- Deux districts ne sont pas comparables tels quels
- La standardisation directe
- Ce qu'est la population standard, et pourquoi elle compte
- La standardisation indirecte, et quand elle est nécessaire
- L'âge n'est pas le seul facteur de confusion
- Rapportez la paire
- La suite

Deux districts ne sont pas comparables tels quels — En 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))
```

Deux districts ne sont pas comparables tels quels — En R

```
library(dplyr)

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

Deux districts ne sont pas comparables tels quels

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

Deux districts ne sont pas comparables tels quels

- 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 . . .

La standardisation directe — En Python (suite)

```
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 = (
```

La standardisation directe — En Python (suite)

```
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))
```

La standardisation directe — En 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)
```

La standardisation directe

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. . .

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

- **La population combinée de l'étude**, comme ci-dessus. Simple, défendable, et interne — les taux standardisés sont. . .
- **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. . .
- **Énoncez le standard** — Un taux standardisé ne s'interprète que face à d'autres standardisés sur la même population, et. . .

Ce qu'est la population standard, et pourquoi elle compte — En Python

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

Ce qu'est la population standard, et pourquoi elle compte — En R

Say it in the table caption, every time.

La standardisation indirecte, et quand elle est nécessaire — En Python

```
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))
```

La standardisation indirecte, et quand elle est nécessaire — En R

```
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)
```

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

- Standardiser sur l'âge puis affirmer que la différence restante est la performance du programme est l'erreur que cette. . .

Rapportez la paire — Exemple

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.

- La standardisation a retiré une explication alternative.

Lire la leçon complète, avec le code exécutable

[https://data-analysis.cassion.dev/fr/cours/public-health-epidemiology/
epi-06-age-standardisation](https://data-analysis.cassion.dev/fr/cours/public-health-epidemiology/epi-06-age-standardisation)