

La moyenne de votre échantillon n'est la moyenne de rien

32,8 % contre 29,1 % sur les mêmes 996 entretiens, et 62,8 % contre 69,8 % sur les mêmes ménages. L'écart, c'est le plan de sondage, et ce n'est pas du bruit.

Pierre Robentz Cassion

Leçon 1 sur 8

Analyse d'enquêtes, échantillonnage et pondération — L'échantillon n'est pas la population

Ce que couvre cette leçon

- Deux nombres tirés d'un seul fichier
- D'où vient l'écart
- Pourquoi concevoir le plan ainsi
- Les estimations par strate n'ont pas besoin de pondérations
- Le sens n'est pas toujours le même
- Que faire lorsqu'il n'y a pas de base de sondage
- La suite

Deux nombres tirés d'un seul fichier — En Python

```
import pandas as pd

survey = pd.read_csv("household-survey-2025.v1.csv")
frame = pd.read_csv("household-survey-frame-2025.v1.csv")

unweighted = (survey["food_insecure"] == "true").mean()
print(f"unweighted: {unweighted:.1%}")
```

Deux nombres tirés d'un seul fichier — En R

```
library(dplyr)
library(readr)

survey <- read_csv("household-survey-2025.v1.csv")
frame  <- read_csv("household-survey-frame-2025.v1.csv")

mean(survey$food_insecure == "true")
```

D'où vient l'écart

Strate	Ménages dans la base	Zones tirées	Ménages enquêtés	Insécurité alimentaire
Urbain	21 270	25	316	14,6 %
Rural accessible	28 958	25	337	36,2 %
Rural reculé	6 200	25	343	46,4 %

D'où vient l'écart — En Python

```
by_stratum = (  
    survey.assign(insecure=survey["food_insecure"] == "true")  
    .groupby("stratum")  
    .agg(interviews=("insecure", "size"), rate=("insecure", "mean"))  
)  
frame_totals = frame.groupby("stratum")["households"].sum()  
  
comparison = by_stratum.join(frame_totals.rename("frame_households"))  
comparison["sample_share"] = comparison["interviews"] / comparison["interviews"].sum()  
comparison["population_share"] = (  
    comparison["frame_households"] / comparison["frame_households"].sum()  
)  
print(comparison.round(3))
```

```
survey |>
  summarise(interviews = n(), rate = mean(food_insecure == "true"), .by = stratum) |>
  left_join(summarise(frame, frame_households = sum(households), .by = stratum),
            by = "stratum") |>
  mutate(sample_share = interviews / sum(interviews),
         population_share = frame_households / sum(frame_households))
```

D'où vient l'écart

Strate	Part de l'échantillon	Part de la population
Urbain	31,7 %	37,7 %
Rural accessible	33,8 %	51,3 %
Rural reculé	34,4 %	11,0 %

D'où vient l'écart

- Lorsque ces deux colonnes diffèrent, une estimation non pondérée est biaisée — et l'ampleur comme le sens du biais se. . .

Pourquoi concevoir le plan ainsi

- **L'allocation proportionnelle** donne la meilleure estimation nationale et des estimations infranationales faibles.
- **L'allocation égale** donne des estimations infranationales comparables et exige une pondération pour tout chiffre. . .

Les estimations par strate n'ont pas besoin de pondérations — En Python

```
print(by_stratum["rate"].round(3))
```

Les estimations par strate n'ont pas besoin de pondérations — En R

```
survey |> summarise(rate = mean(food_insecure == "true"), .by = stratum)
```

Le sens n'est pas toujours le même — En Python

```
for outcome in ["food_insecure", "improved_water_source"]:  
    print(f"{outcome:24} unweighted {(survey[outcome] == 'true').mean():.1%}")
```

Le sens n'est pas toujours le même — En R

```
survey |>  
  summarise(across(c(food_insecure, improved_water_source), ~ mean(.x == "true")))
```

Le sens n'est pas toujours le même

Résultat	Non pondéré	Pondéré
Insécurité alimentaire	32,8 %	29,1 %
Source d'eau améliorée	62,8 %	69,8 %

Le sens n'est pas toujours le même

- On ne peut donc pas corriger un chiffre non pondéré à vue de nez — et un rapport qui en présente un avec la mention «...»

Que faire lorsqu'il n'y a pas de base de sondage

- **Rapporter par strate seulement**, si les strates sont enregistrées. Les chiffres infranationaux sont d'ailleurs. . .
- **Reconstituer des pondérations approximatives** à partir d'une source de population externe, en documentant qu'elles. . .
- **Dire que l'enquête ne permet pas d'estimation de population.** C'est une vraie réponse, meilleure qu'un nombre qui. . .

- Les pondérations sont calculables, à partir de la base que ce jeu de données livre.

Lire la leçon complète, avec le code exécutable

[https://data-analysis.cassion.dev/fr/cours/survey-analysis-sampling/
survey-01-why-weights](https://data-analysis.cassion.dev/fr/cours/survey-analysis-sampling/survey-01-why-weights)