

The mean of your sample is not the mean of anything

32.8% against 29.1% on the same 996 interviews, and 62.8% against 69.8% on the same households. The gap is the design, and it is not noise.

Pierre Robentz Cassion

Lesson 1 of 8

Survey Analysis, Sampling and Weighting — The sample is not the population

What this lesson covers

- Two numbers from one file
- Where the gap comes from
- Why anyone would design it that way
- The stratum estimates need no weights
- The direction is not always the same
- What to do when there is no frame
- What comes next

Two numbers from one file — In 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%}")
```

Two numbers from one file — In 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")
```

Where the gap comes from

Stratum	Households in the frame	Areas sampled	Households interviewed	Food insecure
Urban	21,270	25	316	14.6%
Rural accessible	28,958	25	337	36.2%
Rural remote	6,200	25	343	46.4%

Where the gap comes from — In 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))
```

Where the gap comes from — In R

```
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))
```

Where the gap comes from

Stratum	Sample share	Population share
Urban	31.7%	37.7%
Rural accessible	33.8%	51.3%
Rural remote	34.4%	11.0%

Where the gap comes from

- When those two columns differ, an unweighted estimate is biased — and the size and direction of the bias are entirely...

Why anyone would design it that way

- **Proportional allocation** gives the best national estimate and weak sub-national ones.
- **Equal allocation** gives comparable sub-national estimates and requires weighting for anything national.

The stratum estimates need no weights — In Python

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

The stratum estimates need no weights — In R

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

The direction is not always the same — In Python

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

The direction is not always the same — In R

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

The direction is not always the same

Outcome	Unweighted	Weighted
Food insecure	32.8%	29.1%
Improved water source	62.8%	69.8%

The direction is not always the same

- So you cannot correct an unweighted figure by rule of thumb — and a report that presents one with a note saying. . .

What to do when there is no frame

- **Report by stratum only**, if the strata are recorded. Sub-national figures are usually what the audience wanted. . .
- **Reconstruct approximate weights** from an external population source, and document that they are approximate.
- **Say the survey cannot support a population estimate.** This is a real answer and it is better than a number that will. . .

What comes next

- The weights are computable, from the frame this dataset ships.

Read the full lesson, with runnable code

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