

Les ménages que vous n'avez pas eus

Cinquante-quatre entretiens manquants concentrés dans la strate la moins semblable aux autres, deux grappes de remplacement qui changent une probabilité que personne ne recalcule, et l'hypothèse sur laquelle repose tout ajustement.

Pierre Robentz Cassion

Leçon 6 sur 8

Analyse d'enquêtes, échantillonnage et pondération — Estimer correctement

Ce que couvre cette leçon

- La non-réponse n'est pas une donnée manquante
- L'ajustement, et l'hypothèse qui le porte
- Ajustez au bon niveau
- Les grappes de remplacement
- Les taux de réponse ont leur place dans le rapport
- Quand la réponse est assez mauvaise pour s'arrêter
- La suite

La non-réponse n'est pas une donnée manquante — En Python

```
selected = frame[frame["selected"] == True]

response = selected.groupby("stratum").agg(
    selected=("households_selected", "sum"),
    interviewed=("households_interviewed", "sum"),
)
response["rate"] = response["interviewed"] / response["selected"]
print(response)
```

La non-réponse n'est pas une donnée manquante — En R

```
frame |>
  filter(selected) |>
  summarise(selected = sum(households_selected),
            interviewed = sum(households_interviewed), .by = stratum) |>
  mutate(rate = interviewed / selected)
```

La non-réponse n'est pas une donnée manquante

Strate	Tirés	Enquêtés	Réponse
Urbain	350	316	90,3 %
Rural accessible	350	337	96,3 %
Rural reculé	350	343	98,0 %

L'ajustement, et l'hypothèse qui le porte — En Python

```
survey["nr_adjustment"] = 14 / survey["ea_id"].map(  
    selected.set_index("ea_id")["households_interviewed"]  
)  
print(survey.groupby("stratum")["nr_adjustment"].mean().round(3))
```

L'ajustement, et l'hypothèse qui le porte — En R

```
survey |> summarise(mean_adjustment = mean(14 / households_interviewed), .by = stratum)
```

L'ajustement, et l'hypothèse qui le porte

- L'hypothèse est que les non-répondants ressemblent aux répondants de la même zone — Elle n'est pas gratuite et n'est. . .

L'ajustement, et l'hypothèse qui le porte

La non-réponse a été ajustée à l'intérieur de chaque zone de dénombrement. La réponse urbaine était de 90,3 % contre 96 à 98 % en rural. Si les non-répondants urbains sont systématiquement mieux lotis que les répondants urbains — les concessions fermées et l'absence en journée le suggèrent — l'estimation ajustée reste légèrement trop élevée.

- **Dans la zone.** Ce que fait cette enquête. Le meilleur quand la réponse varie entre zones, ce qui est toujours le cas.
- **Dans la strate.** Plus grossier, et cela jette l'information selon laquelle une zone a eu quatre refus et sa voisine. . .
- **Dans une classe de réponse.** Regroupez les zones selon un facteur prédictif de la réponse — densité urbaine, jour de. . .

```
weights_by_area = survey.groupby("ea_id")["weight"].first().sort_values()
print(weights_by_area.tail(3))
print(f"largest weight / smallest: {weights_by_area.max() / weights_by_area.min():.1f}x")
```

```
survey |>  
  summarise(weight = first(weight), n = n(), .by = ea_id) |>  
  arrange(desc(weight)) |>  
  head(3)
```

- Écrêtez ou signalez toute pondération très éloignée des autres — Une pondération valant trois fois la base de sa strate. . .

Les grappes de remplacement — En Python

```
replacements = frame[frame["replacement_for"].notna()]  
print(replacements[["ea_id", "stratum", "households", "replacement_for"]])
```

Les grappes de remplacement — En R

```
frame |> filter(!is.na(replacement_for)) |>  
  select(ea_id, stratum, households, replacement_for)
```

Les grappes de remplacement

- Traiter le remplacement comme l'original — Le choix pragmatique, et ce que fait presque tout le monde
- Traiter l'original comme une non-réponse au niveau de la grappe — Plus honnête et plus difficile : la strate compte en...

Les grappes de remplacement — En Python

```
remote_areas = (frame["stratum"] == "rural-remote") & (frame["selected"] == True)
responding = int(remote_areas.sum()) - len(replacements)
print(f"25 selected, {responding} reached, cluster-level adjustment "
      f"{25 / responding:.3f}")
```

Les grappes de remplacement — En R

```
c(selected = 25, reached = 23, adjustment = 25 / 23)
```

Signalez le remplacement même si vous le traitez comme échangeable. « Deux des 25 grappes du rural reculé ont été remplacées pour cause d'inaccessibilité ; les deux remplaçantes étaient accessibles, de sorte que l'estimation du rural reculé est probablement conservatrice » est une phrase dont le lecteur a besoin et qu'il ne peut pas reconstituer.

Les taux de réponse ont leur place dans le rapport — En Python

```
report = response.reset_index()
report["non_response_bias_risk"] = ["high", "low", "low"]
report["adjustment"] = "within enumeration area"
print(report)
```

Les taux de réponse ont leur place dans le rapport — En R

```
tibble::tribble(  
  ~stratum,          ~selected, ~interviewed, ~rate, ~risk,  
  "urban",           350,          316, 0.903, "high",  
  "rural-accessible", 350,          337, 0.963, "low",  
  "rural-remote",    350,          343, 0.980, "low"  
)
```

Quand la réponse est assez mauvaise pour s'arrêter

- **Au-dessus de 90 %** — ajustez et rapportez. Le biais résiduel est en général faible devant l'erreur d'échantillonnage.
- **De 80 à 90 %** — ajustez, rapportez, et décrivez le sens probable du biais résiduel. Ne comparez pas à un tour. . .
- **En dessous de 80 %** — l'ajustement porte trop de charge. Rapportez l'estimation avec une réserve bien visible, ou. . .
- **En dessous de 60 % dans une strate** — l'estimation de cette strate n'est pas utilisable, et le dire est le constat. . .

- Vous avez une estimation, un plan et un compte rendu honnête de ce qui lui manque.

Lire la leçon complète, avec le code exécutable

[https://data-analysis.cassion.dev/fr/cours/survey-analysis-sampling/
survey-06-non-response-and-replacement](https://data-analysis.cassion.dev/fr/cours/survey-analysis-sampling/survey-06-non-response-and-replacement)