

Choisir les six formations sanitaires à visiter

Trois stratégies d'échantillonnage qui répondent à trois questions différentes, combien d'unités suffisent, et la phrase que vous avez le droit d'écrire sur les trente-deux formations non visitées.

Pierre Robentz Cassion

Leçon 4 sur 8

Évaluation de la qualité des données — Confronter à la source

Ce que couvre cette leçon

- Vous avez une semaine et trente-huit formations sanitaires
- Aléatoire — la seule qui se généralise
- Fondée sur le risque — là où sont les problèmes
- Recensement des plus grosses
- Combien suffisent
- LQAS — conçu exactement pour cela
- Combinez, et étiquetez
- La suite

Vous avez une semaine et trente-huit formations sanitaires

Stratégie	Répond à	Ne peut pas répondre à
Aléatoire	« Quelle est la qualité des données du district ? »	« Où sont les problèmes ? »
Fondée sur le risque	« Les formations suspectées le sont-elles vraiment ? »	« Quelle est la qualité du district ? »
Recensement des plus grosses	« Le total du district est-il juste ? »	Quoi que ce soit sur les petites formations

Aléatoire — la seule qui se généralise — En Python

```
import pandas as pd

facilities = vax[["facility_id", "facility_type"]].drop_duplicates()

sample = facilities.sample(n=6, random_state=20260728)
print(sample)
```

Aléatoire — la seule qui se généralise — En R

```
set.seed(20260728)
```

```
sample <- facilities |> dplyr::slice_sample(n = 6)
```

Aléatoire — la seule qui se généralise

- Fixez la graine et consignez-la — Une évaluation dont l'échantillon n'est pas reproductible appelle la question de...

Aléatoire — la seule qui se généralise — En Python

```
sample = (  
    facilities.groupby("facility_type", group_keys=False)  
    .apply(lambda g: g.sample(n=min(2, len(g)), random_state=20260728))  
)
```

```
sample <- facilities |>  
  group_by(facility_type) |>  
  slice_sample(n = 2) |>  
  ungroup()
```

Fondée sur le risque — là où sont les problèmes — En Python

```
risk = (  
    vax.groupby("facility_id")  
    .agg(reporting_rate=("reported", "mean"))  
    .assign(round_share=lambda d: d.index.map(round_number_share))  
)  
risk["score"] = (1 - risk["reporting_rate"]) + risk["round_share"]  
print(risk.sort_values("score", ascending=False).head(6))
```

```
risk <- vax |>  
  summarise(reporting_rate = mean(report_submitted), .by = facility_id) |>  
  left_join(round_shares, by = "facility_id") |>  
  mutate(score = (1 - reporting_rate) + round_share) |>  
  arrange(desc(score))
```

Six formations ont été retenues sur la base des signalements de l'analyse de bureau. Les constats les décrivent et ne sont pas représentatifs du district.

Recensement des plus grosses — En Python

```
volume = (  
    vax[vax["reported"]]  
    .groupby("facility_id")["doses_administered"].sum()  
    .sort_values(ascending=False)  
)  
share = volume.cumsum() / volume.sum()  
print(share.head(10).round(3))
```

Recensement des plus grosses — En R

```
volume <- vax |>
  filter(report_submitted) |>
  summarise(doses = sum(doses_administered), .by = facility_id) |>
  arrange(desc(doses)) |>
  mutate(cumulative = cumsum(doses) / sum(doses))
```

Combien suffisent — En Python

```
p = 0.30
n = 6
print(f"chance of finding none: {(1 - p) ** n:.1%}")
```

$(1 - 0.30)^6$

Visiter 12 formations. Si 3 ou plus ont un facteur de vérification hors de 0,95-1,05, le rapportage du district est classé comme nécessitant une action corrective.

Combinez, et étiquetez — En Python

```
plan = pd.DataFrame({  
    "facility_id": ["FAC012", "FAC033", "FAC004", "FAC019", "FAC007", "FAC021"],  
    "reason": ["random", "random", "risk: reporting 58%", "risk: round numbers",  
              "volume: top 5", "volume: top 5"],  
})
```

```
plan <- tibble::tribble(  
  ~facility_id, ~reason,  
  "FAC012", "random",  
  "FAC033", "random",  
  "FAC004", "risk: reporting 58%",  
  "FAC019", "risk: round numbers",  
  "FAC007", "volume: top 5",  
  "FAC021", "volume: top 5"  
)
```

- La colonne `reason` est la section méthodologie — Avec elle, un lecteur sait quels constats se généralisent et lesquels. . .

- L'échantillonnage fondé sur le risque exige un classement de risque, et vous n'avez pour l'instant que des taux de reportage.

Lire la leçon complète, avec le code exécutable

[https://data-analysis.cassion.dev/fr/cours/data-quality-assessment/
dqa-04-sampling-for-verification](https://data-analysis.cassion.dev/fr/cours/data-quality-assessment/dqa-04-sampling-for-verification)