

Choosing which six facilities to visit

Three sampling strategies that answer three different questions, how many is enough, and the sentence you are allowed to write about the thirty-two facilities you did not visit.

Pierre Robentz Cassion

Lesson 4 of 8

Data Quality Assessment — Checking against the source

What this lesson covers

- You have a week and thirty-eight facilities
- Random: the only one that generalises
- Risk-based: where the problems are
- Census of the largest
- How many is enough
- LQAS: designed for exactly this
- Combine, and label
- What comes next

You have a week and thirty-eight facilities

Strategy	Answers	Cannot answer
Random	"What is the district's data quality?"	"Where are the problems?"
Risk-based	"Are the facilities we suspect actually wrong?"	"What is the district's data quality?"
Census of the largest	"Is the district total right?"	Anything about small facilities

Random: the only one that generalises — In Python

```
import pandas as pd

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

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

Random: the only one that generalises — In R

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

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

Random: the only one that generalises

- Seed it and record the seed — A DQA whose sample cannot be reproduced invites the question of whether the facilities. . .

Random: the only one that generalises — In Python

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

Random: the only one that generalises — In R

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

Risk-based: where the problems are — In 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-based: where the problems are — In R

```
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))
```

Risk-based: where the problems are

Six facilities were selected on the basis of desk-review flags. Findings describe those facilities and are not representative of the district.

Census of the largest — In 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))
```

Census of the largest — In R

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

How many is enough — In Python

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

How many is enough — In R

$$(1 - 0.30)^6$$

LQAS: designed for exactly this

Visit 12 facilities. If 3 or more have a verification factor outside 0.95–1.05, the district's reporting is classified as needing corrective action.

Combine, and label — In 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"],  
})
```

Combine, and label — In R

```
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"  
)
```

Combine, and label

- The reason column is the methods section — With it, a reader knows which findings generalise and which do not

What comes next

- Risk-based sampling needs a risk ranking, and so far you have only reporting rates.

Read the full lesson, with runnable code

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