

Le recomptage, et le nombre qu'il produit

Un facteur de vérification est le recomptage divisé par le chiffre rapporté. Le district est à 1,012, ce que personne ne contesterait, et une école sur vingt-quatre est à 1,42.

Pierre Robentz Cassion

Leçon 3 sur 8

Évaluation de la qualité des données — Confronter à la source

Ce que couvre cette leçon

- L'exactitude est la dimension qui exige un déplacement
- Le facteur de vérification
- Calculez-le sur un registre dont vous disposez
- Le chiffre du district est le piège
- La bande de tolérance
- Le recomptage est un protocole, pas une après-midi
- Ce qu'un facteur de vérification ne peut pas dire
- La suite

L'exactitude est la dimension qui exige un déplacement

- Les quatre autres dimensions sortent de l'extraction.

Le facteur de vérification — Exemple

`verification factor = recount / reported`

- Les deux sens sont des constats, et ce ne sont pas les mêmes
- **Facteur supérieur à 1** — le registre contient plus que ce qui a été rapporté.
Sous-déclaration. Généralement une. . .
- **Facteur inférieur à 1** — le rapport revendique plus que le registre ne soutient.
Surdéclaration. Parfois un double. . .

Calculez-le sur un registre dont vous disposez — En Python (suite)

```
import pandas as pd

attendance = pd.read_csv("school-attendance-2024.v1.csv")
roster = pd.read_csv("school-roster-2024.v1.csv")

# The two students on the roster twice are ambiguous; exclude them from the
# verification and say so, rather than assigning them arbitrarily.
duplicated = roster["student_id"].duplicated(keep=False)
clean_roster = roster[~duplicated]

marks = attendance.merge(
    clean_roster[["student_id", "school_id"]], on="student_id",
    how="inner", validate="many_to_one",
)

vf = (
```

Calculez-le sur un registre dont vous disposez — En Python (suite)

```
marks.assign(  
    reported=marks["present"] == "true",  
    recount=marks["present"].isin(["true", "Y"]),  
)  
marks.groupby("school_id")[["recount", "reported"]]  
    .sum()  
)  
vf["verification_factor"] = vf["recount"] / vf["reported"]  
print(vf.sort_values("verification_factor", ascending=False).head())
```

Calculez-le sur un registre dont vous disposez — En R

```
library(dplyr)

clean_roster <- roster |>
  group_by(student_id) |> filter(n() == 1) |> ungroup()

vf <- attendance |>
  inner_join(select(clean_roster, student_id, school_id), by = "student_id",
             relationship = "many-to-one") |>
  summarise(
    recount = sum(present %in% c("true", "Y")),
    reported = sum(present == "true"),
    .by = school_id
  ) |>
  mutate(verification_factor = recount / reported) |>
  arrange(desc(verification_factor))
```

Calculez-le sur un registre dont vous disposez

École	Recomptage	Rapporté	Facteur
SCH09	2 456	1 734	1,416
SCH01	2 945	2 945	1,000
SCH02	2 569	2 569	1,000
... 22 autres			1,000
District	61 764	61 042	1,012

Le chiffre du district est le piège — En Python

```
print(vf["verification_factor"].describe())  
outliers = vf[(vf["verification_factor"] - 1).abs() > 0.05]  
print(f"{len(outliers)} of {len(vf)} schools outside the tolerance band")
```

Le chiffre du district est le piège — En R

```
summary(vf$verification_factor)
vf |> filter(abs(verification_factor - 1) > 0.05)
```

- **Élargissez-la pour les petits nombres.** Une formation rapportant 12 cas a un facteur de 1,08 si un cas a été manqué...
- **Resserrez-la quand il y a de l'argent.** Lorsque le chiffre déclenche un paiement — financement basé sur les. . .

La bande de tolérance — En Python

```
TOLERANCE = 0.05
MIN_ABSOLUTE = 5

vf["difference"] = vf["recount"] - vf["reported"]
vf["finding"] = (
    ((vf["verification_factor"] - 1).abs() > TOLERANCE)
    & (vf["difference"].abs() > MIN_ABSOLUTE)
)
print(vf[vf["finding"]])
```

```
vf <- vf |>
  mutate(
    difference = recount - reported,
    finding = abs(verification_factor - 1) > 0.05 & abs(difference) > 5
  )
```

- Inscrivez la bande dans le protocole avant la visite — Un seuil choisi après avoir vu les chiffres n'est pas un seuil,...

Le recomptage est un protocole, pas une après-midi

- **Fixez la période exactement.** « Mars » n'est pas une période. « Du 1er au 31 mars, par date de service et non par... »
- **Fixez la définition exactement.** Comptez ce que dit le numérateur de l'indicateur, non ce que dit l'en-tête de... »
- **Recomptez indépendamment.** La personne qui a compilé le rapport ne doit pas faire le recomptage. Non par défiance —... »
- **Notez à partir de quoi vous avez compté.** Registre, fiche de pointage, cartes individuelles. Deux sources d'une même... »

Ce qu'un facteur de vérification ne peut pas dire

- Il ne dit pas que le registre est juste — Un facteur de 1,00 signifie que le rapport correspond au registre
- Il ne se généralise pas au-delà des formations visitées — sauf si vous avez échantillonné correctement
- Ce n'est pas une accusation — Un facteur de 1,42 dans une école, avec vingt-trois autres à exactement 1,00, a la...

- Vous ne pouvez pas visiter trente-huit formations.

Lire la leçon complète, avec le code exécutable

```
https://data-analysis.cassion.dev/fr/cours/data-quality-assessment/  
dqa-03-verification-factor
```