

The recount, and the number it produces

A verification factor is the recount over the reported figure. The district sits at 1.012, which nobody would question, and one school in twenty-four sits at 1.42.

Pierre Robentz Cassion

Lesson 3 of 8

Data Quality Assessment — Checking against the source

What this lesson covers

- Accuracy is the dimension that needs a visit
- The verification factor
- Compute it on a register you have
- The district figure is the trap
- The tolerance band
- The recount is a protocol, not an afternoon
- What a VF cannot tell you
- What comes next

Accuracy is the dimension that needs a visit

- The other four dimensions come out of the extract.

The verification factor — Example

`verification factor = recount / reported`

The verification factor

- Both directions are findings, and they are not the same finding
- **VF above 1** — the register holds more than was reported. Under-reporting. Usually a transcription or aggregation. . .
- **VF below 1** — the report claims more than the register supports. Over-reporting. Sometimes double-counting or a . . .

Compute it on a register you have — In Python (cont.)

```
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 = (
```

Compute it on a register you have — In Python (cont.)

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

Compute it on a register you have — In 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))
```

Compute it on a register you have

School	Recount	Reported	VF
SCH09	2,456	1,734	1.416
SCH01	2,945	2,945	1.000
SCH02	2,569	2,569	1.000
... 22 more			1.000
District	61,764	61,042	1.012

The district figure is the trap — In 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")
```

The district figure is the trap — In R

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

The tolerance band

- **Widen it for small numbers.** A facility reporting 12 cases has a VF of 1.08 if one case was missed. On small counts,...
- **Narrow it for money.** Where the figure drives a payment — results-based financing, a per-beneficiary reimbursement —...

The tolerance band — In 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"]])
```

The tolerance band — In R

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

The tolerance band

- Write the band into the protocol before the visit — A threshold chosen after seeing the numbers is not a threshold, and...

The recount is a protocol, not an afternoon

- **Fix the period exactly.** "March" is not a period. "1 to 31 March, by date of service, not date of entry" is.
- **Fix the definition exactly.** Count what the indicator's numerator says, not what the register's column heading says. . . .
- **Recount independently.** The person who compiled the report should not do the recount. Not because of dishonesty —. . .
- **Record what you counted from.** Register, tally sheet, individual cards. Two sources in one facility often disagree, . . .

What a VF cannot tell you

- It cannot tell you the register is right — A VF of 1.00 means the report matches the register
- It does not generalise from the facilities you visited — unless you sampled properly
- It is not an accusation — A VF of 1.42 in one school, with twenty-three others at exactly 1.00, is the shape of a . . .

What comes next

- You cannot visit thirty-eight facilities.

Read the full lesson, with runnable code

[https://data-analysis.cassion.dev/courses/data-quality-assessment/
dqa-03-verification-factor](https://data-analysis.cassion.dev/courses/data-quality-assessment/dqa-03-verification-factor)