

The cases that have not closed are the long ones

Median time to closure is four months on the 636 cases that closed. 472 more are still open, 140 of them already past four months, and child protection is 58% censored against general protection's 32% — so the same statistic means different things in different columns.

Pierre Robentz Cassion

Lesson 6 of 8

Protection and GBV Data — The caseload

What this lesson covers

- The obvious calculation is biased
- Fix the impossible durations first
- Three honest ways to report it
- Censoring is not evenly spread
- The closure reason is a judgement
- Report it whole
- What comes next

The obvious calculation is biased — In Python

```
import pandas as pd

cases = pd.read_csv("protection-case-management-2024.v1.csv")

opened = cases["opened_month"].str[5:].astype(int)
closed = cases["closed_month"].str[5:].astype("Int64")
duration = closed - opened

print(f"closed cases: {duration.notna().sum()}")
print(f"median months to closure: {duration.median()}")
print(f"still open: {duration.isna().sum()} ({duration.isna().mean():.1%})")
```

The obvious calculation is biased — In R

```
library(dplyr)

cases |>
  mutate(opened = as.integer(substr(opened_month, 6, 7)),
         closed = as.integer(substr(closed_month, 6, 7)),
         duration = closed - opened) |>
  summarise(closed = sum(!is.na(duration)),
            median = median(duration, na.rm = TRUE),
            open = sum(is.na(duration)))
```

The obvious calculation is biased

- Four months, on 636 of 1,108 cases — The other 472 are still open at the December cut-off, and they are not missing —...

The obvious calculation is biased — In Python

```
still_open = 12 - opened[duration.isna()] + 1  
print(f"open cases already past 4 months: {(still_open > 4).sum()}")  
print(f"longest open: {still_open.max()} months and counting")
```

The obvious calculation is biased — In R

```
cases |> filter(is.na(closed_month)) |>  
  mutate(so_far = 12 - as.integer(substr(opened_month, 6, 7)) + 1) |>  
  summarise(past_four = sum(so_far > 4), longest = max(so_far))
```

The obvious calculation is biased

- 140 of the open cases have already been open longer than the median closed case
— and one has been open eleven months

Fix the impossible durations first — In Python

```
impossible = duration.notna() & (duration <= 0)
print(f"cases closing on or before the month they opened: {impossible.sum()}")
print(cases.loc[impossible, ["opened_month", "closed_month"]])
```

Fix the impossible durations first — In R

```
cases |> filter(!is.na(closed_month),  
               closed_month <= opened_month) |> nrow()
```

Fix the impossible durations first

- Seven cases close in a month earlier than they opened — The contradiction is invisible in either column alone and...

Three honest ways to report it

- Report the median with the censoring stated — The cheapest correct option

Three honest ways to report it — Example

Median time to closure: 4 months (636 closed cases).

472 cases (42.6%) were still open at the cut-off and are excluded; 140 of them have already been open longer than 4 months, so the true median is higher.

Three honest ways to report it

- Report a completion-by-month curve — What share of cases opened in month m had closed within k months, computed...

Three honest ways to report it — In Python

```
def closed_within(k):
    eligible = cases[opened <= 12 - k]
    dur = (eligible["closed_month"].str[5:].astype("Int64")
           - eligible["opened_month"].str[5:].astype(int))
    return (dur <= k).sum() / len(eligible), len(eligible)

for k in (3, 6, 9):
    share, n = closed_within(k)
    print(f"closed within {k} months: {share:.1%} (n={n})")
```

Three honest ways to report it — In R

For each k, restrict to cases with k months of possible follow-up.

Three honest ways to report it

- Report time to closure only for a cohort with full follow-up — Cases opened in January have eleven months of. . .
- All three are defensible and the first is not enough on its own — A median with no censoring statement beside it is the. . .

Censoring is not evenly spread — In Python

```
by_category = pd.DataFrame({
    "closed": duration.notna().groupby(cases["case_category"]).sum(),
    "open": duration.isna().groupby(cases["case_category"]).sum(),
})
by_category["censored"] = by_category["open"] / by_category.sum(axis=1)
print(by_category.round(3))
```

Censoring is not evenly spread — In R

```
cases |> summarise(closed = sum(!is.na(closed_month)),  
                  open = sum(is.na(closed_month)), .by = case_category) |>  
mutate(censored = open / (open + closed))
```

Censoring is not evenly spread

Category	Closed	Open	Censored	Median of the closed
Child protection	131	179	58%	5 months
GBV	255	173	40%	4 months
General protection	250	120	32%	4 months

Censoring is not evenly spread

- Child protection is 58% censored and general protection 32% — So the two medians in that table are computed on very. . .
- Uneven censoring makes a comparison of medians misleading in a specific direction — the group with more censoring is. . .

The closure reason is a judgement — In Python

```
print(cases["closure_reason"].value_counts(normalize=True).round(3))
```

The closure reason is a judgement — In R

```
cases |> filter(!is.na(closure_reason)) |> count(closure_reason) |>  
  mutate(share = n / sum(n))
```

The closure reason is a judgement

Reason	Share of closures
Case plan objectives met	29.1%
Lost contact	27.2%
Survivor withdrew	13.1%
Closed administratively	11.3%
Relocated	9.7%
Transferred to another agency	9.6%

The closure reason is a judgement

- Only 29.1% of closures are a case plan completed — That is the headline, and it is the number a supervision. . .

The closure reason is a judgement — In Python

```
by_area = pd.crosstab(cases["admin2"], cases["closure_reason"], normalize="index")  
print((by_area[["closed-administratively", "lost-contact"]] * 100).round(1))
```

The closure reason is a judgement — In R

```
cases |> filter(!is.na(closure_reason)) |>  
  count(admin2, closure_reason) |> mutate(share = n / sum(n), .by = admin2)
```

The closure reason is a judgement

- Hinche files 36.2% of its closures as closed-administratively against 3.8% to 9.8% everywhere else — and its...
- A closure reason is a caseworker's judgement, not an observation — So a distribution that differs sharply between...

Report it whole — Example (cont.)

Case closure, 1,108 cases

Closed by the cut-off	636	57.4%	
Still open	472	42.6%	censored, not missing
Median months to closure	4		of closed cases only
open cases already past 4 months	140		so the true median is higher

Censoring by category: child protection 58%, GBV 40%, general 32%.

Medians are not comparable across those columns.

Closure reasons (of 636 closures)

Case plan objectives met	29.1%	
Lost contact	27.2%	
Survivor withdrew	13.1%	
Closed administratively	11.3%	36.2% in Hinche, 3.8-9.8% elsewhere
Relocated	9.7%	

Report it whole — Example (cont.)

Transferred

9.6%

7 cases record a closing month before their opening month and are excluded.

What comes next

- Everything in this unit is about cases already in the system.

Read the full lesson, with runnable code

[https://data-analysis.cassion.dev/courses/protection-and-gbv-data/
prot-06-closure-and-censoring](https://data-analysis.cassion.dev/courses/protection-and-gbv-data/prot-06-closure-and-censoring)