

Dix-neuf cellules à un

Croisez zone, catégorie, tranche d'âge et sexe et ce jeu de données anonyme produit 200 cellules, dont 85 contiennent quatre cas ou moins et 19 en contiennent exactement un. Dans un jeu de données VBG, une cellule à un est une personne, et le tableau ne peut pas sortir des murs.

Pierre Robentz Cassion

Leçon 2 sur 8

Données de protection et VBG — Ce qu'il ne faut pas collecter

Ce que couvre cette leçon

- Le jeu de données est anonyme et le tableau ne l'est pas
- Pourquoi quatre et non un
- Une règle, écrite comme du code
- Que faire au lieu de publier la cellule
- Le constat d'équité survit à la règle
- La demande qu'il faut refuser
- La suite

Le jeu de données est anonyme et le tableau ne l'est pas — En Python

```
import pandas as pd

referrals = pd.read_csv("protection-referrals-2024.v1.csv")

cells = referrals.groupby(
    ["admin2", "case_category", "age_band", "sex"], dropna=False
).size()

print(f"{len(cells)} cells")
print(f"  holding 1 case: {(cells == 1).sum()}")
print(f"  holding 4 or fewer: {(cells <= 4).sum()}")
```

Le jeu de données est anonyme et le tableau ne l'est pas — En R

```
library(dplyr)
```

```
referrals |>
```

```
  count(admin2, case_category, age_band, sex) |>
```

```
  summarise(cells = n(), of_one = sum(n == 1), of_four_or_fewer = sum(n <= 4))
```

Le jeu de données est anonyme et le tableau ne l'est pas

- 200 cellules, dont 85 avec quatre cas ou moins, et 19 avec exactement un

Pourquoi quatre et non un

- Une cellule à deux identifie les deux personnes l'une à l'autre — et à quiconque en connaît une
- La différenciation récupère les cellules supprimées — Si un total de ligne est publié et que toutes les cellules de...

Pourquoi quatre et non un — En Python

```
by_area = referrals.groupby(["admin2", "case_category"]).size().unstack()
print(by_area)
print("\nIf one cell is suppressed and the row total is printed,")
print("the suppressed cell is the total minus the others.")
```

Pourquoi quatre et non un — En R

```
referrals |> count(admin2, case_category) |>  
  tidyr::pivot_wider(names_from = case_category, values_from = n)
```

Pourquoi quatre et non un

- Une règle de suppression exige donc un seuil et une suppression secondaire — supprimer les petites cellules, puis en. . .

Une règle, écrite comme du code — En Python (suite)

```
THRESHOLD = 5
```

```
def suppress(counts, threshold=THRESHOLD):  
    """Primary suppression, then a secondary pass so rows cannot be differenced."""  
    table = counts.copy()  
    small = table < threshold  
    table = table.mask(small, other=pd.NA)  
  
    # Secondary: if a row has exactly one suppressed cell, the row total gives  
    # it away. Suppress the next smallest cell in that row as well.  
    for index, row in table.iterrows():  
        if row.isna().sum() == 1:  
            remaining = counts.loc[index][~small.loc[index]]  
            if len(remaining):  
                table.loc[index, remaining.idxmin()] = pd.NA  
    return table
```

Une règle, écrite comme du code — En Python (suite)

```
print(suppress(by_area))
```

Une règle, écrite comme du code — En R

```
suppress <- function(x, threshold = 5) ifelse(x < threshold, NA, x)  
# Plus the secondary pass: a row with exactly one NA is not suppressed at all.
```

Une règle, écrite comme du code

- Écrivez la règle comme une fonction et appliquez-la à chaque tableau — Une règle appliquée à l'œil est appliquée de...

Que faire au lieu de publier la cellule

- Agréger vers le haut — Retirez une dimension

Que faire au lieu de publier la cellule — En Python

```
coarse = referrals.groupby(["admin2", "case_category"]).size()
print(f"cells under 5: {(coarse < 5).sum()} of {len(coarse)}")
```

Que faire au lieu de publier la cellule — En R

```
referrals |> count(admin2, case_category) |> summarise(small = sum(n < 5))
```

Que faire au lieu de publier la cellule

- Rapporter le taux, non l'effectif, sur un dénominateur assez grand pour le porter — « L'aboutissement est de 31 % pour... »
- Répondre à la question plutôt que fournir le tableau — Un demandeur qui réclame zone par âge par sexe veut généralement...

Le constat d'équité survit à la règle — En Python

```
def completion(frame):
    consenting = frame[frame["consent_to_refer"]]
    reached = consenting[consenting["referral_accepted"] &
                        consenting["days_to_first_service"].notna()]
    return len(reached), len(consenting), len(reached) / len(consenting)

disability = referrals["disability_reported"].isin([True, "true", "Yes"])
print("reported:      %d/%d = %.1f%%" % completion(referrals[disability]))
print("not reported: %d/%d = %.1f%%" % completion(referrals[~disability]))
```

Le constat d'équité survit à la règle — En R

```
referrals |>
  filter(consent_to_refer) |>
  summarise(reached = sum(referral_accepted & !is.na(days_to_first_service)),
            n = n(), .by = disability_reported) |>
  mutate(completion = reached / n)
```

Le constat d'équité survit à la règle

- 26,7 % contre 46,2 %, sur des dénominateurs de 202 et 1 436 — L'écart d'équité de dix-neuf points — le constat le plus. . .
- Les analyses qui échouent au test de divulgation ne sont presque jamais celles qui comptent — Une désagrégation à. . .

La demande qu'il faut refuser — Exemple (suite)

Request: case counts by commune, month and incident category.

Response: declined.

Commune-by-month cells hold zero to three cases. At that resolution a count identifies individuals to anyone working in the area, and the dataset cannot be re-identified back but a person can be.

What I can provide instead:

- the same counts at district by quarter, no cell below five
- the completion rate and time to service for each commune, on annual denominators
- a written answer to the question the request is for, if you tell me what decision it feeds

If the requirement is genuinely commune-level operational planning, the right route is a data sharing agreement with the case management agency,

La demande qu'il faut refuser — Exemple (suite)

not an extract.

La demande qu'il faut refuser

- Proposez trois solutions de remplacement et demandez quelle décision la demande alimente — Un refus sans alternative. . .

- Les règles sont posées.

Lire la leçon complète, avec le code exécutable

[https://data-analysis.cassion.dev/fr/cours/protection-and-gbv-data/
prot-02-disclosure-risk](https://data-analysis.cassion.dev/fr/cours/protection-and-gbv-data/prot-02-disclosure-risk)