

Un taux de survie qu'il ne faut pas publier

Chaînez les taux de passage d'une année et 19,3 % des enfants atteignent la dernière année. Comptez les redoublants comme retenus et cela fait 40,3 %. Les deux viennent du même fichier, et aucun n'est un taux d'achèvement.

Pierre Robentz Cassion

Leçon 6 sur 8

Analyse de programmes d'éducation — La cohorte

Ce que couvre cette leçon

- Le calcul vers lequel tout le monde se tourne
- Pourquoi il est faux
- Que calculer à la place
- Qui part
- Les données manquantes ne le sont pas au hasard
- Rapportez-le comme un profil de risque
- La suite

Le calcul vers lequel tout le monde se tourne — En Python (suite)

```
import pandas as pd

enrolment = pd.read_csv("school-enrolment-2024.v1.csv")
clean = enrolment[enrolment["age_years"] <= 20]
cohort = clean[clean["school_year"] == 2023]

survival, cumulative = {}, 1.0
for grade in range(1, 7):
    grade_rows = cohort[cohort["grade"] == grade]
    promoted = grade_rows["end_of_year_status"].isin(
        ["promoted", "completed-final-grade"]).mean()
    survival[grade] = cumulative
    cumulative *= promoted
    print(f"grade {grade}: promotion {promoted:.1%}, "
          f"survival to here {survival[grade]:.1%}")
```

Le calcul vers lequel tout le monde se tourne — En Python (suite)

```
print(f"\nsurvival to grade 6 entry: {cumulative:.1%}")
```

Le calcul vers lequel tout le monde se tourne — En R

```
library(dplyr)

cohort |>
  filter(between(grade, 1, 6)) |>
  summarise(promotion = mean(end_of_year_status %in%
    c("promoted", "completed-final-grade")), .by = grade) |>
  arrange(grade) |>
  mutate(survival = cumprod(lag(promotion, default = 1)))
```

Le calcul vers lequel tout le monde se tourne

Année	Passage	Survie cumulée
1	89,3 %	100,0 %
2	83,0 %	89,3 %
3	71,7 %	74,1 %
4	76,2 %	53,1 %
5	70,9 %	40,5 %
6	67,3 %	28,7 %
—	—	19,3 %

Le calcul vers lequel tout le monde se tourne

- 19,3 % des enfants entrant en première année atteignent la fin du primaire —
C'est un nombre frappant, c'est ce que la...

- Les redoublants sont comptés comme des échecs — Un enfant qui redouble la deuxième année n'est pas déscolarisé — il est. . .

Pourquoi il est faux — En Python

```
promotion_or_repeat = cohort[cohort["grade"].between(1, 6)].groupby("grade")["end_of_year_status"].apply(lambda s: s.isin(["promoted", "completed-final-grade", "repeated"]).mean())

still_enrolled = 1.0
for grade, rate in promotion_or_repeat.items():
    still_enrolled *= rate
print(f"survival counting repeaters as retained: {still_enrolled:.1%}")
```

Pourquoi il est faux — En R

Retention, not promotion, is what a survival rate needs.

Pourquoi il est faux

- Les transferts sont eux aussi comptés comme des échecs — 2,9 % sont partis dans une autre école et la chaîne les lit. . .
- Une seule année est prise pour six — Le taux de passage de sixième observé en 2023 appartient à des enfants entrés en. . .
- Les effectifs par année ne forment pas une cohorte — La première année compte 149 élèves et la sixième 110, et une. . .

Que calculer à la place

- La rétention, non le passage — s'il faut employer une seule année

Que calculer à la place — En Python

```
def chained(statuses):
    rate, product = {}, 1.0
    for grade in range(1, 7):
        rows = cohort[cohort["grade"] == grade]["end_of_year_status"]
        product *= rows.isin(statuses).mean()
        rate[grade] = product
    return product

promotion_only = chained(["promoted", "completed-final-grade"])
retained = chained(["promoted", "completed-final-grade", "repeated"])
print(f"chained promotion: {promotion_only:.1%}")
print(f"chained retention: {retained:.1%}")
```

Que calculer à la place — En R

Same chain, different numerator. The gap is repetition compounding.

Que calculer à la place

- 19,3 % contre 40,3 % — Plus de vingt points de la perte apparente étaient du redoublement lu comme une sortie, et le...
- Suivez de vrais élèves quand vous le pouvez — Ce registre a deux années et les mêmes identifiants, si bien qu'une...

Que calculer à la place — En Python

```
years = clean.pivot_table(index="student_id", columns="school_year",
                           values="grade", aggfunc="first")

both = years.dropna()
print(f"students observed in both years: {len(both)}")
print(f"  advanced a grade: {(both[2024] > both[2023]).mean():.1%}")
print(f"  same grade:      {(both[2024] == both[2023]).mean():.1%}")
```

Que calculer à la place

- 1 103 élèves figurent dans les deux années : 88,5 % ont avancé d'une année et 11,5 % sont deux fois dans la même —...

Que calculer à la place — En R

```
enrolment |>  
  select(student_id, school_year, grade) |>  
  tidyr::pivot_wider(names_from = school_year, values_from = grade) |>  
  filter(!is.na(`2023`), !is.na(`2024`))
```

- Une transition d'un an observée sur de vrais élèves vaut mieux qu'une chaîne de six ans assemblée à partir des taux. . .

Qui part — En Python

```
year23 = cohort.assign(
    over_age=cohort["age_years"] > cohort["grade"] + 5,
    dropped=cohort["end_of_year_status"].eq("dropped-out"),
)
for column in ("over_age", "disability_reported", "displacement_status", "sex"):
    table = year23.groupby(column, dropna=False)["dropped"].agg(["mean", "size"])
    print((table * [100, 1]).round(1), "\n")
```

```
cohort |>  
  mutate(dropped = end_of_year_status == "dropped-out") |>  
  summarise(dropout = mean(dropped), n = n(), .by = displacement_status)
```

Ventilation	Abandon	n
En retard d'âge	19,5 %	478
À l'âge attendu	4,8 %	814
Retournés	19,7 %	71
Handicap signalé	17,9 %	112
Déplacés internes	14,7 %	163
Communauté d'accueil	11,6 %	95
Résidents	8,6 %	940
Garçons	10,4 %	634

...

- Le retard d'âge est le prédicteur le plus fort et le groupe le plus grand — Le handicap et le déplacement doublent à . . .
- Le sexe ne montre rien — 10,4 % contre 10,0 % sur 634 et 658 élèves

Les données manquantes ne le sont pas au hasard — En Python

```
print(year23["displacement_status"].isna().sum(), "students with no status")  
print(year23.loc[year23["displacement_status"].isna(), "dropped"].mean())
```

Les données manquantes ne le sont pas au hasard — En R

```
cohort |> filter(is.na(displacement_status)) |> nrow()
```

Les données manquantes ne le sont pas au hasard

- Soyez précis sur ce que cela biaise et ne biaise pas — Les vides sont retirés à peu près au hasard *à l'intérieur* de. . .
- le **taux d'abandon** des déplacés internes et des retournés n'est pas biaisé — 14,7 % et 19,7 % sont les bons nombres. . .
- leurs **effectifs et leur part de population** sont sous-estimés d'environ un dixième. 234 élèves sont enregistrés. . .

Rapportez-le comme un profil de risque — Exemple (suite)

Dropout, 2023 cohort, 1,292 students

Overall	10.2%	
Over-age for grade	19.5%	478 students
In-age	4.8%	814
Returnee	19.7%	71 students
Disability reported	17.9%	112
Internally displaced	14.7%	163
Resident	8.6%	940

No meaningful difference by sex: 10.4% boys, 10.0% girls.

42 students have no displacement status, all drawn from displaced and returnee households. The rates above are unbiased for those groups; their

Rapportez-le comme un profil de risque — Exemple (suite)

counts are understated by about a tenth.

Not reported: survival to the final grade. A reconstructed cohort from a single year gives 19.3% on promotion and 40.3% on retention, neither of which is a completion rate. A true cohort needs six years of the register or a household survey. The measured one-year transition is 88.5%.

Rapportez-le comme un profil de risque

- « Non rapporté, et pourquoi » est la ligne la plus utile de ce bloc — Quelqu'un demandera un taux d'achèvement, et la...

- L'inscription, la présence et la rétention disent si les enfants sont à l'école.

Lire la leçon complète, avec le code exécutable

[https://data-analysis.cassion.dev/fr/cours/education-programme-analysis/
edu-06-who-drops-out](https://data-analysis.cassion.dev/fr/cours/education-programme-analysis/edu-06-who-drops-out)