

## Ce que dit le dernier chiffre

Préférence de chiffres, attraction des valeurs rondes, et l'étalonnage qui sépare un constat du bruit. Les tailles d'une équipe se terminent par ,0 ou ,5 dans soixante-neuf pour cent des cas ; la formation la plus suspecte du district se révèle être du hasard.

---

Pierre Robentz Cassion

Leçon 6 sur 8

Évaluation de la qualité des données — Repérer les sites qui méritent une visite

## Ce que couvre cette leçon

- Les mesures ont une texture
- Le contrôle
- L'attraction des âges, et pourquoi ce n'est la faute de personne
- Étalonnez sur ce que produit le hasard
- Pourquoi le cas négatif compte davantage
- Ce que vous avez le droit d'écrire
- La suite

## Les mesures ont une texture

- Un nombre produit par la lecture d'un instrument a une texture particulière — son dernier chiffre est quasi uniforme, car la valeur vraie a autant de chances de tomber n'importe où dans la plus petite division de l'échelle.

```
import pandas as pd

smart = pd.read_csv("smart-nutrition-survey-2024.v1.csv")

smart["terminal"] = (smart["height_cm"] * 10).round() % 10
by_team = (
    smart.assign(rounded=smart["terminal"].isin([0, 5]))
    .groupby("team")
    .agg(rounded_share=("rounded", "mean"), n=("rounded", "size"))
)
print((by_team["rounded_share"] * 100).round(0))
```

```
library(dplyr)

smart |>
  mutate(terminal = round(height_cm * 10) %% 10,
         rounded = terminal %in% c(0, 5)) |>
  summarise(rounded_share = mean(rounded), n = n(), .by = team)
```

| Équipe   | Tailles finissant par ,0 ou ,5 | Enfants mesurés |
|----------|--------------------------------|-----------------|
| 1        | 18 %                           | 217             |
| <b>2</b> | <b>69 %</b>                    | <b>248</b>      |
| 3        | 18 %                           | 248             |
| 4        | 23 %                           | 217             |

- Ce n'est pas un résultat marginal et cela n'exige aucun test — Rien dans une population ne fait tomber les tailles des...

## L'attraction des âges, et pourquoi ce n'est la faute de personne — En Python

```
ages = smart["age_months"].dropna()
whole_years = (ages % 12 == 0).mean()

print(f"{whole_years:.0%} of ages fall on an exact multiple of 12 months")
print(ages.value_counts().reindex([22, 23, 24, 25, 26]))
```

## L'attraction des âges, et pourquoi ce n'est la faute de personne — En R

```
smart |>  
  filter(!is.na(age_months)) |>  
  summarise(whole_years = mean(age_months %% 12 == 0))  
  
smart |> count(age_months) |> filter(age_months %in% 22:26)
```

## L'attraction des âges, et pourquoi ce n'est la faute de personne

| Âge en mois | 22 | 23 | 24 | 25 | 26 |
|-------------|----|----|----|----|----|
| Enfants     | 21 | 15 | 66 | 19 | 17 |

## Étalonnez sur ce que produit le hasard — En Python

```
reported = vax[vax["reported"]].copy()
base_rate = (reported["doses_administered"] % 10 == 0).mean()

by_facility = (
    reported.assign(round_number=reported["doses_administered"] % 10 == 0)
    .groupby("facility_id")
    .agg(rounds=("round_number", "sum"), n=("round_number", "size"))
)
by_facility["share"] = by_facility["rounds"] / by_facility["n"]
print(f"district base rate: {base_rate:.3f}")
print(by_facility.nlargest(3, "share"))
```

## Étalonnez sur ce que produit le hasard — En R

```
base_rate <- mean(vax$doses_administered[vax$report_submitted] %% 10 == 0)

by_facility <- vax |>
  filter(report_submitted) |>
  summarise(rounds = sum(doses_administered %% 10 == 0), n = n(), .by = facility_id) |>
  mutate(share = rounds / n) |>
  arrange(desc(share))
```

## Étalonnez sur ce que produit le hasard — En Python

```
from scipy.stats import binomtest

worst = by_facility.nlargest(1, "share").iloc[0]
p = binomtest(int(worst["rounds"]), int(worst["n"]), base_rate,
              alternative="greater").pvalue
print(f"p = {p:.4f}, Bonferroni across 38 facilities: {min(1, p * 38):.3f}")
```

## Étalonnez sur ce que produit le hasard — En R

```
worst <- by_facility[1, ]  
p <- binom.test(worst$rounds, worst$n, base_rate, alternative = "greater")$p.value  
c(p = p, bonferroni = min(1, p * 38))
```

## Étalonnez sur ce que produit le hasard

- $p = 0,003$  pris isolément, et 0,11 après correction pour avoir regardé trente-huit formations — Trois formations passent. . .

## Étalonnez sur ce que produit le hasard

Le résultat est : **aucun indice de préférence de chiffres dans le rapportage vaccinal**. C'est un vrai constat, il a pris quatre lignes, et c'est celui qu'il faut espérer.

## Pourquoi le cas négatif compte davantage

- **Calculez le taux de base sur les données**, pas sur la théorie. Les chiffres terminaux ne sont pas uniformes quand. . .
- **Corrigez du nombre d'unités examinées**. Tester trente-huit formations à 5 % produit environ deux positifs sur des. . .
- **Exigez un mécanisme**. Les 69 % de l'équipe 2 en ont un — l'équipe a lu la toise au demi-centimètre le plus proche. . . .

# Ce que vous avez le droit d'écrire

| N'écrivez pas                                     | Écrivez                                                                                                                                                                                   |
|---------------------------------------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| « Les données de FAC020 semblent fabriquées »     | « FAC020 rapporte des nombres ronds plus souvent que la moyenne du district ; l'écart n'est pas significatif après ajustement pour le nombre de formations examinées »                    |
| « Les mesures de l'équipe 2 ne sont pas fiables » | « Les tailles de l'équipe 2 se terminent par ,0 ou ,5 dans 69 % des cas contre 18 à 23 % pour les autres équipes, ce qui est compatible avec une lecture de la toise au demi-centimètre » |
| « Les âges sont inexacts »                        | « 24 % des âges tombent sur une année entière exacte contre 9 % attendus, ce qui indique que l'âge a été estimé plutôt que documenté pour une part substantielle des enfants »            |

- Vous avez maintenant des constats — un effondrement de complétude, un facteur de vérification de 1,42, une équipe qui arrondit ses tailles.

Lire la leçon complète, avec le code exécutable

```
https://data-analysis.cassion.dev/fr/cours/data-quality-assessment/  
dqa-06-digit-preference
```