

Une moyenne de 29,6 et une médiane de zéro

Le dénombrement moyen d'E. coli dans les ménages testés vaut 29,6 UFC/100 mL. La médiane vaut 0. Plus de la moitié de l'échantillon n'a aucune contamination détectable, et la moyenne décrit une queue qui monte à 608.

Pierre Robentz Cassion

Leçon 1 sur 8

Statistiques appliquées aux programmes — Regardez-la d'abord

Ce que couvre cette leçon

- Calculez le résumé en dernier
- Regardez la forme
- Trois formes et ce que chacune exige
- Quand l'asymétrie est l'erreur de données
- Ce qu'il faut rapporter pour chaque forme
- L'histogramme qu'il faut toujours tracer
- La suite

Calculez le résumé en dernier — En Python

```
import pandas as pd

wash = pd.read_csv("wash-household-survey-2024.v1.csv")
ecoli = wash["ecoli_cfu_100ml"].dropna()

print(ecoli.describe().round(1))
```

Calculez le résumé en dernier — En R

```
library(dplyr)

wash |> filter(!is.na(ecoli_cfu_100ml)) |>
  summarise(n = n(), mean = mean(ecoli_cfu_100ml),
            median = median(ecoli_cfu_100ml), sd = sd(ecoli_cfu_100ml))
```

Calculez le résumé en dernier

	Valeur
n	802
Moyenne	29,6
Médiane	0,0
Écart-type	85,6
Maximum	608

- La moyenne vaut 29,6 et la médiane 0 — Ces deux nombres décrivent les mêmes 802 ménages, et un seul des deux en décrit. . .

Regardez la forme — En Python

```
bins = [-1, 0, 10, 100, 1000]
labels = ["0 (none detected)", "1-10", "11-100", ">100"]
print(pd.cut(ecoli, bins, labels=labels).value_counts().reindex(labels))
print(f"\nskew: {ecoli.skew():.2f}")
```

Regardez la forme — En R

```
wash |> filter(!is.na(ecoli_cfu_100ml)) |>  
  count(band = cut(ecoli_cfu_100ml, c(-1, 0, 10, 100, Inf)))
```

Regardez la forme

Bande	Ménages	Part
0 — rien de détecté	430	53,6 %
1-10	173	21,6 %
11-100	139	17,3 %
>100	60	7,5 %

Regardez la forme

- Asymétrie +4,04 — Une distribution symétrique a une asymétrie proche de zéro ; au-delà d'environ +1, la moyenne et la...
- C'est pourquoi le cours EAH rapportait des classes de risque et non une moyenne — Les classes ne sont pas un choix de...

Trois formes et ce que chacune exige — En Python (suite)

```
import numpy as np

def profile(series, name):
    s = series.dropna()
    return {
        "indicator": name, "n": len(s),
        "mean": round(s.mean(), 1), "median": round(s.median(), 1),
        "skew": round(s.skew(), 2),
    }

points = pd.read_csv("water-point-monitoring-2024.v1.csv")
protection = pd.read_csv("protection-referrals-2024.v1.csv")

print(pd.DataFrame([
    profile(wash["ecoli_cfu_100ml"], "E. coli, CFU/100mL"),
    profile(points["days_since_breakdown"], "days a water point is down"),
```

Trois formes et ce que chacune exige — En Python (suite)

```
profile(protection["days_to_first_service"], "days to first service"),  
profile(wash["litres_per_person_day"], "litres per person per day"),  
]))
```

Trois formes et ce que chacune exige — En R

```
# One function, four indicators, four different answers about which summary  
# to report.
```

Trois formes et ce que chacune exige

Indicateur	n	Moyenne	Médiane	Asymétrie
E. coli, UFC/100 mL	802	29,6	0,0	+4,04
Jours d'arrêt d'un point d'eau	709	125,7	56,0	+1,64
Jours jusqu'au premier service	728	10,1	9,0	+0,52
Litres par personne et par jour	2 403	24,2	23,7	+8,54

Trois formes et ce que chacune exige

- Le délai jusqu'au premier service est presque symétrique — moyenne 10,1, médiane 9,0, asymétrie +0,52
- Les jours d'arrêt d'un point d'eau sont fortement asymétriques à droite — une moyenne de 125,7 contre une médiane de . . .
- Les litres par personne et par jour paraissent presque symétriques et portent une asymétrie de +8,54 — C'est cette . . .

Quand l'asymétrie est l'erreur de données — En Python

```
litres = wash["litres_per_person_day"].dropna()
print(f"median {litres.median():.1f}, p90 {litres.quantile(0.90):.1f}, "
      f"max {litres.max():.1f}")
print(f"above 80: {(litres > 80).sum()} households")
```

Quand l'asymétrie est l'erreur de données — En R

```
wash |> summarise(median = median(litres_per_person_day),  
                  p90 = quantile(litres_per_person_day, 0.9),  
                  max = max(litres_per_person_day))
```

Quand l'asymétrie est l'erreur de données — En Python

```
clean = litres[litres <= 80]
print(f"after removing 11 rows: skew {clean.skew():.2f}, "  
      f"mean {clean.mean():.1f}, median {clean.median():.1f}")
```

Quand l'asymétrie est l'erreur de données — En R

Recompute after the correction and see whether the shape was real.

Quand l'asymétrie est l'erreur de données

- Onze lignes sur 2 403 et l'asymétrie passe de +8,54 à +0,06 — Moyenne 23,5, médiane 23,6 — une distribution symétrique. . .
- Calculez l'asymétrie, puis décidez si c'est un constat ou un bogue — Les deux sont courants et se distinguent en. . .

Ce qu'il faut rapporter pour chaque forme

Forme	Rapportez	Ne rapportez pas
À peu près symétrique	Moyenne et écart-type	—
Asymétrique à droite	Médiane et écart interquartile	Une moyenne sans la médiane à côté
Gonflée en zéros	La part à zéro , puis la distribution du reste	Une moyenne, tout court
Proportion bornée	La proportion et son intervalle	Un écart-type

Ce qu'il faut rapporter pour chaque forme

- E. coli est gonflée en zéros — ce qui est un cas distinct de la simple asymétrie : 53,6 % de l'échantillon se trouve. . .

Ce qu'il faut rapporter pour chaque forme — Exemple

E. coli at point of collection, 802 households tested

No detectable E. coli	53.6%	430 households
-----------------------	-------	----------------

Among the 372 with any detected:

median	13 CFU/100 mL
interquartile range	5 to 49
maximum	608

Mean over all 802 households is 29.6 CFU/100 mL. It is not reported as a summary because 53.6% of the sample sits at zero and the mean falls in a range occupied by almost no household.

L'histogramme qu'il faut toujours tracer — En Python

```
import matplotlib.pyplot as plt

fig, axes = plt.subplots(1, 2, figsize=(9, 3))
axes[0].hist(ecoli, bins=40)
axes[0].set_title("E. coli, raw")
axes[1].hist(np.log1p(ecoli), bins=40)
axes[1].set_title("log(1 + E. coli)")
plt.tight_layout()
```

L'histogramme qu'il faut toujours tracer — En R

```
hist(wash$ecoli_cfu_100ml, breaks = 40)  
hist(log1p(wash$ecoli_cfu_100ml), breaks = 40)
```

L'histogramme qu'il faut toujours tracer

- Tracez-le avant de résumer, à chaque fois, et ne le mettez pas dans le rapport — L'histogramme est un instrument pour. . .
- Deux minutes d'observation auraient évité chaque erreur de cette leçon — et c'est tout l'argument en faveur de cette. . .

- Vous savez maintenant quel nombre unique décrit un indicateur.

Lire la leçon complète, avec le code exécutable

`https://data-analysis.cassion.dev/fr/cours/`

`applied-statistics-for-programmes/stat-01-the-mean-that-describes-nobody`