

$r = 0,039$, et tout le monde attendait 0,5

La présence et la littératie de fin d'année corrèlent à 0,039 sur 659 élèves — un intervalle de $-0,04$ à $+0,12$, ce qui est aussi proche de rien qu'un jeu de données peut l'être. La corrélation que tout le monde suppose présente dans le registre n'y est pas, et le rapporter est le constat.

Pierre Robentz Cassion

Leçon 7 sur 8

Statistiques appliquées aux programmes — Ce que le rapport affirme

Ce que couvre cette leçon

- La corrélation qui n'y est pas
- Un intervalle sur une corrélation
- La corrélation qui y est et ne dit rien
- Élevez-la au carré avant de la décrire
- Pearson, Spearman, et onze mauvaises lignes
- Corrélés parce qu'autre chose a bougé les deux
- Et le regroupement de la leçon précédente
- Écrivez la phrase
- Rapportez-le en entier
- La suite

La corrélation qui n'y est pas — En Python (suite)

```
import pandas as pd
import numpy as np

assessment = pd.read_csv("learning-assessment-2024.v1.csv")
attendance = pd.read_csv("school-attendance-2024.v1.csv")
roster = pd.read_csv("school-roster-2024.v1.csv")

MARKS = {"true": True, "Y": True, "false": False, "N": False}
marked = attendance[attendance["present"].isin(MARKS)].copy()
marked["attended"] = marked["present"].map(MARKS)
rate = marked.groupby("student_id")["attended"].mean().rename("attendance")

endline = (assessment[assessment["assessment_round"] == "endline"]
           .pivot_table(index="student_id", columns="domain", values="raw_score"))
joined = endline.join(rate, how="inner").join(roster.set_index("student_id"))
```

La corrélation qui n'y est pas — En Python (suite)

```
print(joined[["literacy", "numeracy", "attendance"]].corr().round(3))
```

La corrélation qui n'y est pas — En R

```
library(dplyr)
```

```
joined |> select(literacy, numeracy, attendance) |> cor(use = "complete.obs")
```

La corrélation qui n'y est pas

Paire	r	n
Présence et littératie de fin d'année	0,039	659
Présence et numératie de fin d'année	−0,006	659
Littératie et numératie	0,849	659

La corrélation qui n'y est pas

- La présence et la littératie corrèlent à 0,039 — Chaque document de programme éducatif suppose cette relation et la...

Un intervalle sur une corrélation — En Python

```
def fisher_ci(r, n, z=1.96):  
    zr = np.arctanh(r)  
    se = 1 / np.sqrt(n - 3)  
    return np.tanh(zr - z * se), np.tanh(zr + z * se)  
  
for label, r, n in [("attendance vs literacy", 0.039, 659),  
                    ("literacy vs numeracy", 0.849, 659)]:  
    lo, hi = fisher_ci(r, n)  
    print(f"{label:24} r = {r:+.3f} 95% CI [{lo:+.3f}, {hi:+.3f}]" )
```

Un intervalle sur une corrélation — En R

```
cor.test(joined$attendance, joined$literacy)
```

Un intervalle sur une corrélation

- $r = 0,039$, IC à 95 % $-0,037$ à $+0,115$ — L'intervalle contient zéro et toute valeur qu'il contient est négligeable ; ce. . .
- Une corrélation sans intervalle est le même défaut qu'une proportion sans le sien — et `cor.test` le donne gratuitement

La corrélation qui y est et ne dit rien

- C'est un enfant, un jour d'épreuve, deux copies de quarante items — Un enfant malade, à jeun, incapable de lire les. . .
- Demandez ce qu'il faudrait pour que la corrélation soit nulle — Ici la réponse est « le même enfant devrait réussir. . .
- Une corrélation n'est informative que si son absence était plausible — et c'est un jugement sur le monde plutôt que sur. . .

Élevez-la au carré avant de la décrire — En Python

```
for label, r in [("attendance vs literacy", 0.039),  
                 ("attendance vs literacy, school level", 0.199),  
                 ("literacy vs numeracy", 0.849)]:  
    print(f"{label:38} r = {r:+.3f}    r-squared = {r**2:.3f}")
```

Élevez-la au carré avant de la décrire — En R

r^2 is the share of variance, and it is much smaller than r looks.

Élevez-la au carré avant de la décrire

Paire	r	r^2	Se lit
Présence et littératie	0,039	0,002	Deux dixièmes de un pour cent
Présence et littératie, niveau école	0,199	0,040	Quatre pour cent
Littératie et numératie	0,849	0,721	Soixante-douze pour cent

Élevez-la au carré avant de la décrire

- Un r de 0,199 sonne comme une relation et explique quatre pour cent de la variation — Élever au carré est la défense la...

Pearson, Spearman, et onze mauvaises lignes — En Python

```
wash = pd.read_csv("wash-household-survey-2024.v1.csv")
pair = wash.dropna(subset=["round_trip_minutes", "litres_per_person_day"])
clean = pair[pair["litres_per_person_day"] <= 80]

for label, frame in [("all rows", pair), ("11 rows removed", clean)]:
    p = frame["round_trip_minutes"].corr(frame["litres_per_person_day"])
    s = frame["round_trip_minutes"].corr(frame["litres_per_person_day"],
                                         method="spearman")
    print(f"{label:18} n = {len(frame):5}  Pearson {p:+.3f}  Spearman {s:+.3f}")
```

Pearson, Spearman, et onze mauvaises lignes — En R

```
cor(pair$round_trip_minutes, pair$litres_per_person_day)           # Pearson  
cor(pair$round_trip_minutes, pair$litres_per_person_day, method = "spearman")
```

Pearson, Spearman, et onze mauvaises lignes

Lignes	n	Pearson	Spearman
Toutes	2 403	−0,185	−0,285
Onze retirées	2 392	−0,293	−0,287

Pearson, Spearman, et onze mauvaises lignes

- Onze lignes sur 2 403 amputent Pearson de plus d'un tiers. Spearman n'a pas bougé
- Rapportez Spearman quand l'une des deux variables est asymétrique ou porte des valeurs extrêmes que vous n'avez pas. . .
- La réponse corrigée est celle qui sert — des trajets plus longs vont avec moins d'eau consommée, $r = -0,29$ sur 2 392. . .

Corrélés parce qu'autre chose a bougé les deux — En Python

```
prices = pd.read_csv("market-prices-2024.v1.csv")
wide = prices.pivot_table(index="period", columns="commodity",
                           values="price_htg", aggfunc="mean").sort_index()

print(f"beans and maize, levels:           {wide['beans-black'].corr(wide['maize']):.3f}")
changes = wide.diff().dropna()
print(f"beans and maize, month-on-month:   {changes['beans-black'].corr(changes['maize']):.3f}")
```

Corrélés parce qu'autre chose a bougé les deux — En R

```
cor(wide$`beans-black`, wide$maize)  
cor(diff(wide$`beans-black`), diff(wide$maize))
```

Corrélés parce qu'autre chose a bougé les deux

- Le haricot et le maïs corréleront à 0,986 en niveaux et à 0,876 en variations mensuelles — Aucun des deux nombres ne...
- Corréler deux séries dans le temps donnera presque toujours un grand nombre — parce que presque tout suit une tendance

Et le regroupement de la leçon précédente — En Python

```
by_school = joined.groupby("school_id")[["attendance", "literacy"]].mean()
print(f"student level  r = {joined['attendance'].corr(joined['literacy']):.3f}  n = {len(
    joined)}")
print(f"school level  r = {by_school['attendance'].corr(by_school['literacy']):.3f}  n = {
    len(by_school)}")
```

Et le regroupement de la leçon précédente — En R

Same two variables, two units of analysis, two answers.

Et le regroupement de la leçon précédente

- 0,039 sur 659 élèves et 0,199 sur 24 écoles — Le chiffre au niveau école a un intervalle de $-0,22$ à $+0,56$ sur 24. . .
- Dites entre quelles unités porte la corrélation — à chaque fois

Écrivez la phrase

- Pas ceci — « La présence est corrélée aux résultats de littératie
- Ni ceci — « Il n'y a aucune relation entre présence et littératie
- Ceci — « Sur 659 élèves, la présence du trimestre février-avril ne montre aucune association avec la littératie de fin. . .
- La troisième phrase est celle à copier — Elle énonce le constat, l'intervalle, ce qu'il signifie pour le programme, et. . .

Rapportez-le en entier — Exemple

Attendance and learning outcomes, endline 2024

Attendance vs endline literacy $r = 0.04$ 95% CI -0.04 to +0.12 $n = 659$

Attendance vs endline numeracy $r = -0.01$ 95% CI -0.08 to +0.07 $n = 659$

Between students within the term. No association at student level.

Aggregated to the 24 schools, $r = 0.20$ (95% CI -0.22 to +0.56), which is also consistent with no relationship.

Attendance is compressed: the interquartile range is 86.7% to 97.8%, so students with the low attendance that a learning effect would show up at are rare in this cohort. This is a null result about the range observed, not about attendance in general.

- Le dernier paragraphe est ce qui empêche de surlire le résultat nul — et c'est le pendant de la phrase « sous-puissant, . . .

- La dernière leçon assemble tout cela en une section statistique de rapport — ce qui y entre, ce qui reste en annexe, et ce qu'un relecteur demandera et que ce cours a déjà calculé.

Lire la leçon complète, avec le code exécutable

`https://data-analysis.cassion.dev/fr/cours/
applied-statistics-for-programmes/stat-07-correlation-and-the-sentence`