r = 0.039, and everyone expected 0.5

Attendance and endline literacy correlate at 0.039 across 659 students — an interval from −0.04 to +0.12, which is as close to nothing as a dataset gets. The correlation everyone assumes is in the register is not in this one, and reporting that is the finding.

Pierre Robentz Cassion

Lesson 7 of 8

What this lesson covers

- The correlation that is not there
- An interval on a correlation
- The correlation that is there and says nothing
- Square it before you describe it
- Pearson, Spearman, and eleven bad rows
- Correlated because something else moved both
- And the clustering from the last lesson
- Write the sentence
- Report it whole
- What comes next

The correlation that is not there — In Python (cont.)

```
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"))
```

The correlation that is not there — In Python (cont.)

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

The correlation that is not there — In R

```
library(dplyr)

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

The correlation that is not there

Pair	r	n
Attendance and endline literacy	**0.039**	659
Attendance and endline numeracy	**−0.006**	659
Literacy and numeracy	0.849	659

The correlation that is not there

- Attendance and literacy correlate at 0.039 — Every education programme document assumes that relationship and most...

An interval on a correlation — In 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}]")
```

An interval on a correlation — In R

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

An interval on a correlation

- $r = 0.039$, 95% CI -0.037 to $+0.115$ — The interval contains zero and every value inside it is negligible, so this is not…
- A correlation with no interval is the same defect as a proportion with none — and `cor.test` gives it for free

The correlation that is there and says nothing

- It is one child, one test day, two forty-item papers — A child who was ill, who had not eaten, who could not read the...
- Ask what would have to be true for the correlation to be zero — Here the answer is "the same child would have to...
- A correlation is only informative when its absence was plausible — and that is a judgement about the world rather than...

Square it before you describe it — In 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}")
```

Square it before you describe it — In R

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

Square it before you describe it

Pair	r	r^2	Reads as
Attendance and literacy	0.039	**0.002**	Two tenths of one per cent
Attendance and literacy, school level	0.199	0.040	Four per cent
Literacy and numeracy	0.849	0.721	Seventy-two per cent

Square it before you describe it

- An r of 0.199 sounds like a relationship and explains four per cent of the variation — Squaring is the cheapest defence. . .

Pearson, Spearman, and eleven bad rows — In 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, and eleven bad rows — In R

```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, and eleven bad rows

Rows	n	Pearson	Spearman
All	2,403	$-\mathbf{0.185}$	-0.285
Eleven removed	2,392	$-\mathbf{0.293}$	-0.287

Pearson, Spearman, and eleven bad rows

- Eleven rows in 2,403 cut Pearson by more than a third. Spearman did not move
- Report Spearman when either variable is skewed or has outliers you have not resolved, and Pearson when both are roughly...
- The corrected answer is the useful one — longer trips to water go with less water used, $r = -0.29$ on 2,392 households,...

Correlated because something else moved both — In 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']):.3
    f}")
```

Correlated because something else moved both — In R

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

Correlated because something else moved both

- Beans and maize correlate at 0.986 in levels and 0.876 in month-on-month changes — Neither number means the price of...
- Correlating two series over time will nearly always give a large number — because almost everything trends

And the clustering from the last lesson — In 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)}")
```

And the clustering from the last lesson — In R

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

And the clustering from the last lesson

- 0.039 across 659 students and 0.199 across 24 schools — The school-level figure has an interval from -0.22 to $+0.56$ on...
- Say which unit the correlation is between — every time

Write the sentence

- Not this — "Attendance is correlated with literacy outcomes." True of the school level, false of the student level, and...
- Nor this — "There is no relationship between attendance and literacy." Stronger than the data supports; the interval...
- This — "Across 659 students, attendance over the February–April term shows no association with endline literacy ($r =$...
- The third sentence is the one to copy — It states the finding, the interval, what it means for the programme, and the...

Report it whole — Example

```
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.
```

Report it whole

- The last paragraph is what stops the null being over-read — and it is the counterpart of the "underpowered, not null"...

What comes next

- The last lesson assembles all of this into the statistics section of a report — what goes in, what stays in the appendix, and what a reviewer will ask for that this course has already computed.

Read the full lesson, with runnable code

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