Minus 0.96 points, and the assumption you cannot test

Subtracting each group's own baseline gives a difference-in-differences of -0.96 points. It rests on the two groups having been about to change by the same amount, which is untestable with two rounds — so the lesson is how to argue for it rather than how to check it.

Pierre Robentz Cassion

What this lesson covers

- Two differences, subtracted
- Fit it as a regression, because you need the interval
- The assumption, stated exactly
- Why two rounds cannot test it
- Four arguments to make when you cannot test it
- Where difference-in-differences goes wrong
- Report it whole
- What comes next

Two differences, subtracted — In Python

```
import pandas as pd
import numpy as np

cells = d.groupby("feeding_programme")[["baseline", "endline"]].mean()
print(cells.round(4))

did = ((cells.loc[True, "endline"] - cells.loc[True, "baseline"])
       - (cells.loc[False, "endline"] - cells.loc[False, "baseline"]))
print(f"difference-in-differences: {did:+.4f}")
```

Two differences, subtracted — In R

```
library(dplyr)

d |> summarise(baseline = mean(baseline), endline = mean(endline),
               .by = feeding_programme) |>
  mutate(gain = endline - baseline)
```

Two differences, subtracted

	Baseline	**Endline**	**Change**
Feeding programme	57.75%	64.41%	+6.66
No programme	54.67%	62.29%	+7.62
Difference	+3.09	+2.12	**−0.96**

Two differences, subtracted

- The estimate can be read down the last column or across the last row and it is the same number — That symmetry is what...
- The baseline gap of 3.09 points is what the design removes — A cross-sectional comparison at endline would have...

Fit it as a regression, because you need the interval — In Python

```
import statsmodels.formula.api as smf

d = d.assign(gain=d["endline"] - d["baseline"])

naive = smf.ols("gain ~ feeding_programme", data=d).fit()
clustered = naive.get_robustcov_results(cov_type="cluster",
                                        groups=d["school_id"])

by_school = d.groupby(["school_id", "feeding_programme"])["gain"].mean().reset_index()
aggregated = smf.ols("gain ~ feeding_programme", data=by_school).fit()
```

Fit it as a regression, because you need the interval — In R

```
lm(gain ~ feeding_programme, data = d)          # naive
lm(gain ~ feeding_programme, data = by_school)  # school level
```

Fit it as a regression, because you need the interval

Approach	Estimate	SE	95% CI
Student level, naive	−0.96 pts	0.98	−2.89 to +0.96
Student level, clustered on school	**−0.96 pts**	**1.29**	**−3.48 to +1.56**
School level, 15 vs 9	−0.68 pts	1.18	−2.99 to +1.63

Fit it as a regression, because you need the interval

- The clustered interval is the one to report — for the reason the regression course established: the programme was...
- Every version crosses zero comfortably — There is no effect to report, and the interval says the study could not have...

The assumption, stated exactly

- Parallel trends: in the absence of the programme, the two groups' outcomes would have changed by the same amount
- It does not require the groups to be similar — They start 3.09 points apart and that is fine — the design differences...
- It does not require the trends to be flat — Both groups can be improving; the assumption is that they would have...
- It does not require the same variance, sample size or composition — Those affect the interval, not the identification
- What it does require is unobservable — because it is a statement about a world in which the programme did not happen

Why two rounds cannot test it — In Python

```
rounds = assessment["assessment_round"].unique()
print(rounds)
```

Why two rounds cannot test it — In R

```
unique(assessment$assessment_round)
```

Why two rounds cannot test it

- There are two: baseline and endline — With two points you can draw exactly one line through each group, so any...
- With three or more pre-programme rounds you can look — Plot each group's outcome over the pre-period and see whether...
- The design implication is a data-collection decision made years earlier — If you expect to evaluate by...

Four arguments to make when you cannot test it

- Name why the programme schools were chosen — If selection was on something time-invariant — where they are, who runs...
- Show that other outcomes moved in parallel — Numeracy is measured on the same children and is not what a feeding...

Four arguments to make when you cannot test it — In Python

```
def panel(domain):
    wide = (assessment[assessment["domain"] == domain]
            .pivot_table(index="student_id", columns="assessment_round",
                         values="pct").dropna())
    out = wide.join(roster.set_index("student_id"), how="inner").reset_index()
    return out.assign(gain=out["endline"] - out["baseline"])

for domain in ("literacy", "numeracy"):
    frame = panel(domain)
    fit = smf.ols("gain ~ feeding_programme", data=frame).fit(
        cov_type="cluster", cov_kwds={"groups": frame["school_id"]})
    print(f"{domain}: {fit.params['feeding_programme[T.True]']:+.4f}")
```

Four arguments to make when you cannot test it — In R

```
# A placebo outcome: it should show nothing, and here it does.
```

Four arguments to make when you cannot test it

- Numeracy gives +0.39 points, 95% CI −1.16 to +1.94 — the same design applied to an outcome the programme was not...
- Test a placebo period if one exists — Two pre-programme rounds should give a difference-in-differences of zero
- Show the result is not driven by one unit — Drop each school in turn and refit; if the estimate swings, the design is...

Four arguments to make when you cannot test it — In Python

```
for school in sorted(d["school_id"].unique()):
    subset = d[d["school_id"] != school]
    est = smf.ols("gain ~ feeding_programme", data=subset).fit()
    print(f"without {school}: {est.params['feeding_programme[T.True]']:+.4f}")
```

Four arguments to make when you cannot test it — In R

```
# 24 refits, one line each. Cheap, and a reviewer will ask.
```

Four arguments to make when you cannot test it

- Dropping any one school moves the estimate between -1.61 and -0.37 points — No single school carries it, which is worth...

Where difference-in-differences goes wrong

- Different timing — If the programme started in different schools in different months, a single before-after split...
- Composition change — The 585 children are those with both rounds
- Anticipation — If schools changed behaviour once they knew the programme was coming, the baseline is already...
- A control group that is affected — Children moving between schools, teachers transferring, or a nearby school changing...

Report it whole — Example (cont.)

School feeding and literacy, difference-in-differences

585 students with both assessment rounds, in 24 schools.

	Baseline	Endline	Change
Feeding (15 sch)	57.75%	64.41%	+6.66
No programme (9)	54.67%	62.29%	+7.62
Difference	+3.09	+2.12	-0.96

Estimate -0.96 points, 95% CI -3.48 to +1.56, clustered on 24 schools.
No effect on literacy is detected.

The estimate assumes the two groups would have changed by the same amount without the programme. Only two assessment rounds exist, so this cannot be examined and is argued rather than tested: assignment appears to follow district and school size rather than a trend in results, and the same

Report it whole — Example (cont.)

```
design applied to numeracy also shows nothing.

The interval excludes effects larger than 1.6 points in either direction
but not smaller ones. The design's minimum detectable effect was 4.9
points, so this is a null result about large effects only.
```

Report it whole

- The last paragraph converts a null into a statement with a boundary — which is what the statistics course asked for and...

What comes next

- Difference-in-differences uses the baseline by subtracting it.

Read the full lesson, with runnable code

https://data-analysis.cassion.dev/courses/impact-evaluation-methods/ie-03-difference-in-differences