

Data and causality: what a comparison of two groups can show

The ideas in five lines

1. A difference between two groups is a fact. Whether one thing *caused* it is a separate question.
2. A causal effect compares the same person with and without the treatment. Only one of the two is ever seen.
3. When people choose their group, the groups can differ in other ways. This is **selection**.
4. A **confounder** (omitted variable) affects both the choice and the outcome.
5. **Random assignment** makes the groups alike on average. Then the difference in means is the causal effect.

The formula

Outcome: $y = \beta T + \gamma m + e$, where $T = 1$ for the treated group and m is the omitted variable.

$$\bar{y}_{\text{treated}} - \bar{y}_{\text{others}} = \beta + \gamma \Delta m$$

The **bias** is $\gamma \Delta m$: the effect of the omitted variable times how much the groups differ in it.

Effect of m on y	Gap in m	The comparison...
$\gamma > 0$	$\Delta m > 0$	overstates β
$\gamma > 0$	$\Delta m < 0$	understates β
any	$\Delta m = 0$	equals β

Worked example

An app's users score 12 points higher. Suppose the app adds $\beta = 4$ points, one standard deviation of motivation adds $\gamma = 8$ points, and users are $\Delta m = 1$ standard deviation more motivated.

- Comparison: $4 + 8 \times 1 = 12$ points.
- Bias: $8 \times 1 = 8$ points. Only 4 of the 12 points are the app's.
- With a lottery, $\Delta m \approx 0$, so the difference is about 4 points.

Four traps

- "The sample is huge, so it is causal." No. A large sample cuts noise, not bias.
- "Bias always makes effects look bigger." No. It can go either way.
- "There is confounding, so there is no effect." No. Confounding blurs the effect; it does not remove it.
- "Ask people whether it worked." No. Opinions are not a comparison.

Words to know

treatment · outcome · counterfactual · selection · confounder (omitted variable) · randomised controlled trial

Try these

1. Students who attend extra classes score 9 points higher than those who do not. Name one confounder. Which way does it push the comparison?
2. $\beta = 2$, $\gamma = 5$, $\Delta m = 0.4$. What difference does the naive comparison show?
3. Same as question 2, but $\Delta m = -0.4$. What does the comparison show now?
4. The naive difference is 10 points, $\gamma = 6$ and $\Delta m = 1$. What is β ?
5. True or false: with 100,000 students, the comparison of users and non-users measures the causal effect.
6. A school has 200 free app licences and 600 applicants. How could it measure the app's effect?

Answers

1. For example, motivation or earlier grades. If more motivated students attend, and motivation raises scores, the comparison *overstates* the effect of the classes.
2. $2 + 5 \times 0.4 = 4$ points.
3. $2 + 5 \times (-0.4) = 0$. The comparison shows no difference, although the true effect is 2 points.
4. $10 = \beta + 6 \times 1$, so $\beta = 4$.
5. False. Sample size reduces noise; the bias $\gamma \Delta m$ stays the same.
6. Draw lots among the 600 applicants. Compare the average scores of the 200 winners and the 400 others. Chance decided who got the app, so the difference estimates its effect for applicants.