HUMANITARIAN ACADEMY · HUMANITAIRES JOBS

Mini MBA Programmes and MEAL: design, manage and prove impact

Course handout · Week 10 : Measuring impact: counterfactual, methods and contribution

A programme of exactly 3 months (12 weeks, 144 hours) to design, manage and evaluate humanitarian programmes: needs assessment and severity, results chain and indicators, targeting and cash transfers, planning and budget, MEAL plan and data quality, sampling, analysis, accountability, evaluation, impact measurement, value for money and adaptive management. Corrected numerical exercises, three monthly exams with case studies, and a verifiable diploma from the Humanitarian Academy.

https://humanitairejobs.org/en/academy/mini-mba-programmes-meal/semaine-10/

Free document from the Humanitarian Academy. It may be printed and shared freely for training purposes, without modification.

Week 10 : Measuring impact: counterfactual, methods and contribution

Objectives

- Understand the counterfactual problem and why a before-after comparison does not measure an effect.
- Choose and apply a quasi-experimental or experimental method suited to the humanitarian context, with its validity conditions.
- Use contribution approaches when attribution is impossible, and decide what it is reasonable to prove.

The fundamental problem: what would have happened without the programme?

The **impact** of a programme is the difference between what happened with the programme and what would have happened without it: the **counterfactual**. It is never possible to observe the same household directly both with and without the programme; all impact evaluation methods are ways of building a credible estimate of this counterfactual.

Two naive comparisons are misleading:

- **Before-after** among beneficiaries: the observed change mixes the effect of the programme with everything else that changed at the same time (harvest, prices, security, other aid). If the harvest was good, food security improves even without the programme; if it was poor, an effective programme can appear to have no effect.
- **Beneficiaries versus non-beneficiaries** after the programme: the two groups were different at the outset, precisely because targeting selected the most vulnerable. The comparison then underestimates the effect, or overestimates it if it was the most motivated who signed up (selection bias).

> The question to ask of any impact claim: "compared with what?". If the answer is "compared with before", it is not a measure of impact.

The randomised controlled trial and its use in humanitarian settings

In a **randomised controlled trial** (RCT), among the eligible people, those who receive the programme (treatment group) and those who do not receive it or receive it later (control group) are selected by lottery. Random assignment makes the two groups comparable on average on all characteristics, observed or not. The effect is simply the difference in average outcomes at the end.

In humanitarian settings, denying life-saving aid to eligible people in order to form a control group is **unethical**. But RCTs remain possible and have produced major knowledge (on cash, nutrition, mental health) in specific situations:

- **Insufficient resources** for everyone in any case: random selection is then a fair way of allocating, and the group not served acts as the control.
- **Phased roll-out**: everyone will receive the programme, but the order is drawn by lot (stepped wedge design).
- **Comparison of two variants** of a programme, both considered acceptable: monthly or lump-sum cash, with or without accompanying support.

RCTs are expensive, take time and answer specific questions; they are justified when the answer will serve many programmes. For most projects, quasi-experimental methods are sufficient.

Difference-in-differences

Difference-in-differences compares the change in an assisted group with the change in a non-assisted comparison group: effect = (after − before) in the assisted group − (after − before) in the comparison group. It eliminates stable initial differences between the groups and common trends (harvest, prices).

Example. Proportion of households with acceptable food consumption. Assisted group: 28% before, 52% after (+24 points). Comparison group: 35% before, 44% after (+9 points). Estimated effect: 24 − 9 = 15 points. A before-after comparison would have attributed 24 points to the programme; a comparison of final results alone would have attributed 8.

Validity condition: the **parallel trends** assumption. Without the programme, the two groups would have changed in the same way. It is checked where possible with several measurements before the programme, and one makes sure that the comparison group does not receive other significant aid during the period (contamination) and does not suffer a specific shock.

In humanitarian settings, the comparison group is often a neighbouring area not covered for lack of funding, or eligible households on a waiting list. A comparison group is never created by denying aid for the purposes of the study.

Matching and regression discontinuity

Matching builds a comparison group by pairing each assisted household with one or more non-assisted households that resemble it on characteristics measured before the programme (size, assets, income, status, area). The **propensity score** method summarises these characteristics as a probability of being assisted, and households with close scores are compared. Its limitation: it corrects only observed differences. If motivation or social networks influenced registration and are not measured, the bias remains.

Regression discontinuity applies when eligibility depends on a threshold on a score: for example, households with a vulnerability score of 60 or more are assisted, those below are not. Households just above and just below the threshold (57-59 and 60-62) are almost identical; the difference in outcomes between them estimates the effect of the programme around the

threshold. Conditions: the threshold is strictly applied, and households cannot manipulate their score to get above it.

Example. Households with a score of 57-59: 38% with acceptable consumption. Households with a score of 60-62 (assisted): 51%. Estimated effect around the threshold: about 13 points. This result holds for households close to the threshold, not necessarily for the most vulnerable.

The **interrupted time series** compares the trend of an indicator before and after the start of a programme, over many measurement points (for example weekly nutrition admissions). It is useful when no comparison group exists, provided there are enough points before the intervention.

Contribution approaches: when attribution is impossible

Many important effects do not lend themselves to statistical methods: advocacy that changes a law, capacity strengthening of local organisations, improved coordination. They cannot be compared with a quantified counterfactual. The aim is then to establish credibly that the programme **contributed** to the change.

- **Contribution analysis** (Mayne): set out the theory of change, gather evidence for each link, examine alternative explanations, and build a "contribution story" that is then critiqued by stakeholders.
- **Process tracing**: causal hypotheses are tested by looking for evidence whose presence or absence makes it possible to confirm or rule out an explanation (for example, the word-for-word reuse of an advocacy note in a decree).
- **Outcome harvesting**: first identify the changes observed among actors, then work backwards to determine whether and how the programme contributed to them.
- **Most significant change**: stories of change are collected from people, and panels select the most significant ones, explaining why. Useful for discovering unintended effects.

> Rigour does not mean quantitative. A well-conducted contribution analysis that seriously examines alternative explanations is more credible than a quantified before-after comparison.

Choosing the method and deciding what needs proving

The choice of method depends on four questions:

- **What decision?** Scaling up a pilot model nationally justifies strong evidence; adjusting an ongoing programme can make do with sufficient evidence.
- **Does the evidence already exist?** The effect of cash on food consumption is very well documented: there is no need to demonstrate it again with an RCT in every project. It is often enough to check that the programme is faithful to the model and that the intermediate effects are present.
- **What is feasible and ethical?** Is there a comparison group without denying aid? A clear eligibility threshold? A phased roll-out?

- **What budget and what timeframe?** A serious impact evaluation requires a baseline before the start: it is designed with the programme, not at the end.

As regards **sample size**, an impact evaluation must have sufficient power to detect the expected effect. Detecting a 15-point effect requires far fewer households than a 5-point effect. If the expected effect is small and the sample is small, the study will almost certainly conclude "no significant effect", which will be misinterpreted as "no effect".

A final point: present the **level of evidence** honestly. "Assisted households improved their consumption by 24 points; compared with a comparison area, the attributable effect is estimated at 15 points, under the parallel trends assumption supported by data from two previous rounds" is a sentence that inspires confidence.

Exercise: estimating an effect and critiquing the method

Situation. A livelihoods programme supported 1,200 households. Median monthly income: assisted households 42 USD before, 71 USD after. Neighbouring area not covered (for lack of funding): 47 USD before, 58 USD after. During the programme, another NGO distributed seeds in half of the villages in the neighbouring area. The final report states: "the programme increased incomes by 69%".

Task. 1) Where does the 69% figure come from and why is it misleading? 2) Calculate the effect by difference-in-differences. 3) What is the likely effect of the seed distribution on your estimate, and what do you propose?

Answer. 69%: (71 − 42) / 42, a before-after comparison that attributes the entire increase to the programme, including the increase due to the context. Difference-in-differences: (71 − 42) − (58 − 47) = 29 − 11 = 18 USD per month, i.e. about 43% of the initial income of assisted households. The seeds probably increased incomes in the comparison area, which inflates its trend and underestimates the programme's effect (contamination violates the parallel trends assumption). Proposal: analyse separately the comparison villages without seeds, or control for this variable; present the estimate as a lower bound and state the limitation clearly.

Key takeaways

- Impact is measured against a counterfactual; before-after is not a measure of impact.
- Difference-in-differences = change in the assisted group − change in the comparison group, under the parallel trends assumption.
- An RCT in humanitarian settings is ethical only if it deprives no one of available aid: limited resources, phased roll-out, or comparison of variants.
- Regression discontinuity: compare households just above and just below a strict eligibility threshold.
- When attribution is impossible: contribution analysis, process tracing, outcome harvesting, most significant change.

Quiz

1. Assisted group: 30% before, 50% after. Comparison: 32% before, 40% after. Effect by difference-in-differences:
   A. 20 points
   B. 12 points
   C. 10 points
   D. 8 points

2. Which assumption is essential to difference-in-differences?
   A. The groups are identical at the outset
   B. Without the programme, the two groups would have changed in the same way
   C. The sample exceeds 1,000 households
   D. The programme lasted more than a year

3. In which situation is random selection ethically acceptable in humanitarian settings?
   A. To deny available life-saving aid to half of the eligible people
   B. When resources are in any case insufficient for all eligible people, or for the order of a phased roll-out
   C. Never
   D. Only if the donor requires it

4. The main limitation of propensity score matching is:
   A. It only works with RCTs
   B. It corrects only measured differences, not unobserved differences such as motivation
   C. It requires an eligibility threshold
   D. It is prohibited by the OECD

5. To show that advocacy contributed to the adoption of a law, the most suitable method is:
   A. A randomised trial
   B. A difference-in-differences
   C. A contribution analysis or process tracing
   D. A household survey

Quiz answers : 1. B · 2. B · 3. B · 4. B · 5. C