ACADÉMIE HUMANITAIRE · HUMANITAIRES JOBS

Mini MBA Programmes et MEAL : concevoir, piloter et prouver l'impact

Support de cours · Semaine 10 : Mesurer l'impact : contrefactuel, méthodes et contribution

Programme de 3 mois exactement (12 semaines, 144 heures) pour concevoir, piloter et évaluer des programmes humanitaires : évaluation des besoins et sévérité, chaîne de résultats et indicateurs, ciblage et transferts monétaires, planification et budget, plan MEAL et qualité des données, échantillonnage, analyse, redevabilité, évaluation, mesure d'impact, optimisation des ressources et gestion adaptative. Exercices chiffrés corrigés, trois examens mensuels avec études de cas, et un diplôme vérifiable de l'Académie Humanitaire.

https://humanitairejobs.org/fr/academie/mini-mba-programmes-meal/semaine-10/

Document gratuit de l'Académie Humanitaire. Il peut être imprimé et partagé librement à des fins de formation, sans modification.

Semaine 10 : Mesurer l'impact : contrefactuel, méthodes et contribution

Objectifs

- Comprendre le problème du contrefactuel et pourquoi une comparaison avant-après ne mesure pas un effet.
- Choisir et appliquer une méthode quasi expérimentale ou expérimentale adaptée au contexte humanitaire, avec ses conditions de validité.
- Utiliser les approches de contribution quand l'attribution est impossible, et décider ce qu'il est raisonnable de prouver.

Le problème fondamental : que se serait-il passé sans le programme ?

L'**impact** d'un programme est la différence entre ce qui s'est passé avec le programme et ce qui se serait passé sans lui : le **contrefactuel**. On ne peut jamais observer directement le même ménage avec et sans programme ; toutes les méthodes d'évaluation d'impact sont des façons de construire une estimation crédible de ce contrefactuel.

Deux comparaisons naïves sont trompeuses :

- **Avant-après** chez les bénéficiaires : le changement observé mélange l'effet du programme et tout ce qui a changé en même temps (récolte, prix, sécurité, autres aides). Si la récolte a été bonne, la sécurité alimentaire s'améliore même sans programme ; si elle a été mauvaise, un programme efficace peut paraître sans effet.
- **Bénéficiaires contre non-bénéficiaires** après le programme : les deux groupes étaient différents au départ, justement parce que le ciblage a choisi les plus vulnérables. La comparaison sous-estime alors l'effet, ou le surestime si ce sont les plus motivés qui se sont inscrits (biais de sélection).

> Question à poser devant toute affirmation d'impact : « par rapport à quoi ? ». Si la réponse est « par rapport à avant », ce n'est pas une mesure d'impact.

L'essai randomisé contrôlé et son usage en humanitaire

Dans un **essai randomisé contrôlé** (ECR), on tire au sort, parmi les personnes éligibles, celles qui reçoivent le programme (groupe de traitement) et celles qui ne le reçoivent pas ou plus tard (groupe de contrôle). Le tirage au sort rend les deux groupes comparables en moyenne sur toutes les caractéristiques, observées ou non. L'effet est simplement la différence des résultats moyens à la fin.

En humanitaire, refuser une aide vitale à des personnes éligibles pour constituer un groupe de contrôle est **contraire à l'éthique**. Mais des ECR restent possibles et ont produit des

connaissances majeures (sur le cash, la nutrition, la santé mentale) dans des situations précises :

- **Ressources insuffisantes** de toute façon pour tout le monde : le tirage au sort est alors une façon équitable de répartir, et le groupe non servi sert de contrôle.
- **Mise en place progressive** : tout le monde recevra le programme, mais l'ordre est tiré au sort (conception en escalier, « stepped wedge »).
- **Comparaison de deux variantes** d'un programme, toutes deux jugées acceptables : cash mensuel ou en une fois, avec ou sans accompagnement.

Les ECR coûtent cher, prennent du temps et répondent à des questions précises ; ils se justifient quand la réponse servira à de nombreux programmes. Pour la plupart des projets, des méthodes quasi expérimentales suffisent.

La différence de différences

La **différence de différences** compare l'évolution d'un groupe aidé à l'évolution d'un groupe de comparaison non aidé : effet = (après − avant) dans le groupe aidé − (après − avant) dans le groupe de comparaison. Elle élimine les différences initiales stables entre les groupes et les tendances communes (récolte, prix).

Exemple. Proportion de ménages avec une consommation acceptable. Groupe aidé : 28 % avant, 52 % après (+24 points). Groupe de comparaison : 35 % avant, 44 % après (+9 points). Effet estimé : 24 − 9 = 15 points. Une comparaison avant-après aurait attribué 24 points au programme ; une comparaison des seuls résultats finaux en aurait attribué 8.

Condition de validité : l'hypothèse des **tendances parallèles**. Sans le programme, les deux groupes auraient évolué de la même façon. On la vérifie si possible avec plusieurs mesures avant le programme, et l'on s'assure que le groupe de comparaison ne reçoit pas une autre aide importante pendant la période (contamination) et ne subit pas un choc spécifique.

En humanitaire, le groupe de comparaison est souvent une zone voisine non couverte faute de financement, ou des ménages éligibles inscrits sur liste d'attente. On ne crée jamais un groupe de comparaison en refusant l'aide pour les besoins de l'étude.

Appariement et régression sur discontinuité

L'**appariement** construit un groupe de comparaison en associant à chaque ménage aidé un ou plusieurs ménages non aidés qui lui ressemblent sur les caractéristiques mesurées avant le programme (taille, biens, revenus, statut, zone). La méthode du **score de propension** résume ces caractéristiques en une probabilité d'être aidé, et l'on compare des ménages au score proche. Sa limite : elle ne corrige que les différences observées. Si la motivation ou les réseaux sociaux ont influencé l'inscription et ne sont pas mesurés, le biais demeure.

La **régression sur discontinuité** s'applique quand l'éligibilité dépend d'un seuil sur un score : par exemple, ménages avec un score de vulnérabilité de 60 ou plus aidés, en dessous non. Les ménages juste au-dessus et juste en dessous du seuil (57-59 et 60-62) sont presque identiques ; la différence de résultats entre eux estime l'effet du programme autour du seuil. Conditions :

le seuil est appliqué strictement, et les ménages ne peuvent pas manipuler leur score pour passer au-dessus.

Exemple. Ménages au score 57-59 : 38 % avec une consommation acceptable. Ménages au score 60-62 (aidés) : 51 %. Effet estimé autour du seuil : environ 13 points. Ce résultat vaut pour les ménages proches du seuil, pas forcément pour les plus vulnérables.

La **série chronologique interrompue** compare la tendance d'un indicateur avant et après le début d'un programme, sur de nombreux points de mesure (par exemple les admissions hebdomadaires en nutrition). Elle est utile quand il n'existe aucun groupe de comparaison, à condition d'avoir assez de points avant l'intervention.

Approches de contribution : quand l'attribution est impossible

Beaucoup d'effets importants ne se prêtent pas aux méthodes statistiques : un plaidoyer qui change une loi, un renforcement de capacités d'organisations locales, une coordination améliorée. On ne peut pas les comparer à un contrefactuel chiffré. On cherche alors à établir de façon crédible que le programme a **contribué** au changement.

- **Analyse de contribution** (Mayne) : on expose la théorie du changement, on rassemble les preuves pour chaque lien, on examine les explications alternatives, et l'on construit un « récit de contribution » que l'on fait critiquer par des parties prenantes.
- **Traçage de processus** (process tracing) : on teste des hypothèses causales en cherchant des preuves dont la présence ou l'absence permet de confirmer ou d'écarter une explication (par exemple, la reprise mot pour mot d'une note de plaidoyer dans un décret).
- **Récolte des effets** (outcome harvesting) : on identifie d'abord les changements observés chez les acteurs, puis on remonte pour déterminer si et comment le programme y a contribué.
- **Changement le plus significatif** : on collecte des récits de changement auprès des personnes, et des panels sélectionnent les plus significatifs en expliquant pourquoi. Utile pour découvrir des effets non prévus.

> Rigueur ne veut pas dire quantitatif. Une analyse de contribution bien menée, qui examine sérieusement les explications alternatives, est plus crédible qu'une comparaison avant-après chiffrée.

Choisir la méthode et décider ce qu'il faut prouver

Le choix d'une méthode dépend de quatre questions :

- **Quelle décision ?** Étendre un modèle pilote à l'échelle nationale justifie une preuve forte ; ajuster un programme en cours se contente d'une preuve suffisante.
- **Les preuves existent-elles déjà ?** L'effet du cash sur la consommation alimentaire est très documenté : inutile de le redémontrer avec un ECR dans chaque projet. Il suffit souvent de vérifier que le programme est fidèle au modèle et que les effets intermédiaires sont présents.

- **Qu'est-ce qui est faisable et éthique ?** Existe-t-il un groupe de comparaison sans refuser l'aide ? Un seuil d'éligibilité net ? Une mise en place progressive ?
- **Quel budget et quel délai ?** Une évaluation d'impact sérieuse exige une référence avant le démarrage : elle se conçoit avec le programme, pas à la fin.

Pour la **taille d'échantillon**, une évaluation d'impact doit avoir une puissance suffisante pour détecter l'effet attendu. Détecter un effet de 15 points demande beaucoup moins de ménages qu'un effet de 5 points. Si l'effet attendu est faible et l'échantillon petit, l'étude conclura presque sûrement « pas d'effet significatif », ce qui sera mal interprété comme « pas d'effet ».

Dernier point : présenter honnêtement le **niveau de preuve**. « Les ménages aidés ont amélioré leur consommation de 24 points ; par rapport à une zone de comparaison, l'effet attribuable est estimé à 15 points, sous l'hypothèse de tendances parallèles que les données de deux campagnes précédentes soutiennent » est une phrase qui inspire confiance.

Exercice d'application : estimer un effet et critiquer la méthode

Situation. Un programme de moyens d'existence a soutenu 1 200 ménages. Revenu mensuel médian : ménages aidés 42 USD avant, 71 USD après. Zone voisine non couverte (faute de financement) : 47 USD avant, 58 USD après. Pendant le programme, une autre ONG a distribué des semences dans la moitié des villages de la zone voisine. Le rapport final annonce : « le programme a augmenté les revenus de 69 % ».

Travail demandé. 1) D'où vient le chiffre de 69 % et pourquoi est-il trompeur ? 2) Calculez l'effet par différence de différences. 3) Quel est l'effet probable de la distribution de semences sur votre estimation, et que proposez-vous ?

Corrigé. 69 % : (71 − 42) / 42, comparaison avant-après qui attribue au programme toute la hausse, y compris celle due au contexte. Différence de différences : (71 − 42) − (58 − 47) = 29 − 11 = 18 USD par mois, soit environ 43 % du revenu initial des ménages aidés. Les semences ont probablement augmenté les revenus de la zone de comparaison, ce qui gonfle sa tendance et sous-estime l'effet du programme (la contamination viole l'hypothèse des tendances parallèles). Proposition : analyser séparément les villages de comparaison sans semences, ou contrôler cette variable ; présenter l'estimation comme une borne basse et indiquer clairement la limite.

À retenir

- L'impact se mesure par rapport à un contrefactuel ; avant-après n'est pas une mesure d'impact.
- Différence de différences = évolution du groupe aidé − évolution du groupe de comparaison, sous l'hypothèse de tendances parallèles.
- Un ECR en humanitaire est éthique seulement s'il ne prive personne d'une aide disponible : ressources limitées, mise en place progressive, ou comparaison de variantes.
- Régression sur discontinuité : comparer les ménages juste au-dessus et juste en dessous d'un seuil strict d'éligibilité.
- Quand l'attribution est impossible : analyse de contribution, traçage de processus, récolte des effets, changement le plus significatif.

Quiz

1. Groupe aidé : 30 % avant, 50 % après. Comparaison : 32 % avant, 40 % après. Effet par différence de différences :
   A. 20 points
   B. 12 points
   C. 10 points
   D. 8 points

2. Quelle hypothèse est essentielle à la différence de différences ?
   A. Les groupes sont identiques au départ
   B. Sans le programme, les deux groupes auraient évolué de la même façon
   C. L'échantillon dépasse 1 000 ménages
   D. Le programme a duré plus d'un an

3. Dans quelle situation un tirage au sort est-il éthiquement acceptable en humanitaire ?
   A. Pour refuser une aide vitale disponible à la moitié des éligibles
   B. Quand les ressources ne suffisent de toute façon pas pour tous les éligibles, ou pour l'ordre d'une mise en place progressive
   C. Jamais
   D. Seulement si le bailleur l'impose

4. La principale limite de l'appariement par score de propension est :
   A. Il ne fonctionne qu'avec des ECR
   B. Il ne corrige que les différences mesurées, pas les différences non observées comme la motivation
   C. Il exige un seuil d'éligibilité
   D. Il est interdit par l'OCDE

5. Pour démontrer qu'un plaidoyer a contribué à l'adoption d'une loi, la méthode la plus adaptée est :
   A. Un essai randomisé
   B. Une différence de différences
   C. Une analyse de contribution ou un traçage de processus
   D. Une enquête ménage

Corrigé du quiz : 1. B · 2. B · 3. B · 4. B · 5. C