

GÉNÉRALISATION DU RAPPORT DE RISQUES EN UTILISANT DES DONNÉES OBSERVATIONNELLES

Ahmed Boughdiri¹, Julie Josse² & Erwan Scornet³

¹ *Inria Premedical Inria-Idesp (Inserm - University of Montpellier) team,
ahmed.boughdiri@inria.fr*

² *Inria Premedical Inria-Idesp (Inserm - University of Montpellier) team,
julie.josse@inria.fr*

³ *Sorbonne Université and Université Paris Cité, CNRS, Laboratoire de Probabilités,
Statistique et Modélisation, F-75005 Paris, France, erwan.scornet@sorbonne-universite.fr*

Résumé. Les essais contrôlés randomisés (ECR) sont actuellement la référence pour mesurer empiriquement un effet causal d’une intervention donnée sur une réponse. En pratique, presque tous les nouveaux médicaments sont autorisés grâce à de tels essais. Mais récemment, des préoccupations ont été soulevées concernant la portée limitée des ECR : critères d’éligibilité stricts, conformité non réaliste dans le monde réel, cadre temporel court, taille d’échantillon limitée, etc. De telles limitations menacent la validité externe des études ECR dans d’autres situations ou populations. L’utilisation de données non randomisées complémentaires, appelées observationnelles, est prometteuse. La transportabilité permet de généraliser les réponses de l’essai vers une population cible d’intérêt, dont la distribution diffère par rapport à la population de référence. Cependant, certaines préoccupations ont été soulevées, en particulier lorsque la population cible est très différente de la cohorte ECR. De plus, ces méthodes sont encore à un stade de prototype, avec un grand écart entre la théorie et la pratique, et la confrontation avec les données soulève de nombreux défis méthodologiques. Des développements et validations théoriques, méthodologiques et appliqués sont nécessaires pour mieux comprendre et exploiter l’équilibre entre rapidité et sécurité, et pour concevoir les futurs ECR. La différence de risques (DR) est une mesure causale souvent utilisée pour caractériser l’effet d’un médicament. Différentes méthodes (systèmes de pondération, modélisation des réponses ou approches doublement robustes) ont été proposées et étudiées pour estimer la DR. Cependant, selon le type de réponse, d’autres mesures comme le rapport de risques (RR) peuvent être plus appropriées. Nous visons d’abord à étendre les différentes classes d’estimateurs au rapport de risques ainsi que leurs propriétés théoriques, en particulier pour des échantillons finis. Nous proposons ensuite différents estimateurs de généralisation pour le rapport de risques.

Mots-clés. Inférence causale, Essais Contrôlés Randomisés, Données observationnelles, Généralisation

Abstract. Randomized Controlled Trials (RCT) are the current gold-standard to empirically measure a causal effect of a given intervention on an outcome. In practice, almost all new drugs are authorized through such trials. But recently, concerns have been raised on the limited scope of RCTs: stringent eligibility criteria, unrealistic real-world compliance, short timeframe, limited sample size, etc. Such limitations threaten the external validity of RCT

studies to other situations or populations. The usage of complementary non-randomized data, referred to as observational brings promises as additional sources of evidence. Transportability allows to generalize the trial findings toward a target population of interest, potentially subject to a covariate distributional shift. However, some concerns have been raised, especially when the target population is very different from the RCT cohort. In addition, such methods are still in a prototype stage, with a wide gap between theory and practice and the confrontation with data give rise to many methodological challenges. Theoretical, methodological and applied developments and validations are needed to better understand and leverage the speed-safety balance and to design future RCTs. The Risk Difference (RD) is a causal measure often used to characterize the effect of a drug. Different methods (weighting schemes, outcome modeling, or doubly-robust approaches) have been proposed and studied to estimate the RD. However, depending on the type of output, other measures as the Risk Ratio (RR) may be more appropriate. We aim at first extending the different classes of estimators to the Risk Ratio and derive theoretical guarantees, in particular finite-sample guarantees. We then propose different generalizing estimators for the risk ratio.

Keywords. Causal inference, Randomized Controlled Trials, Observational data, Generalization

1 Cadre et notation

Le cadre des réponses potentiels, initié par Neyman en 1923 et largement diffusé par Rubin (Rubin, 1974) dans les années 1970, joue un rôle crucial dans la compréhension des effets des traitements dans les Essais Contrôlés Randomisés (ECR). Considérons un individu i ; soit $T_i \in \{0, 1\}$ la variable aléatoire qui désigne l’attribution du traitement, avec 1 indiquant la prise du traitement et 0 l’absence de traitement. Les réponses potentiels, notés $Y_i^{(1)}$ pour le cas où le traitement est administré et $Y_i^{(0)}$ pour le cas contraire, représentent respectivement les issues possibles avec et sans le traitement.

La réponse observé Y_i est défini comme suit :

$$Y_i = Y_i^{(T_i)} = \begin{cases} Y_i^{(0)}, & \text{si } T_i = 0 \\ Y_i^{(1)}, & \text{si } T_i = 1 \end{cases}$$

L’effet du traitement individuel, $\delta_i = Y_i^{(1)} - Y_i^{(0)}$, quantifie l’efficacité du traitement. Cependant, seul l’un des réponses potentiels peut être observé pour chaque individu. On suppose que l’on observe n individus, indexés par $i \in \mathcal{I}$. Pour chaque unité i et pour chaque niveau de traitement, il existe des réponses potentiels correspondants $Y_i^{(0)}$ et $Y_i^{(1)}$. Pour $t \in \{0, 1\}$, nous définissons les variables aléatoires $Y^{(t)}$ et T qui associent respectivement à $i \in \mathcal{I}$ $Y_i^{(t)}$ et T_i . On définit alors l’Effet Moyen du Traitement ou la différence de risques, comme l’esperance de l’effet du traitement sur la population \mathcal{I} .

Définition 1 (Effet Moyen du Traitement/différence de risques (RD)).

$$\tau_{RD} := \mathbb{E} [Y^{(1)} - Y^{(0)}]$$

Ce dernier peut être interprété comme la différence entre la réponse moyenne si toute la population avait reçu le traitement et la réponse moyenne si la population n'avait reçu le traitement. De façon similaire, nous définissons le rapport de risques comme le rapport des espérances des réponses potentielles.

Définition 2 (Rapport de risques (RR)).

$$\tau_{RR} := \frac{\mathbb{E}[Y^{(1)}]}{\mathbb{E}[Y^{(0)}]}$$

Le rapport de risques fournit une mesure de la probabilité relative qu'ont les individus du groupe exposé de subir l'événement, par rapport à ceux du groupe non exposé. Un rapport de risques de 1 indique qu'il n'y a pas d'association, un rapport de risques supérieur à 1 suggère un risque accru avec l'exposition, et un rapport de risques inférieur à 1 suggère un risque diminué avec l'exposition.

Afin d'estimer ces quantités causales et garantir l'interprétation valide de la différence de risques (RD) et du rapport de risques (RR), il est essentiel de prendre en considération les hypothèses fondamentales de l'inférence causale.

- **Hypothèse de Consistance (Consistency) :**

$$\forall (t, i) \in \{0, 1\} \times \mathcal{I}, T = t \Rightarrow Y_i = Y_i^{(t)}$$

En d'autres termes, l'assignation du traitement ne modifie pas de manière intrinsèque la réponse potentielle de l'individu, assurant ainsi que la réponse observée sous le traitement est la même que celle qui aurait été observée si, dans toutes les circonstances, l'individu avait été traité de la même manière.

- **Hypothèse d'Absence d'Interférence (No Interference) :**

$$\forall i \in \mathcal{I}, Y_i(t_1, \dots, t_i, \dots, t_n) = Y_i(t_i)$$

L'hypothèse d'absence d'interférence postule que la réponse potentielle d'un individu i , $Y_i^{(t)}$, n'est pas affectée par le traitement administré à un autre individu j .

Dans le cadre d'un traitement binaire, ces deux hypothèses peuvent se traduire par l'hypothèse **SUTVA (Stable Unit Treatment Value Assumption) :**

$$Y = TY^{(1)} + (1 - T)Y^{(0)}$$

- **Hypothèse de Non-Confondance (Unconfoundedness) :**

$$T \perp\!\!\!\perp (Y(0), Y(1)) | X$$

Une fois que l'on contrôle les variables de confusion potentielles, la décision de traiter ou non un individu ne devrait pas être liée aux réponses potentiels. Cette condition est essentielle pour établir une relation causale, car elle permet de comparer les groupes de traitement et de contrôle comme s'ils avaient été randomisés, en supposant que toutes les variables de confusion pertinentes ont été correctement mesurées et ajustées.

- **Hypothèse de Positivité (positivity) :**

$$\exists \eta \in]0; 1[, \forall x, \quad \eta \leq e(x) \leq 1 - \eta,$$

où le score de propension $e(\cdot)$ est la probabilité qu'un individu soit traité sachant ses caractéristiques :

$$e(x) = \mathbb{P}[T = 1|X = x].$$

Dit autrement, cela signifie qu'il n'existe pas de sous-populations avec des probabilités nulles ou quasi nulles de recevoir un traitement spécifique.

- **Hypothèse non-négativité (non-negativity) :**

$$\forall i, j \quad Y_i Y_j \geq 0$$

Cette dernière hypothèse contraint les réponses de tous les individus à avoir le même signe, ce qui permet ainsi au rapport de risques d'être bien défini.

2 Estimation du rapport de risques avec des données observationnelles

Sous ces hypothèses, le rapport de risques peut être identifié à partir des différentes approches suivantes issues des données observationnelles :

1. Approche de la pondération:

$$\tau_{RR} = \frac{\mathbb{E} \left[\frac{TY}{e(X)} \right]}{\mathbb{E} \left[\frac{(1-T)Y}{1-e(X)} \right]}$$

2. Approche de la régression:

$$\tau_{RR} = \frac{\mathbb{E} [\mu_1(X)]}{\mathbb{E} [\mu_0(X)]}$$

À partir de ces différentes approches et de façon similaire au cas de l'effet moyen du traitement, nous pouvons construire des estimateurs du rapport de risques. En effet, de l'approche de la pondération découle le Ratio Inverse Propensity Weighting (R-IPW).

Définition 3 (Ratio Inverse Propensity Weighting (R-IPW)).

$$\tau_{R-IPW} = \frac{\sum_{i=1}^n \frac{T_i Y_i}{\hat{e}(X_i)}}{\sum_{i=1}^n \frac{(1-T_i) Y_i}{1-\hat{e}(X_i)}}$$

avec $\hat{e}(X) = \hat{\mathbb{P}}[T = 1|X]$.

Cet estimateur s'inspire de l'estimateur IPW introduit par (Hirano, Imbens, and Ridder, 2003). Ainsi comme l'IPW, le R-IPW augmente le poids des observations traitées ayant un faible score de propension (et inversement) pour équilibrer les deux groupes, traités et non traités, en fonction de leurs covariables.

De l'approche de la regression on peut également contruire un estimateur similaire à la G-formula dans le cas de la différence.

Définition 4 (Ratio G-formula).

$$\tau_{R-G} = \frac{\sum_{i=1}^n \hat{\mu}_1(X_i)}{\sum_{i=1}^n \hat{\mu}_0(X_i)},$$

où $\mu_a(X) = \mathbb{E}[Y|T = a, X]$.

Cet estimateur, tout comme pour la différence, estime les deux réponses potentielles pour chaque individu, les moyennes ensuite, puis calcule le ratio. En plus de ces deux approches, comme détaillé dans (Laan and Rose, 2011), (Kennedy, 2015) et (Chernozhukov, Chetverikov, Demirer, Duflo, Hansen, and Newey, 2017) dans le cadre de la différence pour une estimation non-paramétrique, on peut également construire un estimateur dit double robuste.

Définition 5 (Ratio AIPW).

$$\tau_{R-AIPW} = \frac{\sum_{i=1}^n \hat{\mu}_1(X_i) + \frac{T_i(Y_i - \hat{\mu}_1(X_i))}{\hat{e}(X_i)}}{\sum_{i=1}^n \hat{\mu}_0(X_i) + \frac{(1-T_i)(Y_i - \hat{\mu}_0(X_i))}{1 - \hat{e}(X_i)}},$$

où $\hat{e}(X) = \hat{\mathbb{P}}[T = 1|X]$ and $\hat{\mu}_a(X) = \mathbb{E}[Y|T = a, X]$.

Nous démontrons la convergence en loi vers une distribution Gaussienne pour chacun de ces estimateurs, dont les variances seront explicitées.

3 Generalisation du rapport de risque

La généralisation a pour but d'élargir la portée des conclusions causales au-delà des conditions initiales de l'étude menée par essai de contrôle randomisé, permettant ainsi une application plus large et plus pertinente des réponses de recherche. Dans notre cas, on cherche à généraliser une étude ECR mené sur une population peu représentative que l'on note $\mathcal{R} = \{1, \dots, n\}$ vers une population plus représentative dite cible correspondant aux données observations $\mathcal{O} = \{n + 1, \dots, n + m\}$. On suppose de plus que seules les covariables des données observationnelles sont connues. On note $p_R(x)$ (resp $p_T(x)$) la densité des covariables dans l'ECR (resp les données observations). On suppose alors l'hypothèses suivante.

- **Hypothèse d’Inclusion sur les support (Overlap) :**

$$\text{supp}(P_T(X)) \subset \text{supp}(P_R(X)).$$

Cela implique la nécessité d’une représentation adéquate de la distribution des covariables au sein de l’échantillon de l’essai comparativement à la population cible, signifiant que chaque membre de la population cible doit avoir une probabilité non nulle d’être choisi pour l’essai. Il est également crucial d’assurer une répartition équilibrée des covariables entre les différents groupes de traitement au sein de l’échantillon de l’essai. L’attribution du traitement dans l’essai doit respecter le principe de positivité, reprenant ainsi l’hypothèse de positivité évoquée précédemment.

Une dernière hypothèse concernant le comportement des réponses entre la population cible et l’ECR est alors nécessaire afin de correctement estimer les différentes quantités causales sur la population cible. Deux approches sont alors possibles.

1. **Hypothèse de Transportabilité (Transportability) :**

$$\forall t \in \{0, 1\} \quad \mathbb{E}_R[Y(t) | X] = \mathbb{E}_T[Y(t) | X]$$

Cette hypothèse traduit le fait que le comportement des réponses d’un individu ne varie pas selon qu’il se trouve dans l’ECR ou dans la population cible.

2. **Hypothèse de Transportabilité de l’effets de traitement individuel (Transportability of the CATE) :**

$$\mathbb{E}_R[Y(1) - Y(0) | X] = \mathbb{E}_T[Y(1) - Y(0) | X]$$

Cette dernière hypothèse un peu plus faible que la Transportabilité, affirme que seul l’effet de traitement pour un individu ne varie pas selon qu’il se trouve dans l’ECR ou dans la population cible.

Sous l’hypothèse 1 (Colnet, Josse, Varoquaux, and Scornet, 2022) montre que l’estimateur plug-in G-formula suivant converge en norme L^1 vers la différence de risques au sein de la population cible.

Définition 6 (Plug-in G-formula).

$$\hat{\tau}_{g,n,m} = \frac{1}{m} \sum_{i=n+1}^{n+m} (\hat{\mu}_{1,n}(X_i) - \hat{\mu}_{0,n}(X_i)),$$

avec $\hat{\mu}_t(x) = \mathbb{E}_R[Y | X = x, T = t]$.

On montre alors une formule similaire pour le rapport de risque avec le plug-in R-G-formula.

Définition 7 (Plug-in R-G-formula).

$$\hat{\tau}_{r-g,n,m} = \frac{\sum_{i=n+1}^{n+m} \hat{\mu}_{1,n}(X_i)}{\sum_{i=n+1}^{n+m} \hat{\mu}_{0,n}(X_i)}$$

avec $\hat{\mu}_t(x) = \mathbb{E}_R[Y \mid X = x, T = t]$.

Sous l’hypothèse 2, (Colnet et al., 2022) montre que l’estimateur IPSW suivant converge en norme L^1 vers la différence de risques au sein de la population cible.

Définition 8 (IPSW: inverse propensity sampling weighting).

$$\hat{\tau}_{IPSW,n,m} = \frac{1}{n} \sum_{i \in \mathcal{R}} \frac{\hat{p}_T(X_i)}{\hat{p}_R(X_i)} \left(\frac{TY}{e_R(x)} - \frac{(1-T)Y}{1-e_R(x)} \right).$$

On montre alors une formule similaire pour le rapport de risque avec le R-IPSW:

Définition 9 (IPSW: inverse propensity sampling weighting).

$$\hat{\tau}_{R-IPSW,n,m} = \frac{\sum_{i \in \mathcal{R}} \frac{\hat{p}_T(X_i)}{\hat{p}_R(X_i)} \frac{TY}{e_R(x)}}{\sum_{i \in \mathcal{R}} \frac{\hat{p}_T(X_i)}{\hat{p}_R(X_i)} \frac{(1-T)Y}{1-e_R(x)}}.$$

References

- Chernozhukov, V., D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, and W. Newey (2017, 01). Double/debiased/neyman machine learning of treatment effects. *American Economic Review* 107.
- Colnet, B., J. Josse, G. Varoquaux, and E. Scornet (2022, 11). Causal effect on a target population: A sensitivity analysis to handle missing covariates. *Journal of Causal Inference* 10, 372–414.
- Hirano, K., G. Imbens, and G. Ridder (2003, 02). Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica* 71, 1161–1189.
- Kennedy, E. (2015, 10). Semiparametric theory and empirical processes in causal inference.
- Laan, M. and S. Rose (2011, 01). *Targeted Learning: Causal Inference for Observational and Experimental Data*.
- Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational psychology* 66(5), 688–701.