Peut-on estimer un effet de mesure sur une enquête à partir d'un essai croisé AB/BA : la question de la non-réponse non ignorable dans l'enquête test Emploi du temps

Document de travail

N° M2025-04 – Juin 2025

Loreline COURT
Simon QUANTIN

INSTITUT NATIONAL DE LA STATISTIQUE ET DES ÉTUDES ÉCONOMIQUES
Série des documents de travail « Méthodologie Statistique »
de la Direction de la Méthodologie et de la Coordination Statistique et Internationale

M 2025/04

Peut-on estimer un effet de mesure sur une enquête à partir d'un essai croisé AB/BA : la question de la non-réponse non ignorable dans l'enquête test Emploi du temps

Loreline COURT
Simon QUANTIN
Insee

JUIN 2025

Remerciements :
Les auteurs souhaitent remercier Hélène Chaput, Eric Lesage et Corinne Prost pour leur relecture attentive de ce document de travail, les participants au 13e colloque international francophone sur les sondages pour avoir par leurs échanges nourri notre réflexion ainsi que Anne Pla et Barbara Mettetal pour leur contribution à ces travaux.
Ces travaux ont partiellement été effectués dans le cadre d'un projet financé par l'Union européenne.

Direction de la méthodologie et de la coordination statistique et internationale
Département des Méthodes Statistiques -Timbre L001 -
88 Avenue Verdier - CS 70058 - 92541 Montrouge Cedex - France -
Tél. : 33 (1) 87 69 55 00 - E-mail : DG75-L001@insee.fr - Site Web Insee : http://www.insee.fr

Résumé

Parce qu'il conduit à interroger à deux reprises sous un mode de collecte différent un même enquêté, un essai croisé AB/BA apparaît comme l'outil adéquat pour mettre en évidence un effet de mesure. Cependant, la présence d'une possible sélection endogène (non-réponse non ignorable), comme c'est le cas pour l'enquête test Emploi du temps, biaise les résultats obtenus. Il est malgré tout possible de discuter l'existence d'un effet de mesure en comparant les réponses données en première interrogation par des personnes similaires enquêtées avec un mode de collecte différent. Pour cela, il est nécessaire de mener une analyse de sensibilité à la Rosenbaum sur les conclusions obtenues. En ce qui concerne l'enquête test Emploi du temps, nous trouvons que (i) la collecte par internet impacte les durées déclarées de sommeil, de repas, de loisirs, (ii) la collecte papier celle des trajets, mais (iii) seulement pour un nombre restreint d'enquêtés. Ces résultats robustes à une possible sélection endogène confirment ceux obtenus par l'analyse de l'essai croisé en supposant de prime abord la non-réponse comme ignorable.

Mots clés : Enquêtes multimodes, essai croisé, sélection endogène, analyse de sensibilité

Classification JEL : C12, C14, C21, C83

Abstract

Because it involves interviewing the same respondent twice, using a different collection method, an AB/BA crossover trial appears to be an appropriate tool for highlighting a measurement effect. However, the presence of a possible endogenous selection (non-ignorable non-response), as is the case for the Time Use test survey, biases the results obtained. It is nevertheless possible to discuss the existence of a measurement effect by comparing the responses given in the first interrogation by similar people surveyed using a different collection method. To do this, it is necessary to carry out a Rosenbaum sensitivity analysis on the conclusions obtained. With regard to the Time Use test survey, we found that (i) Internet data collection had an impact on the declared duration of sleep, meals and leisure activities, (ii) paper data collection had an impact on the declared duration of journeys, but in each case (iii) only for a limited number of respondents. These results, which are robust to a possible endogenous selection, confirm those obtained by the crossover analysis, which initially assumed ignorable nonresponse.

Keywords. Mixed-mode surveys, crossover trial, endogenous selection, sensitivity analysis

Classification JEL : C12, C14, C21, C83

1 Introduction

L'enquête Emploi du temps est une enquête auprès des ménages visant à collecter les emplois du temps des personnes pour mener des analyses quantifiées du temps passé au travail, aux tâches domestiques, aux loisirs, etc. Un test méthodologique préalable à l'introduction d'une collecte multimode pour cette enquête a été mis en place en 2023 pour questionner l'existence d'éventuels effets de mesure liés à la collecte par internet. Usuellement, dans ce cas de figure, le protocole retenu attribue aléatoirement un mode de collecte à deux échantillons représentatifs de la population étudiée afin de comparer, après appariement, les réponses données par les enquêtés répondants. Le protocole du test méthodologique mis en œuvre par l'Insee est différent : il s'agit d'un essai-croisé AB/BA sur un échantillon d'habitants issus de 2 100 logements. Chaque personne détaille son emploi du temps d'une journée sur le carnet papier et sur le carnet numérique, à une semaine d'intervalle, l'ordre de la passation (carnet numérique puis papier ou l'inverse) étant affecté aléatoirement en amont de l'enquête. Ainsi, contrairement à l'analyse de données appariées, les résultats sous différents modes de collecte sont enregistrés *pour la même personne*[1].

Cependant, comme dans toute enquête, une personne interrogée peut ne pas participer à la totalité du test ou ne répondre qu'avec l'un des deux modes de collecte. La plupart des études exploitant des données issues d'un essai croisé pour lesquelles des observations sont manquantes supposent que la non-réponse est ignorable (MAR, *Missing At Random*), c'est-à-dire que le comportement de réponse des enquêtés ne dépend que des caractéristiques observées (voir par exemple Patel, 1985)[2]. Sous cette hypothèse, une estimation sans biais d'un effet de mesure du mode de collecte est possible. Cela n'est plus le cas dès lors que le comportement de réponse des répondants et des non-répondants diffèrent selon une caractéristique inobservée, c'est-à-dire qu'il existe une possible sélection endogène (ou non-réponse non-ignorable dite MNAR, *Missing Not At Random*).

Dans cette étude, nous questionnons justement l'impact d'une possible sélection endogène sur les conclusions obtenues, sous l'hypothèse de non-réponse ignorable, quant à la présence d'un effet de mesure dans l'enquête test Emploi du temps. Pour cela, nous comparons tout d'abord les résultats obtenus par l'analyse classique de l'essai croisé avec ceux issus d'une analyse complémentaire des (seules) réponses données en première période par des enquêtés similaires, i.e. appariés sur des caractéristiques observées, mais interrogés avec des modes différents. Il s'agit dans cette étape de montrer que les deux analyses fournissent des résultats similaires et non statistiquement différents. L'approche par appariement suppose l'absence de caractéristiques inobservées différentes entre les personnes appariées : elle suppose donc, comme l'analyse par essai croisé, l'absence d'un biais de sélection endogène. Néanmoins, avec cette démarche, il est possible de tenir compte de l'impact de cette éventuelle sélection endogène sur les résultats obtenus, en menant une analyse de sensibilité à la Rosenbaum (Rosenbaum, 2002b, 2010). Plus précisément, l'analyse de sensibilité permet de quantifier l'ampleur du biais de sélection endogène nécessaire pour infirmer les conclusions obtenues sur l'existence possible d'un effet de mesure, et donc discuter de l'impact de la non-réponse non ignorable sur les conclusions de l'analyse de l'essai croisé mis en œuvre dans l'enquête test Emploi du temps. Nous trouvons que la collecte par internet impacte les durées déclarées de sommeil, de repas, et de loisirs, la collecte papier

1. Cet avantage d'un essai croisé peut être contrebalancé par des inconvénients qui ne sont pas présents dans l'analyse de données appariées, comme la nécessaire attention à la durée de la période entre les deux modes de collecte, un possible effet de mesure du mode de collecte précédent sur les réponses futures (*carryover effect*), etc. Néanmoins, ceux-ci peuvent être réduits en adoptant un *design* approprié.

2. Certaines ne font pas cette hypothèse, mais s'appuient sur l'absence de non-réponse en première interrogation (Ho *et al.*, 2012; Basu et Santra, 2010), ce qui n'est pas le cas pour l'enquête test Emploi du temps.

celle des trajets, mais dans chaque cas pour un nombre restreint d'enquêtés. Ces résultats sont robustes à une possible sélection endogène et confirment ainsi l'existence d'un effet de mesure déjà mis en évidence par l'analyse de l'essai croisé en supposant les données manquantes par hasard.

Le document de travail est structuré comme suit. La partie 2 présente les résultats sous l'hypothèse d'absence de sélection endogène obtenus à partir (i) de l'analyse de l'essai croisé mené pour l'enquête test Emploi du temps et (ii) de l'approche par appariement réalisée en comparant les seules réponses données en première période par des personnes similaires mais enquêtées sous des modes de collecte différents. Si les résultats obtenus par les deux analyses sont cohérents, nous mettons en évidence qu'elles sont toutes deux impactées par un biais de sélection endogène similaire, en nous appuyant sur l'étude des réponses données à une question de l'enquête (supposée) non impactée par un effet de mesure. La partie 3, plus méthodologique, présente le modèle d'analyse de sensibilité avec des variables continues, dans le cas de paires appariées, en précisant notamment son implémentation avec différentes statistiques de test. Cette partie explicite aussi une interprétation du paramètre de sensibilité du modèle en termes de corrélation d'une caractéristique inobservée avec le mode de collecte et la réponse donnée. Pour conclure, elle introduit la notion d'effets attribuables au mode de collecte et détaille comment les estimer dans le cas d'une variable d'intérêt binaire. Enfin, la partie 4 présente les résultats de l'analyse de sensibilité obtenus sur quatre durées d'activités : la durée de sommeil, la durée de trajet, le temps consacré aux loisirs et le temps consacré aux repas. Si l'analyse de sensibilité nous permet de conclure sur l'existence d'un effet de mesure du mode de collecte en présence de sélection endogène pour chacune de ces durées, des pistes d'explication en lien avec l'application numérique ou la codification des carnets papier sont proposées.

2 Résultats sous l'hypothèse d'absence de sélection endogène

Le protocole mis en place pour le test d'éventuels effets de mesure du mode de collecte sur les résultats de l'enquête Emploi du temps est un essai croisé, ou une enquête par test-retest, aussi appelé *crossover design*. Un essai croisé se distingue d'un test classique en cela que chaque unité enquêtée se voit affectée aléatoirement non pas à un seul mode de collecte, mais à une séquence de modes de collecte. La justification d'un tel *design* d'enquête est la suivante. Supposons que l'on souhaite comparer l'impact de deux modes de collecte A et B sur une même unité enquêtée. Une approche possible est de demander que le questionnaire de l'enquête soit d'abord remplie avec le mode de collecte A, puis avec le mode de collecte B par l'unité enquêtée. Un tel protocole est cependant potentiellement biaisé, parce que les réponses données peuvent refléter un impact lié au temps écoulé entre les deux interrogations, sans lien avec le mode de collecte considéré.

Afin de surmonter cette difficulté et pouvoir identifier un effet de mesure éventuel, un ordre différent des modes de collecte utilisés pour remplir les questionnaires est attribué aléatoirement à deux sous-échantillons d'unités enquêtées, chacun étant représentatif de la population étudiée. Dans le cas de l'enquête Test-emploi du temps, les unités du premier sous-échantillon (dénommé dans ce qui suit AB) répondent d'abord au questionnaire par internet, puis 7 jours plus tard par papier ; les unités du second sous-échantillon BA répondent au questionnaire d'abord par papier avant de répondre au questionnaire par internet, 7 jours plus tard [3]. Parce qu'il comporte deux périodes d'interrogation (et deux

3. Il était préconisé aux enquêteurs de suggérer que le choix du jour de collecte par l'enquêté, pour les deux interrogations, soit un jour ouvré. Quel que soit le jour retenu, il convient de noter qu'il peut être attendu que les activités effectuées lors de la première interrogation soient sensiblement similaires à celles effectuées lors de la deuxième interrogation et que l'impact du temps écoulé entre les deux interrogations

modes de collecte) ce protocole est appelé un essai croisé AB/BA [4]. Dans cette section, nous décrivons sommairement les deux approches méthodologiques usuelles pour analyser un essai croisé AB/BA. Comme nous le verrons, la pertinence des résultats obtenus avec chaque méthode repose sur l'hypothèse d'absence de sélection endogène.

2.1 Approches méthodologiques pour analyser un essai croisé AB/BA

Dans un essai croisé AB/BA, contrairement aux enquêtes test usuelles, les réponses sous différents modes de collecte sont donc enregistrées pour la même personne : la comparaison des réponses entre modes de collecte différents est donc intra-individuelle (*within-subjects analysis*). De ce point de vue, un tel protocole permet théoriquement d'estimer un effet de mesure avec une meilleure précision, à taille d'échantillon donnée, ce qui permet d'implémenter un test à moindre coût. En effet, pour une grandeur donnée, la variance intra-individuelle est souvent plus faible que la variance inter-individuelle qui résulterait par exemple d'un test comparant après affectation aléatoire à un mode de collecte, les réponses d'enquêtés différents (*between-subjects analysis*). Cet avantage est cependant contrebalancé par une exploitation plus difficile des résultats du test. Le premier inconvénient est que les réponses données lors de la seconde interrogation peuvent refléter un impact persistant du premier mode de collecte : on parle alors de *carryover effect*. Dans notre étude, par exemple, les réponses à la deuxième interrogation peuvent refléter un effet d'apprentissage lié au mode de collecte utilisé la première fois. En effet, le remplissage du questionnaire de l'enquête par internet s'appuie sur des listes d'activités et un suggesteur. Leur utilisation en première interrogation peut impacter la nature des activités déclarées en deuxième interrogation lors du remplissage du questionnaire papier, permettant par exemple une codification plus précise que celle issue d'un champ libre « non-entraîné », comme c'est le cas lors d'une première interrogation via le questionnaire papier. De même, les réponses données peuvent refléter un effet du mode de collecte différent entre la première et la deuxième interrogation (*treatment by period interaction*). Dans notre étude, par exemple, un tel effet pourrait être occasionné par la lassitude en deuxième interrogation de l'enquêté, le conduisant à ne pas saisir toutes les plages horaires et les activités associées avec la même attention qu'en première période [5]. Usuellement, de tels effets sont supposés neutralisés en amont par le *design* de l'essai croisé avec le respect d'une période de latence suffisamment longue entre les deux interrogations, dite de *washout*, qui est, comme nous l'avons évoqué, de 7 jours dans le test pour l'enquête Emploi du temps. *Dans cette étude, nous ferons cette hypothèse.*

Une autre difficulté concerne la participation des enquêtés à l'intégralité du test. La non-réponse ne peut en effet être exclue d'un essai croisé, comme pour tout test avec une enquête. Une personne enquêtée peut ne pas participer à la totalité du test pour de nombreuses raisons, comme son expérience directe du mode de collecte une journée donnée, mais aussi sa comparaison avec le mode de collecte proposé plus tôt dans l'essai, et leurs liens avec la grandeur d'intérêt. Cela complexifie grandement l'analyse, même en présence d'un *design* approprié permettant d'exclure les hypothèses d'un effet de *carry-over* ou d'un effet du mode de collecte différent entre la première et la deuxième interrogation. Cette difficulté dans l'interprétation des résultats est l'objet principal - du point de vue méthodologique - de notre étude.

Un modèle usuel d'analyse d'un essai croisé AB/BA (*within-subjects*)

soit nul.

4. Une telle appellation permet de le différencier d'un autre protocole d'essai croisé qui ferait intervenir aussi les séquences de traitement AA et BB.

5. Avec deux périodes, il n'est pas possible d'identifier séparément ces deux effets.

On suppose que la réponse donnée par un enquêté $i, i = 1, \ldots, n$ à la période $j = 1{,}2$ peut être représentée par une variable aléatoire Y_{ij}. Par souci d'alléger les notations, on note $t = t(i,j)$ le mode de collecte alloué à l'enquêté i à la période j. Un modèle largement utilisé dans l'analyse d'un essai croisé AB/BA peut s'écrire :

$$Y_{ij} = \mu + \pi_j + \tau_t + (\tau\pi)_{tj} + \xi_i + \epsilon_{ij} \tag{1}$$

où μ est une moyenne, π_j est l'effet de la période j, τ_t est l'effet de mesure du mode de collecte t et $(\tau\pi)_{tj}$ leur interaction. Classiquement, pour assurer l'identification des paramètres, on suppose que $\pi = \pi_1 = -\pi_2$, $\tau = \tau_A = -\tau_B$, et $(\tau\pi)_{t1} = 0, (\tau\pi)_{B2} = (\tau\pi) = -(\tau\pi)_{A2}$. ξ_i est un effet indépendant (*random effect*) associé à la personne enquêtée.

Usuellement, $(\tau\pi)$ est souvent identifié comme l'effet de *carryover*. Comme nous l'avons dit, on supposera, pour la suite, que le *design* du test permet de le considérer comme nul suite à l'intervalle de temps imposé entre les deux interrogations.

Ce modèle à effets aléatoires individuels peut être estimé en s'appuyant sur l'économétrie des panels. Si, pour un individu i, un des deux carnets (Y_{i1}, Y_{i2}) n'est pas complété, il est toujours possible de l'inclure dans l'analyse. Cette approche n'est valide, du point de vue théorique, que dans deux cas (à notre connaissance) : (i) si les observations manquantes sont manquantes complètement aléatoirement (MCAR) ou manquantes aléatoirement (MAR), selon la classification proposée par Rubin (1976), ou (ii) si la non-réponse est endogène (MNAR) mais dans ce cas, seulement si seules les réponses données en deuxième période sont manquantes (sur cette approche, voir Ho *et al.*, 2012).

Dans l'enquête test Emploi du temps, parmi les 2 100 ménages enquêtés, 29 % des ménages (i.e. 27,5 % pour l'échantillon papier/internet et 29,8 % pour l'échantillon internet-papier) pour un total de 1 070 personnes ont accepté de participer à l'enquête test - sans présager pour autant de leur complétion future des deux questionnaires. Parmi ces personnes, 1 045 ont pu effectivement prendre part au test car elles disposaient d'un accès internet (529 dans le lot internet/papier et 516 dans le lot papier/internet). Dans toute cette étude, nous considérerons que le refus de participer au test ne dépend pas du sous-échantillon [6], de telle sorte que l'on peut toujours supposer l'affectation à une séquence de modes de collecte comme aléatoire entre les deux sous-échantillons de participants [7].

Néanmoins, contrairement à ce qui était prévu par le protocole, toutes les personnes ayant accepté de participer n'ont pas rempli les deux carnets, certaines n'ayant finalement complété aucun carnet ou seulement l'un des deux. Ainsi à l'issue du test, si 77 % des enquêtés papier/internet ont répondu au questionnaire en première période, ils ne sont que 54 % à répondre avec les deux modes de collecte, soit 23 points de pourcentage en moins (cf. tableau 1). Par ailleurs, si 64 % des enquêtés internet/papier ont répondu au questionnaire en première période, ils sont 57 % à répondre avec les deux modes de collecte, soit 7 points de pourcentage en moins.

Dans notre essai croisé, la non-réponse lors de la première interrogation est de 23 % dans l'échantillon papier-internet et de 36 % dans l'échantillon internet-papier. Cela exclut donc une approche, telle que suggérée par Ho *et al.* (2012), autorisant une non-réponse MNAR. De plus, la non-réponse en deuxième interrogation plus élevée lorsque la personne enquêtée a répondu au questionnaire papier en première interrogation soulève la question d'une possible sélection endogène. Par exemple, les enquêtés ayant au cours d'une même journée un nombre élevé d'activités de nature différentes (et donc un nombre de plages horaires

6. La séquence des modes de collecte n'est alors pas connue des personnes enquêtées

7. Les deux sous-échantillons ne sont donc plus supposés représentatifs de la population originale, mais d'une sous-population qui est cependant similaire entre les deux sous-échantillons.

Tableau 1 – Types de carnets remplis par les enquêtés participants par lot (en %)

	Papier-Internet (516 ind.)	Internet-Papier (529 ind.)	Ensemble (1 045 ind.)
Aucun carnet	16	18	17
Carnet internet manquant	23	18	20
Carnet papier manquant	7	7	7
Deux carnets complétés	54	57	56

Champ : participants ayant accès à Internet, France.
Note : La participation au test est définie lorsque le carnet rempli est exploitable (au moins 4 plages horaires renseignées, ce qui correspondait par exemple au cas d'une personne malade alitée ayant participé au test.). Les lots considérés sont les lots effectifs observés et non les lots assignés : un nombre très faible de personnes enquêtées n'ayant pas respecté l'ordre des modes de collecte assigné.

à saisir important) peuvent avoir décidé par lassitude de ne pas participer en deuxième interrogation, et ce d'autant plus si la saisie des carnets papiers a été jugée chronophage. Dès lors, en l'absence de solution théorique pour l'estimation des paramètres en présence de sélection endogène, la pertinence des résultats de l'analyse de l'essai croisé doit être questionnée.

Analyse par appariement (*between-subjects*)

Il est aussi possible d'exploiter un essai-croisé en se restreignant aux seuls répondants en première période. Il s'agit dans cette analyse de comparer les réponses données par des répondants différents sur des modes de collecte différents. Ces répondants seront cependant jugés « similaires », car appariés sur des caractéristiques observées. L'analyse de paires appariées suppose en cela l'absence de différence sur une caractéristique inobservée entre ces personnes appariées, c'est-à-dire l'absence d'un biais de composition caché. Puisque que l'ordre et donc les modes de collecte sont affectés aléatoirement aux enquêtés en amont de leur interrogation, ce possible biais de composition caché résulterait d'une possible sélection endogène des répondants à l'enquête. Ainsi, dans les deux analyses, l'hypothèse d'un possible biais de sélection endogène (non-réponse non-ignorable) fragilise les résultats.

Qualité de l'appariement optimal par paire

Dans ce qui suit nous détaillons l'appariement effectué dans cette étude. L'objectif est de souligner la similarité des personnes enquêtées au sein d'une paire avant de comparer les résultats obtenus avec les deux méthodes décrites précédemment.

L'appariement classique sur le score de propension tend à constituer des groupes traité et de contrôle qui présentent des distributions similaires pour les variables observées. Cependant, si les unités enquêtées au sein de chaque paire ainsi constituée ont un score de propension (estimé) proche, elles peuvent néanmoins différer fortement sur des covariables spécifiques. Dans cette étude, afin de constituer des paires d'unités enquêtées plus semblables, nous construisons une distance qui pénalise les différences importantes de caractéristiques observables, puis nous constituons des paires d'unités aussi similaires que possibles, en utilisant un algorithme d'optimisation. Une description précise de l'appariement mis en œuvre est fournie en Annexe A. Sans entrer dans les détails, relevons cependant deux points importants.

Les durées des différentes activités déclarées par une unité enquêtée dépendent du jour de la semaine où le questionnaire est rempli (week-end ou non) et de sa situation principale (en

emploi, au chômage [8], retraité ou préretraité, en incapacité de travailler, en études, au foyer ou autre situation). Il est donc essentiel de pouvoir s'assurer qu'au sein de chaque paire, les unités enquêtées soient identiques sur ces caractéristiques observées. En ce qui concerne le jour de la semaine, un appariement exact est imposé. Pour tenir compte de la situation principale, une pénalité est prise en compte dans la distance évoquée précédemment. Si deux unités enquêtées k et l n'ont pas la même situation principale, la fonction de pénalité utilisée ajoute à la distance initiale 10 fois la distance maximale observée (sur la distance préalablement calculée) si les deux unités diffèrent sur leur situation principale. Cette pénalité est prise en compte comme suit par l'algorithme d'optimisation : si un appariement exact est possible, il sera considéré ; sinon, un appariement aussi proche que possible d'un appariement exact sera effectué. Enfin, nous incluons également les variables socio-démographiques usuelles disponibles telles que : le sexe, l'âge, le type de logement, le type de ménage et la structure de ses revenus en 2021. Au total, notre appariement se fonde sur 63 covariables.

Tableau 2 – Qualité de l'appariement (% de paires avec appariement exact) sur quelques caractéristiques observées

Caractéristiques du ménage	
Logement	
Maison	67
Quartier prioritaire	95
HLM	83
Propriétaire	67
Structure des revenus du ménage en 2021	
avec au moins une allocation chômage	63
avec au moins une pension retraite	87
nombre d'habitants avec des revenus salariés	54
Type de ménage en 2021	
Être ou non un salarié seul	84
Être ou non un retraité seul	96
Être ou non un couple de retraités	95
Être ou non un couple de salariés sans enfant	85
Être ou non un couple de salariés avec enfant(s)	69
Caractéristiques individuelles	
Sexe	76
Âge	8
De nationalité française	96
Situation principale (emploi, études, chômage, retraité, etc.)	98
Situation matrimoniale (marié, célibataire, concubinage, etc.)	71
Questionnaire rempli un jour ouvré	100

Le tableau 2 vise à illustrer la qualité de l'appariement ainsi obtenu en précisant pour chaque caractéristique observée le pourcentage de paires où l'appariement est exact. Comme attendu, après appariement, 100 % des personnes ayant répondu par internet un jour de semaine sont appariées avec une personne ayant aussi répondu au questionnaire papier un jour de semaine. De même, il est intéressant de souligner qu'en ce qui concerne la

8. inscrit ou non à France Travail (ex. Pôle Emploi).

situation principale qui distingue les personnes en emploi, qui poursuivent des études, sont au chômage, retraité ou invalide, 98 % des paires sont constituées d'un répondant par internet à la situation similaire au répondant par papier auquel il est apparié. La qualité de l'appariement sur cette caractéristique était importante tant la situation principale de la personne enquêtée est corrélée à la nature et la durée des activités déclarées. Plus encore, 100 % des personnes en emploi ayant répondu par internet sont appariées avec un répondant au questionnaire papier lui aussi en emploi ; il nous sera donc possible de concentrer notre analyse sur la durée du travail sur les seules personnes en emploi en nous appuyant dans ce cas de figure sur un appariement exact [9]. De manière générale, quelle que soit la covariable, au moins 63 % des paires sont composées de personnes enquêtées identiques du point de vue de la caractéristique considérée.

En résumé ces quelques résultats montrent qu'après appariement, au sein de chaque paire, les unités répondant par internet et celles répondant par papier ne diffèrent pas plus que ce qui aurait été attendu si le mode de collecte avait été affecté aléatoirement (voir Annexe B pour une présentation plus formelle de cette assertion).

2.2 En supposant l'absence de sélection endogène, des résultats cohérents

Dans cette partie nous présentons les estimations obtenues avec l'approche par essai croisé et par appariement, i.e. **sous l'hypothèse d'absence de biais de sélection endogène** [10], de l'effet de mesure de la collecte par internet sur les durées de différents types d'activités. Avant de discuter ces résultats, il est néanmoins utile d'expliciter comment l'ensemble des activités quotidiennes sont regroupées dans les différents agrégats étudiés. Certaines activités (sommeil, repas, loisirs et trajets) feront l'objet d'une analyse détaillée par la suite. Savoir dans quels agrégats elles se trouvent éclairera ainsi les résultats détaillés obtenus.

Parmi les agrégats présentés, le **temps physiologique et personnel** regroupe la *durée de sommeil*, le *temps consacré aux repas* (mais pas à leur préparation), à l'hygiène ainsi qu'aux soins personnels et médicaux. La **durée de travail et d'études** regroupe les heures consacrées à son activité professionnelle pour les personnes enquêtées en emploi, les heures de formation (scolaire, professionnelle ou relatives à des activités hors pratiques sportives et artistiques [11]) ainsi que les temps de pause entre ces activités, celui dédié à des activités syndicales ou de recherche d'emploi. La **durée des travaux ménagers et des loisirs récréatifs** totalise le temps consacré à toutes les tâches ménagères au sein du domicile, mais aussi aux achats de biens de consommation et aux activités personnelles de création artistique. S'il inclut le temps dédié à son animal domestique, il ne comporte pas celui consacré aux enfants qui est totalisé dans le temps consacré à **s'occuper d'autres personnes** du ménage ou d'un autre ménage. Le temps de **sociabilité** correspond principalement à celui passé à recevoir des amis, discuter ou téléphoner. Les activités de ***loisirs*** recensent celles associées à une pratique culturelle (lecture, télévision, spectacle, etc.) ou sportive, ainsi que les promenades et randonnées ; ces deux dernières activités ne sont donc pas comptabilisées comme des ***trajets***.

Toutes ces durées constituent une partition de la durée totale agrégée déclarée. La durée de la période couverte correspond, elle, au temps qui sépare l'heure de début de la première plage horaire déclarée (normalement 4 heures du matin, le jour de l'enquête) à l'heure de fin de la dernière plage horaire déclarée (normalement 4 heures du matin, le jour suivant). La durée totale d'activités issue de l'agrégation des temps déclarés et la durée de la période

9. Notre appariement assure aussi que 100 % des personnes ayant répondu par internet qui poursuivent des études ou sont au chômage sont également appariées avec un répondant papier dans une situation similaire.

10. Ou de manière similaire l'absence de biais de composition caché

11. Comme les cours de conduite, de cuisine, etc.

couverte peuvent donc différer dès lors que les activités effectuées au cours de certaines plages horaires ne sont pas déclarées par l'enquêté.

Le tableau 3 détaille, pour chaque durée, les estimations (i) d'un effet de mesure de la collecte par internet additif et constant obtenu avec l'analyse des données issues de l'essai croisé et (ii) d'un effet de mesure moyen obtenu avec l'analyse des seuls emplois du temps déclarés en première période après appariement d'enquêtés ayant répondu avec des modes de collecte différents. Ainsi, en supposant l'absence de sélection endogène dans la participation à l'essai croisé [12], détailler son emploi du temps avec le questionnaire internet réduirait, par exemple, la durée totale de trajet déclarée de 20 minutes, avec un intervalle de confiance (IC) à 95 % égal à [-27 ; -14]. De même, si l'on suppose l'absence de biais de composition après appariement, déclarer son emploi du temps par internet réduirait de 32 minutes la durée totale de trajet comptabilisée sur la journée (IC à 95 % [-43 ; -20]).

Tableau 3 – Estimations de l'effet de mesure de la collecte par internet sur différentes durées

	Essai croisé Effet constant	Appariement Effet moyen
Temps physiologique et personnel	−1 [−20; 18]	−5 [−36; 25]
Travail et études	−8 [−24; 8]	−22 [−50; 7]
Travaux ménagers et loisirs récréatifs	−7 [−16; 3]	6 [−11; 23]
S'occuper d'autres personnes	4 [0;8]	4 [−4;11]
Sociabilité	25 [17;33]	27 [14;40]
Loisirs	−51 [−64;−38]	−45 [−67;−24]
Trajet	−20 [−27;−14]	−32 [−43;−20]
Durée totale agrégée	−53 [−73;−34]	−67 [−102;−32]
Période couverte	−27 [−44;−11]	−44 [−73;−14]

Note : Les estimations de l'essai croisé reposent sur les durées déclarées par 866 personnes ayant complété au moins un carnet, erreurs standards robustes à une corrélation sérielle et à l'hétéroscédasticité (General Feasible Generalized Least Squares Analysis Wooldridge, 2002, chapitre 10). L'analyse par appariement repose sur 311 paires. Comme attendu, la variance des estimations est plus faible dans l'analyse de l'essai croisé et conduit à des intervalles de confiance réduits. Effets statistiquement significatifs à 95 % en gris.

De manière générale, quelle que soit la durée considérée, les deux méthodes produisent des résultats semblables [13]. Sous l'hypothèse d'absence de sélection endogène, nos résultats mettent en évidence que répondre par internet aurait un impact significatif sur la durée de sociabilité, de loisirs et de trajets déclarée (et la durée totale des activités). Par ailleurs, avec le questionnaire internet, les emplois du temps ne seraient pas complétés en intégralité, de telle sorte que la période couverte serait inférieure à celle que l'on aurait observée si

12. Pour rappel, nous supposons aussi l'absence d'effet de *carryover* ou d'interaction d'un effet période avec un effet du mode de collecte.

13. Avec comme attendu, une variance des estimateurs plus petite dans l'analyse de l'essai croisé, menant à des intervalles de confiance plus resserrés.

l'enquêté avait utilisé un questionnaire papier.

Pour autant, ces résultats cohérents sont-ils pertinents ? Si la non-réponse est endogène ou qu'après appariement les personnes au sein d'une même paire diffèrent sur une caractéristique inobservée, les estimations obtenues avec ces deux approches sont toutes deux biaisées. Dans un premier temps, il est donc nécessaire de discuter l'existence possible d'un biais de sélection endogène pour l'analyse par essai croisé ou d'un biais de composition (issu d'une possible sélection endogène) pour l'approche par appariement. Le cas échéant, cela justifiera de mener, dans un second temps, une analyse de sensibilité sur les conclusions obtenues.

2.3 Un biais de sélection endogène est-il possible ?

Il peut être utile avant de chercher à mettre en évidence un possible biais de sélection endogène ou de composition d'illustrer par un exemple le(s) problème(s) auquel(auxquels) on se trouve confronté. Pour cela, considérons deux étudiants du même âge, de la même nationalité, vivant dans le même type de ménage, etc. Le répondant papier déclare une durée de sommeil plus longue que le répondant internet. Plusieurs explications sont possibles.

Tout d'abord, le répondant par internet peut avoir simplement mal dormi : la différence de durée observée est alors due au hasard, c'est-à-dire à une caractéristique inobservée non corrélée avec le mode de collecte et avec la participation. Il peut aussi s'agir d'un effet de mesure de la collecte par internet : les deux répondants sont rigoureusement identiques, mais le répondant par internet n'a pas détaillé tout son emploi du temps, par exemple, la plage horaire de 1 heure du matin à 4 heures du matin car il ne s'est pas connecté. Enfin, l'un des deux peut avoir un « travail étudiant » le soir qui réduit de fait sa durée de sommeil : la différence est cette fois due à une caractéristique inobservée corrélée avec la grandeur mesurée. Cette caractéristique inobservée peut aussi être corrélée avec le mode de collecte et la participation, par exemple si les étudiants ayant accepté de participer avec le carnet papier ont moins fréquemment un travail à côté que les répondants utilisant le carnet numérique.

Dans ce dernier cas de figure, comparer les réponses des deux étudiants après appariement ne permet pas de mettre en évidence un effet causal du mode de collecte sur la durée de sommeil déclarée. Mais il convient de souligner que cela n'est pas possible non plus dans l'approche par essai croisé, alors même que dans ce type de test, les enquêtés sont « leur propre contrôle ». En effet, en intégrant les réponses données par les enquêtés qui n'ont complété leur emploi du temps qu'avec un seul mode de collecte, la non-participation n'est plus indépendante de la grandeur étudiée.

Dès lors, quelle que soit l'approche retenue, la difficulté pour l'analyse est que nous ne savons pas quelle situation est reflétée par les données. Il peut néanmoins être envisagé de tirer bénéfice d'une question supposée non impactée par un effet de mesure pour mettre en évidence l'existence d'un possible biais de composition (après appariement) ou de sélection (pour l'essai croisé).

Une réponse supposée non affectée par le mode de collecte

Si l'emploi du temps détaillé par l'enquêté avec le questionnaire internet ne couvre pas la période attendue de 24 heures, il est cohérent d'anticiper que les durées de certaines activités soient inférieures à celles qui auraient été constatées avec le questionnaire papier. La cohérence évoquée dans l'assertion précédente fait référence à la possibilité que le mode de collecte affecte différentes réponses dans une direction connue [14]. À l'inverse, il est

14. Cette cohérence peut être exploitée pour accroître l'insensibilité des conclusions à la présence d'un

possible que l'on pense que le mode de collecte affectera la réponse à une question, mais n'en affectera pas une autre et que nous souhaitions exploiter l'absence anticipée d'effet pour fournir des informations sur l'existence possible d'un biais de sélection endogène (ou de composition caché).

Pour être utile, il nous faut disposer d'une question qui a une forte probabilité de ne pas être affectée par un effet de mesure et qui soit corrélée à une caractéristique inobservée. Dans l'enquête test Emploi du temps, il est demandé à l'enquêté, à la fin du carnet papier et du carnet numérique, son estimation de la durée totale des trajets effectués lors de la journée détaillée. Cette durée, qui ne résulte pas d'une agrégation des temps de trajets déclarés, exclut un impact du *design* du questionnaire numérique, de la codification post-collecte des activités du questionnaire papier, de la non-complétion de la totalité du carnet, etc. Il peut, en ce sens, être possible de supposer que la réponse à cette question ne soit pas impactée par le mode de collecte utilisé.

La durée totale des trajets effectués lors de la journée déclarée à la fin du questionnaire est cependant vraisemblablement fortement corrélée à une variable d'intérêt de notre analyse : la durée de trajets calculée par agrégation des temps correspondant déclarés séparément dans les carnets complétés. Elle est aussi associée de façon plausible à des caractéristiques inobservées comme l'existence d'un emploi (pour les étudiants), éventuellement aussi très éloigné du domicile. Dès lors, la durée totale des trajets effectués lors de la journée et déclarée par l'enquêté à la fin du questionnaire peut être utile pour distinguer un effet de mesure du mode de collecte utilisé (i) d'un biais de sélection endogène dans l'approche par essai croisé et dans l'approche par appariement.

Il est possible de démontrer qu'une réponse non affectée par un effet de mesure peut fournir un test consistant et non biaisé [15] à la présence d'une caractéristique inobservée à laquelle elle est corrélée (voir Rosenbaum (1989b,a) et Rosenbaum (2002b), §6). Il convient néanmoins de garder à l'esprit qu'un biais sur une réponse non affectée ne peut refléter qu'une partie d'une différence plus importante sur une caractéristique inobservée. Plus précisément si un tel biais résulte de différences sur une caractéristique inobservée, alors ces différences sont toujours au moins aussi larges que le biais lui-même (voir Rosenbaum, 1989b, pour une discussion plus formelle) [16].

Une participation corrélée à l'importance des trajets effectués la journée

biais de sélection endogène dans une analyse de sensibilité : cela nécessite formellement de tenir compte de la multiplicité des tests à partir d'une même base de données à la manière de la correction de Bonferroni, voir Rosenbaum (2002b), §9, et Rosenbaum (2020), §18).

15. Il peut être utile de rappeler ici ce que l'on entend par test consistant et sans biais, des notions peut-être moins connues que leurs homologues pour un estimateur. Un test d'une hypothèse nulle H_0 contre une hypothèse alternative H_a est *consistant* si le test « fonctionne » lorsque la taille des échantillons est suffisamment élevée. Un test est *sans biais* s'il est « orienté dans la bonne direction » quelle que soit la taille de l'échantillon. Plus formellement, ces deux notions se réfèrent à la puissance d'un test.

Pour qu'un test soit un test au seuil de significativité de 5 % de H_0, la chance pour que la P-value associée soit inférieure à 0.05 doit être d'au plus 5 % lorsque H_0 est vraie. La *puissance* d'un test de l'hypothèse nulle, H_0, contre une hypothèse alternative, H_a, est la probabilité qu'H_0 soit rejetée, quand H_a est vraie. Si l'on réalise un test au seuil de 5 %, alors la puissance du test est la probabilité que la P-value soit inférieure ou égale à 0,05 quand H_0 est fausse et que H_a est vraie. Il convient donc de disposer d'un test à la puissance la plus élevée possible. De ce point de vue - et donc plus formellement - un test est *consistant* contre l'hypothèse alternative H_a si sa puissance tend vers 1 quand la taille de l'échantillon augmente, c'est-à-dire que le rejet de H_0 en faveur de H_a est presque certain si H_a est vraie et que la taille de l'échantillon est importante. Un test est un test *sans biais* de H_0 contre H_a si la puissance est au moins égale au seuil de significativité quand H_a est vraie. Au seuil de significativité de 5 %, un test est sans biais contre H_a si la puissance du test est d'au moins 5 % quand H_a est vraie.

16. Cela explique que « contrôler » de la réponse non affectée par le mode de collecte ne saurait exclure l'existence d'un biais de composition.

Le tableau 4 présente les résultats obtenus par l'analyse par essai croisé et par l'analyse par appariement si l'on considère le temps total de trajets effectués lors de la journée. Quelle que soit l'approche méthodologique retenue, et dans une ampleur cohérente, le temps total de trajets effectués lors de la journée déclaré à la fin du questionnaire est significativement plus faible pour les carnets numériques comparé aux carnets papier. En d'autres termes, les répondants par internet et les répondants papier diffèrent du point du vue du temps consacré aux trajets le jour de l'enquête, alors qu'il est peu vraisemblable que cette différence soit causée par un effet de mesure du mode de collecte, comme nous avons essayé de le justifier précédemment.

Tableau 4 – Un biais de sélection endogène ou de composition est-il possible ?

	Essai croisé Effet constant	Appariement Effet moyen
Temps de trajet total déclaré	−9 [−15;−3]	−15 [−25;−5]

Note : Les estimations de l'essai croisé reposent sur les durées déclarées par 866 personnes ayant complété au moins un carnet, erreurs standards robustes à une corrélation sérielle et à l'hétéroscédasticité (General Feasible Generalized Least Squares Analysis Wooldridge, 2002, chapitre 10). L'analyse par appariement repose sur 311 paires. Effets statistiquement significatifs à 95 % en gris.

Ce résultat suggère l'existence d'un biais de sélection endogène. Il est donc nécessaire de questionner l'impact de cette sélection endogène sur les effets de mesure mis en évidence au tableau 3. Pour cela, nous implémentons une analyse de sensibilité aux résultats de l'approche par appariement, ce type d'analyse n'étant, à notre connaissance, pas développé dans le cas des tests croisés.

3 L'analyse de sensibilité sur des paires appariées

Dans une expérience aléatoire, la probabilité pour une unité de l'échantillon d'être enquêtée par internet est la même que celle d'être enquêtée *via* le questionnaire papier. Par conséquent, hormis par hasard, aucun type de personne n'est surreprésenté dans le groupe enquêté par internet ou celui enquêté *via* le questionnaire papier. Si la différence de réponses observée est trop importante pour être due au hasard, c'est-à-dire trop importante pour être attribuée de manière plausible à une séquence malchanceuse de tirages à pile ou face, alors l'existence d'un effet de mesure est démontré. Dans le cadre de notre étude, les personnes enquêtées sont affectées aléatoirement à un premier mode de collecte, de telle sorte que le protocole s'apparente à une expérience aléatoire. Cependant, comme toutes les personnes enquêtées n'ont pas répondu, celles *ayant répondu* dans l'un ou l'autre groupe peuvent être différentes. Les personnes ayant répondu peuvent différer sur des caractéristiques observables, mais il est difficile d'exclure la possibilité qu'elles ne diffèrent pas aussi sur une caractéristique qui ne peut être mesurée, comme par exemple, avoir un « travail étudiant ». Dans ce cas de figure, comparer leur durée déclarée de temps de travail conduirait alors à conclure à l'existence d'un effet de mesure de façon erronée.

Le modèle d'analyse de sensibilité proposé par Rosenbaum que nous implémentons dans cette étude permet de répondre en partie à cette éventualité. Il s'agit en effet d'estimer de quelle « ampleur » devrait être la caractéristique non mesurée pour arriver à modifier les conclusions émises sur l'existence d'un effet de mesure, lorsque l'on suppose l'absence de biais caché. Ce possible biais caché découlerait, dans notre étude, d'une possible sélection endogène : parce que le modèle d'analyse de sensibilité considère quelle qu'en soit la cause la possibilité d'un biais de composition caché (et non explicitement issu d'une possible

sélection endogène), nous emploierons ce terme dans cette partie méthodologique.

Nous rappelons tout d'abord dans un cadre unifié - pour les variables continues et binaires - comment est effectuée l'inférence aléatoire dans une expérience randomisée par paires (section 3.1). En cela, cette partie complète la présentation pour le seul cas de variables binaires faite par Court et Quantin (2024). Puis nous décrivons, dans ce cadre unifié, comment mener une analyse de sensibilité (section 3.2). Dans une troisième sous-partie nous revenons sur l'interprétation de l'écart à l'hypothèse d'absence de biais de composition caché dans la cas de l'analyse de paires de répondants. Nous détaillons notamment le lien entre le paramètre d'analyse de sensibilité, introduit dans la sous-partie précédente, et la corrélation entre d'une part la caractéristique inobservée et le mode de collecte et d'autre part entre la caractéristique inobservée et la réponse donnée d'autre part (section 3.3). L'objectif de cette sous-partie est de permettre une compréhension de nos résultats au lecteur plus familier de l'approche économétrique de l'évaluation qui questionne usuellement la présence d'une caractéristique inobservée en fonction de sa corrélation avec le traitement et la variable étudiée. Enfin, nous présentons la notion d'effets attribuables au traitement dans le cas d'une variable d'intérêt binaire qui nous permettra de proposer une autre quantification que l'effet moyen de l'effet de mesure du mode de collecte (section 3.4). Comme nous le verrons dans la présentation des résultats, cette quantification offre un éclairage sur l'ampleur de l'effet de mesure complémentaire, parfois plus explicite. Néanmoins, le lecteur intéressé trouvera en annexe D une présentation théorique des estimateurs d'Hodges-Lehmann (dans le cadre de la statistique de Wilcoxon signée du rang) nécessaires à l'estimation d'un effet additif classique, y compris dans le cadre de l'analyse de sensibilité. Ces différentes parties sont très théoriques ; un résumé non formalisé et plus intuitif de la démarche est donc proposé à la section 3.5. Toutes ces parties s'appuient fortement sur les présentations très complètes que l'on trouve dans les ouvrages et articles de Rosenbaum (2002b); Rosenbaum et Silber (2009); Rosenbaum (2010).

Notations, effet du traitement et assignation au traitement
Dans cette étude, on considère I paires, $i = 1, \ldots, I$, comportant chacune deux personnes enquêtées $j = 1{,}2$, l'une répondant par internet $Z_{ij} = 1$ (ci-après dénommée « traitée »), l'autre répondant par papier $Z_{ij'} = 0$ (ci-après dénommée « contrôle »), appariées sur des caractéristiques observées $\mathbf{x}_{ij}$, de telle sorte que $\mathbf{x}_{i1} = \mathbf{x}_{i2}$ et $Z_{i1} + Z_{i2} = 1$, pour chaque paire i. Cependant, au sein de chaque paire, les personnes peuvent différer sur une caractéristique inobservée, u_{ij}, de telle sorte que $u_{i1} \neq u_{i2}$. Suivant les notations de Neyman (1923) et Rubin (1974), chaque personne enquêtée a deux réponses potentielles, celle r_{Tij} si la personne enquêtée répond par internet (i.e. en présence de traitement) et r_{Cij} si la personne enquêtée répond *via* le questionnaire papier (i.e. en l'absence de traitement). Ainsi, la réponse observée pour l'unité j de la paire i est $R_{ij} = Z_{ij}r_{Tij} + (1 - Z_{ij})r_{Cij}$. L'effet du traitement $r_{Tij} - r_{Cij}$ correspond à l'effet de mesure du mode de collecte alternatif (internet) par rapport au mode de collecte de référence (collecte papier). Il n'est observé pour aucune unité enquêtée ij, notamment car l'on restreint notre analyse dans cette partie aux réponses données uniquement lors de la première période.

On note $\mathbf{R} = (R_{11}, \ldots, R_{I2})^T$, $\mathbf{Z} = (Z_{11}, \ldots, Z_{I2})^T$, $\mathbf{r}_T = (r_{T11}, \ldots, r_{TI2})^T$, $\mathbf{r}_C = (r_{C11}, \ldots, r_{CI2})^T$ et $\mathbf{u} = (u_{11}, \ldots, u_{2I})$ les vecteurs de dimension $2I$ associés aux grandeurs précédentes. L'hypothèse nulle d'absence d'effet du traitement de Fisher (1935) ("*sharp null hypothesis of no treatment effect*") énonce que les réponses de chaque personne enquêtée ij est inchangée si elle répond par internet, soit $H_0 : r_{Tij} = r_{Cij}, \forall i,j$ ou $H_0 : \mathbf{r}_C = \mathbf{r}_T$. Si H_0 est vraie, alors $\mathbf{R} = \mathbf{r}_C$.

On note $\mathcal{F} = \{(r_{Tij}, r_{Cij}, \mathbf{x}_{ij}, u_{ij}), i = 1, \ldots, I, j = 1{,}2\}$ et $\mathcal{Z}$ l'ensemble des 2^I assignations possibles au traitement $\mathbf{Z}$ de telle sorte que $\mathbf{z} \in \mathcal{Z}$ si $z_{ij} = 0$ ou 1 et $z_{i1} + z_{i2} = 1$ pour tout

i. Enfin, le nombre d'éléments d'un ensemble S est noté $|S|$; ainsi $|\mathcal{Z}| = 2^I$.

3.1 Inférence aléatoire dans les expériences randomisées

Dans le cadre d'une expérience randomisée par paires, l'affectation aléatoire au traitement assure que $\forall i, \Pr(Z_{ij} = 1 \mid \mathcal{F}, \mathcal{Z}) = 1/2$ et $\Pr(\mathbf{Z} = \mathbf{z} \mid \mathcal{F}, \mathcal{Z}) = 1/|\mathcal{Z}| = 2^{-I}$ pour tout $\mathbf{z} \in \mathcal{Z}$. Comme le souligne la formule de Fisher (1935), l'affectation aléatoire est l'élément de base de l'inférence ("*reasoned basis for inference*") dans une telle expérience randomisée, au sens où la distribution de toute statistique de test, $t(\mathbf{Z},\mathbf{r})$, sous l'hypothèse nulle H_0 est sa distribution de permutation :

$$\Pr(t(\mathbf{Z}, \mathbf{R}) \geq k \mid \mathcal{F},\mathcal{Z}) = \Pr(t(\mathbf{Z}, \mathbf{r}_C) \geq k \mid \mathcal{F},\mathcal{Z}) = \frac{|\{\mathbf{z} \in \mathcal{Z} : t(\mathbf{Z}, \mathbf{R}) \geq k\}|}{2^I} \tag{2}$$

car sous $H_0, \mathbf{R} = \mathbf{r}_C$ est fixe conditionnellement à $\mathcal{F}$ et $\mathbf{Z}$ est distribuée uniformément sur $\mathcal{Z}$. Ainsi, en inférence aléatoire, dans le cas d'une expérience randomisée, c'est la connaissance du processus d'assignation aléatoire au traitement qui assure que la distribution de la statistique de test est connue, ce qui permet de tester, dans ce cadre, l'hypothèse nulle d'absence d'effet du traitement. En pratique, cependant, lorsque le nombre de paires croît, la taille de $\mathcal{Z}$ augmente et rend le calcul direct de (2) difficile. Dans ce cas, il est usuel d'utiliser une approximation asymptotique de (2). Il est aussi possible d'effectuer un calcul exact (i) en bénéficiant de simplification propre à la statistique utilisée (voir Rosenbaum (2002b) et Court et Quantin (2024) par exemple pour le cas du test de Mc Nemar) ou (ii) sans utiliser la totalité de l'ensemble $\mathcal{Z}$ à partir de la fonction caractéristique de la statistique de test (Pagano et Tritchler, 1983).

Dans ce qui suit, nous illustrons comment implémenter une approche asymptotique pour plusieurs statistiques de test couramment utilisées dans l'analyse de paires appariées [17]. Malgré leur variété, il est en effet possible de présenter cette démarche dans un cadre unifié. Plus précisément, nous discutons le test de Mc Nemar, la statistique de Wilcoxon signée du rang, les statistiques de Stephenson, les U-statistiques de Rosenbaum et les M-statistiques de Hubert-Maritz [18]. Si ces statistiques se différencient en fonction de la nature de la variable d'intérêt, binaire, discrète ou continue, les tests associés ont surtout des puissances différentes [19] en particulier lorsque seule une partie des unités traitées réagit au traitement. Dès lors, le choix d'une statistique de test plutôt qu'une autre peut impacter la robustesse des conclusions. Comme nous le verrons dans les résultats nous utiliserons de fait plusieurs statistiques de test pour discuter de la robustesse des conclusions dans le cadre de l'analyse de sensibilité pour des variables continues.

On note $V_i = Z_{i1} - Z_{i2}$, $Y_i = R_{i1} - R_{i2}$, $y_{Ci} = r_{Ci1} - r_{Ci2}$, et on considère $q_i \geq 0$ une fonction de la valeur absolue de la différence des réponses au sein d'une paire, $|Y_i|$, telle que $q_i = 0$ si $|Y_i| = 0$. Sous cette notation, $V_iY_i = (Z_{i1} - Z_{i2})(R_{i1} - R_{i2})$ correspond à la différence de réponse entre l'unité traitée et l'unité de contrôle au sein de chaque paire, et sous $H_0, Y_i = y_{Ci}$.

La plupart des statistiques dans l'analyse de paires appariées sont de la forme $t(\mathbf{Z}, \mathbf{R}) = \sum_{i=1}^{I} \text{sign}(V_iY_i)q_i$ ou une combinaison linéaire d'une telle statistique, avec $\text{sign}(w)$ qui vaut 1, 0, ou -1 respectivement si $y > 0$, $y = 0$ et $y < 0$. Ainsi, la statistique de Wilcoxon signée du rang se déduit en posant que q_i est égale au rang de $|Y_i|$ si $|Y_i| > 0$ (voir par exemple

17. Le lecteur intéressé trouvera dans les articles en référence comment effectuer le calcul exact des distributions pour un nombre de paires peu élevé.

18. On trouvera en annexe C une description sommaire des U-statistiques de Rosenbaum (et leur lien avec les statistiques de Stephenson, Stephenson (1981)) et des M-statistiques (Huber, 1964; Maritz, 1979, 1981)

19. et notamment du point de vue de l'analyse de sensibilité (Rosenbaum, 2010).

Rosenbaum, 2010, §2.3.3) [20]. Dans le cas particulier où R_{ij} est une variable binaire, Y_1 vaut $1, 0$ ou -1 ; considérer $q_i = |Y_i| = 0$ pour les paires concordantes, et $q_i = |Y_i| = 1$ pour les paires discordantes, permet d'utiliser le test de Mc Nemar (voir par exemple Rosenbaum, 2002b, §2.4.3).

Pour toutes ces statistiques, sous l'hypothèse nulle H_0, $Y_i = y_{Ci} = r_{Ci1} - r_{Ci2}$ et q_i sont fixes dans (2), conditionnellement à $\mathcal{F}$, et dans une expérience randomisée, V_i vaut 1 ou -1 avec pour chaque valeur une probabilité de 1/2. Ainsi, la distribution de $T = t(\mathbf{Z},\mathbf{R}) = \sum_{i=1}^{I} q_i \text{sign}(V_i Y_i)$ dans (2) est celle de la somme de I variables aléatoires indépendantes discrètes, $\text{sign}(V_i Z_i) q_i$ qui prennent les valeurs q_i ou 0 avec des probabilités égales, et $\text{E}(T \mid \mathcal{F},\mathcal{Z}) = \sum q_i/2$ et $\text{var}(T \mid \mathcal{F},\mathcal{Z}) = \sum q_i^2/4$.

3.2 Le modèle d'analyse de sensibilité

La distribution précédente se déduisait de l'assignation aléatoire au traitement qui a lieu dans une expérience randomisée. Cependant, dans notre étude, ce sont les unités enquêtées qui sont affectées aléatoirement à un mode de collecte, mais *pas les unités répondantes*. Dans notre étude, après appariement sur les caractéristiques observées $\mathbf{x}_{ij}$, les probabilités d'être traité $\pi_{ij} = \Pr(Z_{ij} = 1 \mid \mathcal{F})$ peuvent différer au sein d'une même paire, $\pi_{i1} \neq \pi_{i2}$, précisément, car deux personnes similaires *répondantes* peuvent être différentes sur une caractéristique inobservée $u_{i1} \neq u_{i2}$. De telle sorte, qu'après appariement $\Pr(Z_{ij} = 1 \mid \mathcal{F},\mathcal{Z}) \neq 1/2$.

Le modèle d'analyse de sensibilité (Rosenbaum, 2002b, 2010) considère justement des « écarts » d'ampleurs diverses à une affectation aléatoire du traitement - nous précisons dans ce qui suit le sens exact entendu par là - et leur impact sur l'inférence d'un éventuel effet du traitement. Précisément, le modèle d'analyse de sensibilité pose que des unités enquêtées $i1$ et $i2$ ont des probabilités inconnues $\pi_{ij} = \Pr(Z_{ij} \mid \mathcal{F})$ mais telles que pour deux unités ayant les mêmes caractéristiques observées $\mathbf{x}_{i1} = \mathbf{x}_{i2}$ celles-ci peuvent différer dans leurs rapports de chances de répondre par internet par un facteur d'au plus $\Gamma \geq 1$, soit formellement

$$\frac{1}{\Gamma} \leq \frac{\pi_{i1}/(1-\pi_{i1})}{\pi_{i2}/(1-\pi_{i2})} \leq \Gamma \quad \text{quand} \quad \mathbf{x}_{i1} = \mathbf{x}_{i2} \tag{3}$$

pour $i = 1, \ldots, I$. Par ailleurs, le modèle restreint la distribution de $\mathbf{Z}$ à $\mathcal{Z}$ en conditionnant sur $Z_{i1} + Z_{i2} = 1, \forall i$; c'est-à-dire en intégrant la structure par paires (voir Rosenbaum, 2002b, §4).

Comme le montre Rosenbaum (2002b, 2020), ce modèle est similaire si l'on pose que

$$\Pr(\mathbf{Z} = \mathbf{z} \mid \mathcal{F},\mathcal{Z}) = \frac{\exp(\gamma \mathbf{z}^T \mathbf{u})}{\sum_{\mathbf{b} \in \Omega} \exp(\gamma \mathbf{b}^T \mathbf{u})} = \prod_{i=1}^{I} \frac{\exp(\gamma z_{i1} u_{i1} + \gamma z_{i2} u_{i2})}{\exp(\gamma u_{i1}) + \exp(\gamma u_{i2})}, \mathbf{u} \in [0,1]^{2I} \tag{4}$$

où $\mathbf{z} \in \mathcal{Z}$, et $\gamma = \log(\Gamma) \geq 0$. Malgré sa complexité, on peut noter que les I termes dans (4), soit $\Pr(Z_{ij} \mid \mathcal{F},\mathcal{Z}) = \exp(\gamma u_{ij})/(\exp(\gamma u_{i1}) + \exp(\gamma u_{i2}))$ sont bornés par $1/(1+\Gamma)$ et $\Gamma/(1+\Gamma)$ (Rosenbaum, 2002b, §4.2).

Par ailleurs, lorsque $\Gamma = 1$, ou de manière identique $\gamma = 0$, alors $\pi_{i1} = \pi_{i2}$ dans (3) et $\Pr(\mathbf{Z} = \mathbf{z} \mid \mathcal{F},\mathcal{Z}) = 2^{-I}$ dans (4), c'est-à-dire que lorsque $\Gamma = 1$ ou $\gamma = 0$, on se retrouve dans la situation d'une expérience randomisée, c'est-à-dire d'absence de biais de composition

20. Il est possible par des considérations similaires d'étendre cette analogie aux statistiques de Stephenson (Rosenbaum, 2007a, 2010, §16) et aux U-statistiques de Rosenbaum (Rosenbaum, 2011). De même, si $\psi(.)$ est une fonction impaire, $\psi(-y) = -\psi(y)$, alors $\sum_{i=1}^{I} \psi(V_i Y_i) = \sum_{i=1}^{I} \text{sign}(V_i Y_i) q_i$ avec $q_i = \psi(|Y_i|)$ permet de considérer les tests basés sur les M-statistiques (Rosenbaum, 2007b, 2014).)

caché. Par conséquent, Γ est une mesure du degré d'écart par rapport à une étude sans présence de biais caché.

Pour tester l'existence d'un effet du traitement, comme nous l'avons évoqué dans la partie précédente, il est nécessaire de connaître la distribution de la statistique de test T sous l'hypothèse H_0. Cependant en présence d'un biais de composition caché, cette distribution n'est pas connue, puisque la distribution de $\mathbf{Z}$ n'est pas connue (dans (4), par exemple, les valeurs u_{i1} et u_{i2} sont inconnues). Cependant, comme le montre Rosenbaum (2002b), le modèle d'analyse de sensibilité précédent permet de déduire, à Γ fixé, un encadrement de cette distribution inconnue de la statistique de test T, par les distributions de deux statistiques, notées ci-dessous T^+ et T^-, dont les paramètres sont connus et dépendent de Γ.

$$\Pr(T^+ \geq k) \geq \Pr(T \geq k \mid \mathcal{F},\mathcal{Z}) \geq \Pr(T^- \geq k),\ \forall k \tag{5}$$

De manière générale, T^+ est la somme de I variables aléatoires indépendantes où la i^{e} variable prend une valeur q_i avec la probabilité $\Gamma/(1+\Gamma)$ et 0 avec la probabilité $1/(1+\Gamma)$ et T^- la somme de I variables aléatoires indépendantes définies de façon similaire en inversant les rôles de $\Gamma/(1+\Gamma)$ et $1/(1+\Gamma)$; la valeur q_i dépendant de la statistique de test utilisée. Lorsque $I \to \infty$, la probabilité $\Pr(T^+ \geq k)$ peut être approximée en utilisant une approximation Normale de la distribution de T^+ avec $\mathrm{E}(T^+) = \frac{\Gamma}{1+\Gamma}\sum_{i=1}^{I} q_i$ et $\mathrm{var}(T^+) = \frac{\Gamma}{(1+\Gamma)^2}\sum_{i=1}^{I} q_i^2$, et une approximation analogue pour T^- [21].

Cet encadrement de la distribution, sous l'hypothèse nulle, de la statistique de test T permet d'obtenir des *bornes* à la p-value associée à une valeur k de la statistique de test, à Γ donné. Celles-ci définissent donc un intervalle de valeurs possibles pour la p-value [22] qui reflète l'incertitude sur nos conclusions liée à un possible biais de composition caché, dont l'ampleur est caractérisée par Γ. En faisant varier Γ, l'intervalle des valeurs possibles des p-values grandit. Il existe donc une valeur Γ_{max} de l'ampleur du biais de composition caché, au-delà de laquelle l'hypothèse d'absence d'effet de mesure ne peut être rejetée. Discuter la possibilité, après appariement, d'une telle ampleur Γ_{max} du biais permet d'argumenter sur la pertinence ou non de l'existence d'un effet du traitement.

L'analyse de sensibilité proprement dite consistera donc à considérer plusieurs valeurs de Γ et à étudier comment les inférences changent. Ainsi les conclusions seront considérées comme sensibles à l'existence d'un biais caché si les valeurs de Γ proches de 1 conduisent à des inférences très différentes de celles obtenues en supposant l'étude exempte de biais caché. Dans le cas contraire, elles seront considérées comme d'autant plus insensibles que des valeurs élevées de Γ sont nécessaires pour modifier ces conclusions.

3.3 Interpréter l'écart à l'hypothèse d'absence de biais de composition caché

L'expression générale (3) du modèle d'analyse de sensibilité quantifie l'amplitude du biais de composition caché considéré, Γ, comme un odds-ratio des probabilités d'être traitée des deux unités appariées. Il est possible pour en simplifier l'interprétation d'exprimer l'amplitude du biais de composition caché, comme un intervalle des valeurs possibles de la probabilité $\theta_i = \mathrm{P}(Z_{i1} = 1, Z_{i2} = 0 \mid \mathcal{F},\mathcal{Z}, Z_{i1} + Z_{i2} = 1)$ d'observer, au sein d'une paire, l'unité enquêtée 1 (par exemple) répondre par internet. En l'absence de biais de composition

21. Une telle approximation nécessite cependant, comme le rappelle Rosenbaum (2002b), que le nombre de paires « discordantes » augmente lorsque I s'accroît, ce qui est le cas pour les statistiques présentées dans la section suivante.

22. Lorsque $\Gamma = 1$ (qui caractérise l'absence de biais de composition caché), les deux statistiques ont une distribution identique et l'intervalle se réduit à une valeur

caché, $\theta_i = 1/2$. En présence de biais de composition caché, θ_i peut être encadré par :

$$\frac{1}{1+\Gamma} \leq \theta_i \leq \frac{\Gamma}{1+\Gamma} \tag{6}$$

Pour $\Gamma = 1$, on retrouve que $\theta_i = 1/2$. Pour $\Gamma = 1{,}5$, θ_i est compris entre 0,40 et 0,60, ce qui correspond à un écart modéré vis-à-vis d'une situation qui serait analogue à une affectation aléatoire *après appariement*, contrairement à $\Gamma = 3$, puisque θ_i est alors compris entre 0,25 et 0,75.

Amplification d'une analyse de sensibilité

Dans l'analyse de sensibilité, le paramètre Γ décrit l'ampleur de l'association entre u_{ij} et Z_{ij} et l'analyse de sensibilité associée permet de déterminer l'intervalle des inférences possibles. Les bornes de (5) sont atteintes pour une caractéristique inobservée u_{ij} très corrélée, voire parfaitement, aux réponses potentielles r_{Cij} (voir Rosenbaum, 2002b, et plus bas). Cependant, l'interprétation à partir d'une telle covariable n'est pas nécessairement la plus adéquate si une corrélation quasi-parfaite entre u_{ij} et r_{Cij} est peu vraisemblable. Une amplification de l'analyse de sensibilité (Rosenbaum et Silber, 2009) offre une interprétation du paramètre Γ, différente mais équivalente, à partir de deux paramètres, Λ qui contrôle l'ampleur de l'association entre la caractéristique inobservée u_{ij} et Z_{ij} et Δ qui contrôle l'ampleur de l'association entre u_{ij} et r_{Cij}, à l'instar des méthodes classiques d'évaluation économétrique qui questionne l'impact sur les résultats de l'existence d'une caractéristique inobservée corrélée au traitement et à la variable de résultat. En effet, parce qu'u_{ij} est corrélée à Z_{ij} et r_{Cij}, elle implique une association entre Z_{ij} et r_{ij} en l'absence d'ajustement sur u_{ij}. Plus précisément, $\Lambda \geq 1$ définit l'ampleur de l'association entre $Z_{i1} - Z_{i2}$ et (u_{i1}, u_{i2}) : il pose que $\Pr(Z_{i1} = 1 \mid \mathbf{x}_{ij}, u_{ij}, \mathcal{Z})$ est d'au moins $1/(1+\Lambda)$ et d'au plus $\Lambda/(1+\Lambda)$ [23]. De même, le paramètre $\Delta \geq 1$ définit l'ampleur de l'association entre $Y_{Ci} = R_{i1} - R_{i2} = r_{Ci1} - r_{Ci2}$ sous H_0 et $(u_{i1} - u_{i2})$: il pose que $\Pr(Y_{Ci} > 0 \mid \mathbf{x}_{ij}, u_{ij}, \mathcal{Z})$ est d'au moins $1/(1+\Delta)$ et d'au plus $\Delta/(1+\Delta)$.

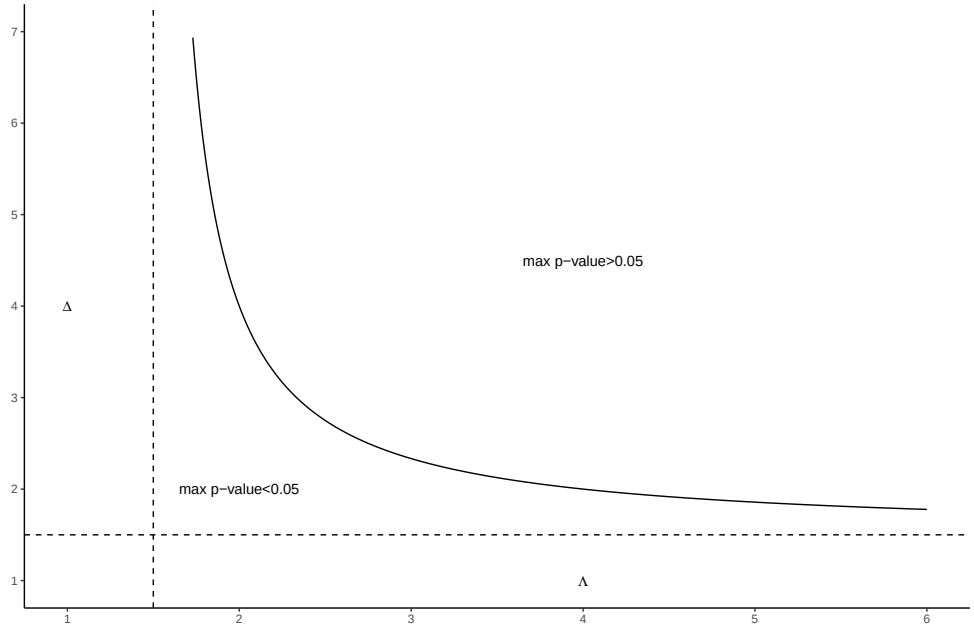
Formellement, dans l'interprétation d'une analyse de sensibilité, un biais d'ampleur Γ est équivalent à un modèle d'analyse de sensibilité défini à partir de (Λ, Δ) dès lors que :

$$\Gamma = \frac{(\Lambda\Delta + 1)}{(\Lambda + \Delta)} \tag{7}$$

Ainsi, une caractéristique inobservée qui doublerait ($\Lambda = 2$) les chances d'être traitée dans une paire et quadruplerait ($\Delta = 4$) les chances d'une différence positive entre les réponses est identique à $\Gamma = 1{,}5$. Mais, il en est de même pour toute valeur (Λ, Δ) telle que $1{,}5 = (\Lambda\Delta + 1)/(\Lambda + \Delta)$, par exemple $(\Lambda, \Delta) = (4{,}2)$ ou $(\Lambda, \Delta) = (2.5{,}2.75)$. La figure 1 représente ainsi (une partie de) l'ensemble des valeurs (Λ, Δ) possibles pour $\Gamma = 1{,}5$ et par extension l'ensemble des valeurs (Λ, Δ) pour lesquelles la borne supérieure de la p-value de l'équation (5) est inférieure à 0,05 si $\Gamma_{\max} = 1{,}5$.

23. Les quantités Γ et Λ sont proches, mais différentes : Γ fait référence à $\pi_{ij} = \Pr(Z_{ij} \mid r_{Tij}, r_{Cij}, \mathbf{x}_{ij}, u_{ij})$, tandis que Λ renvoie à $\Pr(Z_{ij} \mid \mathbf{x}_{ij}, u_{ij})$. Ainsi Γ s'appuie sur u_{ij} et r_{Cij}, et implicitement sur Y_{Ci}, tandis que Λ ne s'appuie que sur u_{ij} et non sur r_{Cij} c'est-à-dire sans inclure Y_{Ci}.

FIGURE 1 – Amplification (Λ,Δ) pour Γ = 1.5



L'équation (7) définit la correspondance entre une analyse de sensibilité à partir d'un paramètre et de deux paramètres. Comme nous l'avons évoqué précédemment, il est intéressant de noter que l'analyse à partir d'un paramètre est aussi la limite d'une analyse à partir de deux paramètres, au sens où $\Delta \to \infty$ dans (7) conduit à $\Lambda \to \Gamma$. Ainsi, les bornes de (5) peuvent être interprétées comme le cas particulier d'un modèle d'analyse de sensibilité à deux paramètres où la caractéristique inobservée u_{ij} est corrélée quasi-parfaitement à Y_{Ci}.

Un autre point important de l'équation (7) est qu'elle permet de définir quelle caractéristique inobservable u_{ij} est pertinente du point de vue du biais dans l'inférence aléatoire. En effet, si $\Gamma = 1$, alors les valeurs possibles de (Λ, Δ) sont les courbes (1,Δ) où $\Delta \in (0,\infty[$ et (Λ,1) où $\Lambda \in (0, \infty[$. Cela signifie classiquement qu'il n'existe un biais que si u_{ij} est corrélé simultanément à Z_{ij} et r_{Cij}.

Au final, le tableau 5 ci-dessous fournit pour différentes valeurs de Γ, l'intervalle des valeurs possible de θ_i et un couple de valeurs possibles pour (Λ, Δ).

TABLEAU 5 – Interpréter le paramètre Γ

Γ	Intervalle des valeurs possibles de θ_i		Λ	Δ
1	0,50	0,50	1	1
1,1	0,48	0,52	1,40	1,80
1,25	0,44	0,56	2	2
1,5	0,40	0,60	2	4
2	0,33	0,67	3	5

Note : issu de (Rosenbaum, 2017a, chapitre 9).

En résumé, Γ est une mesure de l'écart considéré à la situation d'une absence de biais de composition caché après appariement. En l'absence de biais de composition caché, la probabilité que la première personne dans une paire soit traitée est 1/2. En présence d'un biais de composition caché, $\Gamma = 2$ par exemple, signifie que cette probabilité peut n'être

que de 1/3 ou au contraire atteindre au plus 2/3 (soit comprise plus généralement entre $1/(1+\Gamma)$ et $\Gamma/(1+\Gamma)$) ce qui représente un écart assez important par rapport à 1/2. Il est parfois plus utile d'exprimer l'écart à l'hypothèse d'absence de biais caché en fonction de l'ampleur de l'association entre une caractéristique inobservée d'une part et le traitement et la variable d'intérêt considérés d'autre part. Ainsi, par exemple, $\Gamma = 2$ est aussi identique à l'existence d'une caractéristique inobservée qui augmenterait les chances d'être traité, au sein d'une paire, d'un facteur de 3 et celles d'observer une valeur de la variable d'intérêt plus élevée d'un facteur de 5.

3.4 Effets attribuables au traitement - une présentation simplifiée du cas d'une variable d'intérêt binaire

Le modèle d'analyse de sensibilité décrit précédemment permet de questionner l'impact de l'existence d'un biais de composition caché sur la conclusion d'un test de significativité de l'effet *via* la p-value. Dans le cas de variables continues, comme le montre Rosenbaum (2002b, 2010), il est aussi possible à partir de ce même modèle d'analyse de sensibilité d'estimer un *intervalle des valeurs possibles* d'un effet du traitement lorsque l'on relâche d'une ampleur Γ l'hypothèse d'absence de biais de composition caché. Cet intervalle des valeurs possibles de l'effet estimé traduit l'incertitude liée à l'existence d'une caractéristique inobservée, et son amplitude dépend de la valeur du paramètre Γ considéré dans l'analyse de sensibilité. Pour cela la démarche s'appuie sur les estimateurs d'Hodges-Lehmann (Hodges et Lehmann, 1963; Lehmann, 1975), qui « inverse » la statistique de test ajustée c'est-à-dire, par exemple dans le cas d'un effet additif constant τ_0, en considérant $t(\mathbf{Z},\mathbf{R} - \tau_0\mathbf{Z})$. À titre d'exemple, nous décrivons formellement la démarche, avec, pour statistique de test, la statistique de Wilcoxon signée du rang, en Annexe D.

Cette méthode pourrait aussi tester d'autres hypothèses, comme celle d'un effet nul pour certaines unités traitées et non nuls et différents deux à deux pour les autres. Néanmoins, au-delà de la multiplicité des hypothèses à tester, comme le souligne Rosenbaum (2010), il est par ailleurs difficile de comprendre dans un espace qui n'est plus unidimensionnel les intervalles de confiance associés et donc l'interprétation des résultats obtenus. Pour contourner cette difficulté, il est usuel d'estimer l'effet moyen du traitement sur les traités. Cependant, si l'effet moyen du traitement tient compte de l'hétérogénéité de l'effet, il n'en fournit qu'une description imparfaite et est peu robuste, notamment car la démarche s'appuie sur la statistique de la moyenne des différences qui est l'une des statistiques les moins efficaces. Rosenbaum (2002a) propose plutôt de s'intéresser aux effets dits *attribuables* au traitement qui offrent comme nous le verrons une analyse synthétique de l'existence d'effets que l'on pourrait imputer au dispositif, sans effectuer d'hypothèse sur leur ampleur au niveau individuel. Avant de détailler plus précisément la méthodologie mise en œuvre, nous explicitons tout d'abord la notion *d'effet attribuable au traitement* introduite par Rosenbaum (2001).

Considérons comme « succès » la réalisation d'un évènement particulier, par exemple le fait pour une unité enquêtée de ne pas compléter la totalité de son emploi du temps sur 24 heures, c'est-à-dire une variable binaire. Le nombre d'effets attribuables au traitement est le nombre d'évènements parmi les unités traitées qui ne se seraient pas produits si celles-ci n'avaient pas répondu par internet. Cette estimation, comme nous le verrons, peut être réalisée en considérant comme valide l'hypothèse d'absence de biais de composition caché après appariement ($\Gamma = 1$), mais aussi en autorisant un relâchement de cette hypothèse ($\Gamma > 1$).

Par définition, pour chaque unité enquêtée, les réponses potentielles r_{Tij} et r_{Cij} prennent donc la valeur 1 si l'évènement se réalise (c'est-à-dire en cas de « succès ») et 0 sinon. On

suppose dans ce qui suit que répondre par internet peut occasionner un évènement mais pas l'empêcher si cet évènement était survenu en répondant avec le questionnaire papier, soit $r_{Tij} \geq r_{Cij}$ [24]. Par exemple, répondre par internet peut conduire une personne enquêtée à ne pas compléter la totalité de son emploi du temps, mais on suppose que cet évènement se serait obligatoirement réalisé si cela était aussi le cas en répondant avec le questionnaire papier. Cette hypothèse est nécessaire pour l'inférence qui s'appuie sur la distribution de $\boldsymbol{Z}$. Comme $r_{Tij} \geq r_{Cij}$, une personne répondant par internet qui ne complète pas la totalité de son emploi du temps ($r_{Tij} = 0$) ne l'aurait pas non plus complété en totalité en répondant au questionnaire papier ($r_{Cij} = 0$). De même, si une personne répondant au questionnaire papier complète la totalité de son emploi du temps ($r_{Cij} = 1$), elle en aurait fait de même avec le questionnaire internet ($r_{Tij} = 1$).

On note $\delta_{ij} = r_{Tij} - r_{Cij}$ l'effet du mode de collecte et $\boldsymbol{\delta}$ le vecteur de dimension $2I$ associés. Ainsi, une hypothèse $\boldsymbol{\delta} = \boldsymbol{\delta}_0$ est *compatible* avec les données si $\delta_{ij} = 0$ lorsque ($Z_{ij} = 1$ et $R_{ij} = 0$) ou ($Z_{ij} = 0$ et $R_{ij1} = 1$) (voir Rosenbaum, 2001, pour une utilisation des hypothèses compatibles), et *incompatible* le cas échéant. Une hypothèse incompatible peut être rejetée avec certitude, c'est-à-dire avec une erreur de type 1 nulle. Si une hypothèse est vraie, alors elle est compatible pour tout $\boldsymbol{Z}$, tandis qu'une hypothèse fausse peut être compatible pour certains $\boldsymbol{Z}$ et pas pour d'autres. Dès lors, puisque l'on cherchera à tester une hypothèse nulle en s'appuyant sur la distribution de $\boldsymbol{Z}$, il est nécessaire de pouvoir s'assurer qu'elle corresponde à une hypothèse compatible.

L'effet attribuable au mode de collecte est $A = \sum_{i,j} Z_{ij}(r_{Tij} - r_{Cij}) = \sum_{i,j} Z_{ij}\delta_{ij}$ [25]. Si l'on note $T = \sum_{i,j} Z_{ij}R_{ij}$ le nombre de personnes répondant par internet qui connaissent un « succès », alors $T - A = \sum_{i,j} Z_{ij}r_{Cij}$ correspond au nombre de personnes répondant par internet qui auraient connu un succès même sans répondre par internet. Dans ce qui suit, r_{Ci+} désigne le nombre de personnes de la paire i qui auraient connu l'évènement même en ne répondant pas par internet.

Considérons le test de l'hypothèse nulle $H_0 : \boldsymbol{\delta} = \boldsymbol{\delta}_0$ contre l'alternative $H_1 : \boldsymbol{\delta} \geq \boldsymbol{\delta}_0$.

Si l'hypothèse nulle est incompatible, alors elle est rejetée avec une erreur de type 1 nulle. Si l'hypothèse nulle est compatible, alors sous $H_0 : \boldsymbol{\delta} = \boldsymbol{\delta}_0$, r_{Cij} est connue pour tout i,j, car $r_{Cij} = R_{ij} - \delta_{0ij}Z_{ij}$. En notant $A_0 = \sum_{i,j} Z_{ij}\delta_{0ij}$, le nombre d'effets attribuables au mode de collecte parmi les unités répondant par internet associé à $\boldsymbol{\delta}_0$, $T - A_0 = \sum_{i,j} Z_{ij}r_{Cij}$, c'est-à-dire que $T - A_0 = \sum_i B_i$ est la somme de I variables aléatoires indépendantes binaires $B_i = \sum_j Z_{ij}r_{Cij}$.

En l'absence de biais de composition caché

$$\Pr(B_i = 1) = \pi_i = \frac{r_{Ci+}}{2} \quad \text{avec } \Gamma = 1$$

En présence d'un biais de composition caché, $\Pr(B_i = 1)$ est inconnue mais elle peut être encadré par (Rosenbaum, 2002a) :

$$\bar{\pi}_i = \frac{\Gamma r_{Ci+}}{\Gamma r_{Ci+} + 2 - r_{Ci+}} \geq \Pr(B_i = 1) \geq \frac{r_{Ci+}}{r_{Ci+} + \Gamma(2 - r_{Ci+})} \tag{8}$$

Si l'on note $\beta(k,\boldsymbol{\pi})$, la probabilité d'avoir au moins k succès parmi I essais avec $\boldsymbol{\pi} =$

24. On notera qu'un tel modèle ne peut être vérifié ou réfuté car r_{Tij} et r_{Cij} ne sont jamais observés simultanément.

25. A est une grandeur qui n'est pas observée puisque r_{Tij} et r_{Cij} ne le sont pas et sa valeur dépend des assignations au mode de collecte Z_{ij} : c'est donc une variable aléatoire.

$(\pi_1,\ldots,\pi_I)^T$, alors $\Pr(\sum_i B_i \geq k) = \beta(k,\boldsymbol{\pi})$ et correspond au niveau de significativité du test de l'hypothèse nulle de la statistique $k = T - A_0$.

Ainsi, en l'absence de biais de composition caché, la démarche pour tester $H_0 : \boldsymbol{\delta} = \boldsymbol{\delta}_0$, consiste à calculer $r_{Cij} = R_{ij} - Z_{ij}\delta_{0ij}$ sous H_0, puis r_{Ci+}, afin d'obtenir $\pi_i = \bar{\pi}_i = \frac{r_{Ci+}}{2}$, puis $\Pr(\sum_i B_i \geq k) = \beta(k,\boldsymbol{\pi})$.

En présence d'un biais de composition caché, i.e. si $\Gamma > 1$, comme π_i est inconnu à cause de l'existence d'une caractéristique inobservable, $\Pr(\sum_i B_i \geq k) = \beta(k,\boldsymbol{\pi})$ ne peut pas être calculée mais elle peut être majorée par $\beta(k,\bar{\boldsymbol{\pi}}) \geq \beta(k,\boldsymbol{\pi})$ d'après (8) après avoir déterminé r_{Ci+}.

Dans les deux cas, quand $I \to \infty$, $\beta(k,\bar{\boldsymbol{\pi}})$ peut être approximé [26] par

$$\beta(k,\bar{\boldsymbol{\pi}}) \to 1 - \Phi\left(\frac{k - \sum_i \bar{\pi}_i}{\sqrt{\sum \bar{\pi}_i(1-\bar{\pi}_i)}}\right) = 1 - \Phi\left(\frac{T - A_0 - \sum_i \bar{\pi}_i}{\sqrt{\sum \bar{\pi}_i(1-\bar{\pi}_i)}}\right) \tag{9}$$

L'expression générale de $\bar{\pi}_i$ dans (8) peut être explicitée en considérant sous H_0 les différentes valeurs possibles de r_{Ci+}. Il existe quatre types de paires possibles dont les caractéristiques sont présentées dans le tableau 6. D et C indique une paire discordante et concordante respectivement. $D(+,-)$ indique ainsi une paire discordante où l'unité traitée connaît l'évènement, mais pas l'unité de contrôle.

Tableau 6 – Structure des paires concordantes et discordantes sous H_0

	Nombre					sous H_0			
		Z_{i1}	Z_{i2}	R_{i1}	R_{i2}	r_{Ci1}	r_{Ci2}	r_{Ci+}	$\bar{\pi}_i$
$D(+,-)$	n_{D+}	1	0	1	0	1	0	1	$\Gamma/(\Gamma+1)$
$D(-,+)$	n_{D-}	1	0	0	1	0	1	1	$\Gamma/(\Gamma+1)$
$C(-,-)$	n_{C-}	1	0	0	0	0	0	0	0
$C(+,+)$	n_{C+}	1	0	1	1	1	1	2	1

Note : D et C indiquent des paires discordantes et concordantes, et $+$ et $-$ indique si l'unité a connu l'évènement considéré.

Le nombre T de personnes répondant par internet qui connaissent un succès correspond à $n_{D+} + n_{C+}$. Supposons que l'on veuille tester l'hypothèse qu'il y a A_0 effets attribuables au traitement parmi les n_{D+} paires discordantes, $D(+,-)$, où l'unité traitée a connu l'évènement. Si l'on calcule $r_{Cij} = R_{ij} - Z_{ij}\delta_{0ij}$ sous H_0, il vient qu'il y a $r_{Ci+} = n_{D+} - A_0$ personnes de la paire i qui auraient connu l'évènement quel que soit le mode de collecte. Ainsi, d'après le tableau 6, sous H_0, (i) $\bar{\pi}_i = 1$ pour n_{C+} paires concordantes, (ii) $\bar{\pi}_i = 0$ pour n_{C-} paires concordantes, et (iii) $\bar{\pi}_i = \Gamma/(1+\Gamma)$ pour les $n_{D+} + n_{D-} - A_0$ paires toujours discordantes sous H_0. Ainsi $T - A_0 = (n_{D+} + n_{C+} - A_0)$ a comme espérance $\sum_i \bar{\pi}_i = n_{C+} + (n_{D+} + n_{D-} - A_0)\Gamma/(1+\Gamma)$. Au final,

$$\frac{T - A_0 - \sum_i \bar{\pi}_i}{\sqrt{\sum_i \bar{\pi}_i(1-\bar{\pi}_i)}} = \frac{n_{D+} - A_0 - (n_{D+} + n_{D-} - A_0)\Gamma/(1+\Gamma)}{\sqrt{(n_{D+} + n_{D-} - A_0)\Gamma/(1+\Gamma)^2}} \tag{10}$$

ce qui permet de calculer $\beta(k,\bar{\boldsymbol{\pi}})$.

Il est aussi possible de tester qu'il y a A_0 effets attribuables au mode de collecte parmi les I, et non simplement parmi les n_{D+} paires discordantes ; nous ne testerons pas ces hypothèses

26. Une expression exacte du test peut être implémentée à partir de la distribution binomiale.

dans notre étude. De même un intervalle de confiance du nombre d'effets attribuables au mode de collecte peut être calculé : pour toutes ces questions (plus complexes à implémenter), nous renvoyons le lecteur intéressé à Rosenbaum (2002a). Néanmoins, nous aimerions souligner qu'il y a univocité dans l'interprétation du test au sens où deux hypothèses compatibles $\boldsymbol{\delta}_0$ et $\bar{\boldsymbol{\delta}}_0$ conduisant au même nombre A_0 d'effets du traitement ont une conclusion analogue. De manière plus générale et intuitive, Rosenbaum démontre une relation « d'ordre » entre deux hypothèses compatibles avec $\boldsymbol{\delta}_0 - \bar{\boldsymbol{\delta}}_0 \geq \mathbf{0}$, au sens où $\boldsymbol{\delta}_0$ est plus grande que $\bar{\boldsymbol{\delta}}_0$ car il attribue plus d'effets au mode de collecte. Si $H_0 : \boldsymbol{\delta} = \boldsymbol{\delta}_0$ est rejetée contre l'hypothèse alternative $H_1 : \boldsymbol{\delta} \geq \boldsymbol{\delta}_0, \boldsymbol{\delta} \neq \boldsymbol{\delta}_0$ et si $\boldsymbol{\delta}_0$ est plus grande que $\bar{\boldsymbol{\delta}}_0$ alors l'hypothèse $H_0 : \boldsymbol{\delta} = \bar{\boldsymbol{\delta}}_0$ est aussi rejetée.

3.5 Résumé de l'analyse de sensibilité

Le modèle d'analyse de sensibilité pose que deux personnes appariées sur certaines caractéristiques observées peuvent différer dans leurs chances (*odds-ratio*) d'être traitées, au sein de la paire qu'ils constituent, d'un facteur au plus Γ.

Ainsi, $\Gamma = 1$ correspond exactement à l'hypothèse d'absence de biais de composition caché, c'est-à-dire à l'approche inférentielle aléatoire (Fisher, 1935). $\Gamma > 1$ signifie, à l'inverse, que l'on ne connaît pas vraiment la distribution de la probabilité d'être traité, mais que l'écart par rapport à l'inférence aléatoire est restreint dans son ampleur par Γ. Une valeur élevée de Γ autorise ainsi un écart important vis à vis de l'hypothèse d'absence de biais de composition caché.

Comme la distribution de la probabilité d'être traité n'est pas connue lorsque $\Gamma > 1$, l'inférence ne fournit pas une p-value, mais un intervalle de p-value. De même, elle ne fournit pas une estimation de l'effet de mesure du mode de collecte, mais un intervalle de valeurs possibles pour l'effet de mesure du mode de collecte, la longueur de cet intervalle augmentant au fur et à mesure que Γ s'accroît. Ainsi, une inférence est dite sensible à un biais d'ampleur Γ quand l'intervalle inclut qualitativement différentes inférences, par exemple à la fois le rejet et l'acceptation de l'hypothèse nulle au seuil de 5 %. En faisant varier Γ, l'analyse de sensibilité permet donc de questionner l'ampleur du biais de composition caché nécessaire pour modifier les conclusions obtenues sous cette hypothèse, en s'appuyant sur les données. Il est vrai que cela ne résout pas complètement le désaccord entre l'affirmation qu'il existe un effet de mesure et celle d'un biais dû à une covariable non observée. Néanmoins, c'est une chose de dire qu'un biais minuscule et à peine perceptible dans l'attribution des modes pourrait expliquer une association, que de dire que seul un biais important pourrait le faire. Cela étant dit, lorsque cela est possible, nous compléterons cette analyse en essayant de fournir une explication raisonnable à l'effet de mesure.

4 Résultats

Nous détaillons désormais les résultats de notre analyse de l'effet de mesure du mode de collecte sur quatre durées d'activités : la durée de sommeil, la durée de trajet, le temps consacré aux loisirs et le temps consacré au repas. Ces durées ont été choisies par la division Condition de vie des ménages, responsable de l'enquête Emploi du temps, car elles sont les variables d'intérêt historiques de l'enquête. La présentation des résultats sur la durée de sommeil (section 4.1) nous permet d'illustrer l'implémentation de l'analyse de sensibilité pour une variable continue, l'interprétation par amplification (cf. section 3.3) de l'ampleur Γ de l'écart à l'hypothèse d'absence de biais de composition caché, i.e. de sélection endogène ainsi que le calcul d'un effet attribuable au traitement pour une variable binaire. Dans la partie consacrée au temps de trajets (section 4.2), nous présentons la notion de *facteurs d'évidence* proposée par Rosenbaum (2017b) qui permet de combiner deux analyses de

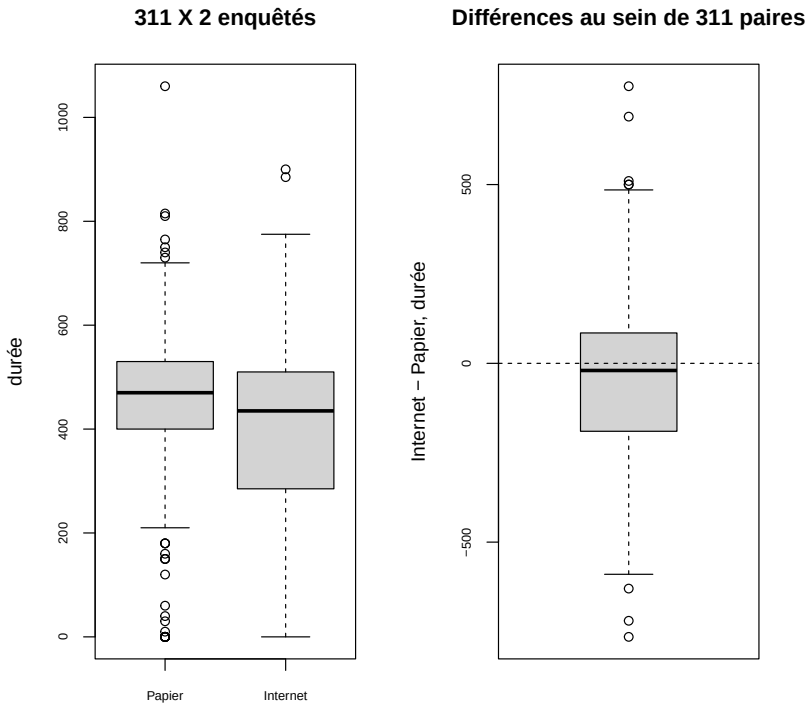
sensibilité afin de renforcer notre conviction sur l'existence d'un effet de mesure. La partie 4.3 se distingue des parties précédentes, car elle explicite l'effet de mesure qui impacte le temps déclaré consacré aux loisirs. Enfin, la partie 4.4 conclut sur l'impact de la collecte numérique sur le temps dédié aux repas déclaré par l'enquêté [27].

4.1 Un exemple détaillé d'analyse de sensibilité : la durée de sommeil

La figure 2 représente les différences de durée de sommeil déclarés par les répondants par internet et par les répondants avec le carnet papier dans les 311 paires constituées.

Dans 28 % de nos 311 paires, la différence des durées de sommeil déclarées par les deux enquêtés (internet-papier) n'excède pas 1 heure. Cependant, la différence de durées médiane est de -25 minutes. Si elle excède 75 minutes pour 25 % des paires, elle est aussi inférieure à 185 minutes dans 25 % des paires. Ainsi, la distribution est asymétrique à gauche et malgré un « pic » assez proche de zéro, de larges différences négatives (comprises entre -150 et -400 minutes) sont plus fréquentes que de larges différences positives. Cela indique que, quand des durées de sommeil très faibles sont observées, elles sont plus fréquemment déclarées par des personnes répondant par internet. Comme le suggère la distribution des différences de durée au sein des paires de la figure 2, en supposant une absence de biais de composition caché, le mode de collecte pourrait donc avoir un effet important sur la durée de sommeil déclarée par certains enquêtés, avec peut-être un effet plus réduit voire nul pour la plupart des autres répondants.

FIGURE 2 – Différence de durées de sommeil (internet - papier) en minutes



Si $\Gamma = 1$, comme expliqué dans la section 3.2, on suppose l'absence de biais de sélection endogène. Sous cette hypothèse, répondre par internet conduirait les répondants à déclarer près d'une heure de moins de sommeil en moyenne (- 53 min., CI 95 % [-79 ; -26], tableau

27. L'implémentation de la plupart des résultats a été réalisée à l'aide des packages R, `DOS2` et `sensitivitymv`. Pour leur utilisation, voir Rosenbaum (2019) et Rosenbaum (2018).

7). Un résultat similaire (-56 min. CI 95 % [-72 ; -40]) est obtenu avec l'analyse par essai croisé, c'est-à-dire en supposant la non-réponse MCAR ou MAR.

Tableau 7 – Analyse de sensibilité : effet moyen du mode de collecte sur la durée de sommeil

Γ	p-value	Effet moyen		I.C. 95 %	
	Borne sup.	Borne inf.	Borne sup.	Borne inf.	Borne sup.
1.0	<0,001	-53	-53	-79	-26
1.1	0,001	-61	-44	-88	-18
1.2	0,008	-69	-36	-96	-10
1.3	0,031	-76	-30	-104	-3
1.4	0,091	-83	-23	-112	4
1.5	0,210	-90	-17	-119	10

Avec $\Gamma = 1{,}2$, on suppose l'existence d'un biais de sélection endogène, d'ampleur réduite : la probabilité d'assignation au mode de collecte internet au sein d'une paire de l'une des deux unités n'est plus de 0,5, mais est comprise entre $(1+\Gamma)^{-1}$ et $\Gamma/(1+\Gamma)$, soit 0,45 et 0,55 (cf. section 3.3). Sous cette hypothèse, il est toujours possible de conclure à l'existence d'un effet de mesure du mode de collecte internet sur la durée de sommeil déclarée, compris entre -69 min. et -36 min. (CI 95 % [-96 ;-10]). Cependant si $\Gamma \geq \Gamma_{max} = 1{,}4$, il n'est plus possible de rejeter l'hypothèse d'absence d'effet de mesure de la collecte internet sur cette durée. Ainsi, par exemple, une covariable non observée qui doublerait, après appariement, les chances de répondre en ligne et triplerait les chances d'une différence négative (cf. section 3.3) pourrait expliquer l'association observée.

Tableau 8 – Analyse de sensibilité (Γ_{max}) sur la durée totale et le nombre de plages de sommeil

Statistique	durée totale	nombre de plages
Moyenne	1,4	2,0
Wilcoxon	1,3	2,0
Stephenson (5,5,5)	1,5	1,9

L'utilisation d'autres statistiques de test (statistique de Wilcoxon signé du rang, statistique de Stephenson, cf. Annexe C) en vue d'accroître la puissance de l'analyse de sensibilité (Rosenbaum, 2002b, §15) ne modifie pas les conclusions obtenues (cf. tableau 8) : l'hypothèse d'un effet de mesure du mode de collecte internet sur la durée de sommeil est plutôt sensible à l'existence d'un biais de sélection endogène. Si l'on considère le nombre de plages de sommeil déclaré [28], l'hypothèse d'un effet de mesure du mode de collecte est moins sensible à l'existence d'un biais de composition : quelle que soit la statistique de test utilisée, Γ_{max} est proche de 2. Cela signifie qu'il faudrait que la probabilité d'assignation au mode de collecte internet au sein d'une paire de l'une des deux unités soit inférieure à 0,33 ou supérieure à 0,67 (au lieu de 0,5) pour expliquer l'association observée.

Ce constat nous incite donc à envisager de quantifier l'effet de mesure de la collecte par internet comme un nombre d'effets attribuables au mode de collecte, c'est-à-dire plus précisément en considérant comme « évènement dû au mode de collecte internet » le fait de déclarer moins de 2 plages horaires de sommeil (cf. section 3.4). En effet, les carnets doivent détailler l'emploi du temps de la personne enquêtée sur 24 heures à partir de 4h du matin. Ainsi, il est vraisemblable (et attendu) que chaque personne déclare au moins 2 plages horaires de sommeil, par exemple entre 4h et 6h puis entre 23h et 4h le lendemain.

28. comme variable d'intérêt continue

Or, 28 % des répondants internet en déclarent moins de 2 contre 14 % des répondants sur carnet papier. Seul un biais de sélection endogène d'ampleur analogue à celui sur le nombre de plages de sommeil ($\Gamma \geq \Gamma_{max} = 1{,}8$ en faisant une analyse de sensibilité avec la statistique de Mc Nemar) pourrait expliquer une telle association.

Plus précisément, après appariement, 96 des 311 paires (soit 31 %) sont discordantes : un des deux répondants a déclaré moins de deux périodes. Dans 70 paires (soit 73 % des paires discordantes), c'est le répondant par internet qui déclare moins de 2 périodes. Parmi ces 70 paires, combien peuvent être attribuées au mode de collecte ? Autrement dit, parmi elles, combien n'auraient pas été discordantes si le répondant qui a répondu par internet avait utilisé le carnet papier ? Avant de déterminer ce nombre d'effets attribuables au mode de collecte, nous illustrons la démarche implémentée telle que présentée dans la section 3.4. Supposons que l'on teste l'hypothèse $H_0 : \boldsymbol{\delta} = \boldsymbol{\delta}_0$ correspondant à un nombre d'effets attribuables au mode de collecte parmi les paires discordantes, $A_0 = 10$. On suppose (dans un premier temps) qu'il n'y a pas de biais de sélection endogène, c'est-à-dire que $\Gamma = 1$. Sous H_0, il y a donc $70 - 10$ paires discordantes qui l'auraient quand même été si le répondant internet avait utilisé le carnet numérique ($n_{D+} - A_0$ pour reprendre les notations de la section 3.4). Comme il y a 26 paires ($n_{D-} = 96 - 70$) discordantes où c'est le répondant par questionnaire papier qui déclare moins de 2 périodes, il vient d'après (10)

$$\frac{T - A_0 - \sum \bar{\pi}_i}{\sqrt{\sum \bar{\pi}_i(1 - \bar{\pi}_i)}} = \frac{70 - 10 - (70 + 26 - 10) \times 1/2}{\sqrt{(70 + 26 - 10) \times (1/2) \times (1/2)}} = 3{,}66314$$

Ainsi, l'hypothèse nulle de A_0 effets attribuables au mode de collecte est largement rejetée au seuil de 5 % ($3{,}66 > 1{,}65 = \Phi(0{,}95)$). En faisant varier A_0, nous pouvons affirmer avec une confiance de 95 % qu'en l'absence de biais de sélection endogène, 43 % des 70 paires discordantes où le répondant internet a déclaré moins de 2 plages horaires de sommeil sont attribuables au mode de collecte internet. C'est-à-dire que la collecte par internet impacterait les réponses de 10 % des 311 paires.

En supposant l'existence d'un biais de sélection endogène d'ampleur $\Gamma = 1{,}4$, avec une démarche similaire, on pourrait toujours être sûr (à 95 %) que le mode de collecte impacte 5 % des 311 paires. Enfin, Γ devrait atteindre 1,9 [29] pour que nous ne puissions pas rejeter l'hypothèse qu'aucune paire discordante ne soit attribuable à un effet de mesure.

Le mode de collecte semble donc avoir un effet significatif sur la durée de sommeil déclarée par certains répondants et un effet plus réduit voire nul pour la plupart des participants. Une explication possible pourrait être la non-complétion de la totalité du carnet par certains répondants par internet, les conduisant à déclarer moins de 2 plages horaires de sommeil sur les 24 heures. En effet, quel que soit le mode de collecte, 90 % des répondants commencent à détailler leur emploi du temps à 4h du matin. Cependant, seuls 74 % des répondants internet le complètent jusqu'à 4h du matin le jour suivant, contre 87 % des répondants avec le carnet papier.

Une analyse de sensibilité sur la probabilité de ne pas achever la complétion de son carnet à 4h du matin aboutit à des résultats similaires à ceux présentés avant. Après appariement, parmi les 311 paires, 105 soit 37 %, sont discordantes sur leur heure de fin et parmi elles, 76 soit 72 % sont dues au répondant par internet. En l'absence de biais de sélection endogène, nous serions sûrs à 95 % que le mode de collecte est à l'origine d'au moins 43 % des paires discordantes où l'enquêté par internet n'a pas un horaire de fin attendu. Cela représente 14 % des 311 paires, un résultat en accord avec l'estimation précédente. Enfin, il faudrait

29. soit, après appariement, une covariable inobservée qui triplerait les chances de répondre par internet et qui augmenterait d'un facteur 5 celles de déclarer moins de 2 plages de sommeil pourrait expliquer les résultats obtenus.

que $\Gamma = 1{,}9$ pour que l'on ne puisse attribuer aucun effet au mode de collecte. De fait, dans 70 % des paires où le nombre de plages de sommeil déclaré sur le carnet numérique est inférieur à 2, celui-ci ne s'achève pas à 4h du matin le lendemain.

4.2 Impact sur la durée des trajets

Dans la section précédente, nous avons détaillé comment interpréter les résultats d'une analyse de sensibilité pour conclure à l'existence d'un effet de mesure de la collecte par internet sur la durée de sommeil déclarée. Cette analyse de sensibilité était nécessaire de par la présence d'un biais de sélection endogène des répondants à l'enquête test : les répondants avec le carnet numérique diffèrent dans leur temps consacré aux trajets le jour de l'enquête des répondants avec le carnet papier (cf. section 2.3), ce qui peut de fait conduire à observer des durées de sommeil différentes, même en l'absence d'effet de mesure. L'existence de ce biais de sélection endogène est bien évidemment encore plus problématique dans l'analyse de l'impact éventuel du mode de collecte sur la durée totale des trajets déclarés dans les carnets numériques et papiers. Dans un premier temps, nous montrerons que cette sélection endogène ne permet pas de rejeter l'hypothèse d'un effet de la collecte papier sur la durée totale de trajets calculée en agrégeant les temps de déplacement déclarés dans le carnet correspondant. Puis nous questionnerons l'existence d'un problème de codification des activités avec le carnet papier comme source de l'effet de mesure mis en évidence.

Comme nous l'avons expliqué dans la section 2.3, nous disposons de deux mesures de la durée totale des trajets dans cette enquête. La première résulte de l'agrégation des temps de trajet renseignés tout au long de la journée dans le carnet numérique ou le carnet papier et est susceptible d'être impactée par le mode de collecte utilisé. Ainsi, nous avons montré dans la partie 2.2, que si l'on omet l'existence d'un biais de sélection endogène, les répondants déclareraient en moyenne 32 min (IC 95 % $[20; 43]$) de temps de trajets en plus quand ils répondraient en utilisant les carnets papiers (cf. tableau 3). Cet effet de mesure est peu sensible à l'hypothèse d'un biais de sélection endogène puisque $\Gamma_{max} = 1{,}7$ [30](cf. tableau 9).

TABLEAU 9 – Analyse de sensibilité Γ_{max} sur la durée totale issue de l'agrégation des temps de trajets et sur le nombre de trajets renseignés dans les carnets

Statistique	durée totale	nombre de plages
Moyenne	1,7	1,3
Wilcoxon	1,6	1,2
Stephenson (5,5,5)	2,0	1,2

Bien que la valeur Γ_{max} nécessaire pour invalider l'hypothèse d'un effet de mesure est élevée, est-elle suffisante alors qu'il existe précisément une sélection endogène des répondants en fonction de leur temps de trajet quotidien, telle que renseignée à la fin du questionnaire ? De fait, en moyenne, les répondants papier font 15 min. de plus (IC 95 % $[5; 25]$) de temps de trajet sur une journée que les répondants papier (cf. tableau 4). Une partie de l'effet de mesure mentionné précédemment correspond donc à un biais dû à une caractéristique inobservée. Une analyse de sensibilité sur cette mesure du temps de trajet effectué qui est déclarée à la fin du questionnaire et est supposée non impactée par un effet de mesure révèle que la déviation minimale, Γ vis-à-vis de l'assignation aléatoire, i.e. l'ampleur du biais de composition pour ne plus rejeter l'absence d'association entre le mode de collecte et le temps de trajet total sur une journée déclarée à la fin du questionnaire est $\Gamma = 1{,}2$.

30. Avec d'autres statistiques, $\Gamma_{max} = 2{,}0$. À l'inverse, l'hypothèse d'un effet de mesure sur le nombre de trajet est sensible à un biais de sélection endogène.

Considérer une telle ampleur du biais de sélection endogène ne permet cependant pas de rejeter l'hypothèse d'un effet de mesure sur le temps total de trajets calculé par agrégation des temps des plages horaires correspondantes ($\Gamma_{max} = 2 > 1{,}2$). Cela signifie donc qu'un effet de mesure peut toujours être considéré comme plausible même en présence de la sélection endogène mise en évidence dans cette étude (voir Rosenbaum, 2023, pour un énoncé plus précis et les considérations théoriques soutenant cette assertion).

Questionner l'existence d'un problème de codification des activités avec le carnet papier

Une explication possible à cet effet de mesure est celle d'un problème de codification des activités déclarées sur le carnet papier. En effet, quel que soit le carnet utilisé, chaque répondant explicite pour chaque plage horaire l'activité correspondante. Avec le carnet papier, l'activité est décrite par un libellé renseigné librement par l'enquêté (cf. figure 10a, Annexe G). À partir de ce libellé, un algorithme développé par l'Insee codifie ensuite l'activité. Avec le carnet numérique, l'enquêté choisit d'abord si la plage horaire qu'il renseigne correspond à un temps de trajet. Dans le cas contraire, un sélecteur lui permet de préciser le type d'activité (cf. figure 11, Annexe G). Ainsi, il se pourrait par exemple qu'avec le carnet papier l'activité « sortir pour faire des courses » ne soit pas scindée en deux activités (le trajet aller-retour pour s'y rendre et le temps consacré aux achats), si l'algorithme de codification est défaillant, mais comptabilisée comme une unique activité de trajet. À l'inverse, le *design* du carnet numérique conduit plus sûrement l'enquêté à séparer le temps consacré à se rendre au supermarché de celui consacré à faire ses courses. Il est possible de questionner cette hypothèse d'un problème de codification en s'appuyant sur la « durée totale des plages horaires à coder » dans les carnets papier, appréhendée comme une mesure de l'intensité « du traitement ». Intuitivement, si les différences entre les temps de trajet au sein des paires résultent d'un problème de codification, il est attendu que celles-ci soient d'autant plus marquées que la durée totale des plages horaires à coder dans le carnet papier est élevée.

Avec le carnet papier, l'enquêté doit préciser pour chaque plage horaire non seulement l'activité effectuée, mais aussi le lieu où celle-ci est effectuée ou le moyen de transport utilisé pour l'exercer le cas échéant. De plus, il lui est demandé si la plage horaire correspond à un trajet domicile-travail (cf. figure 10b, Annexe G). Dans le cadre de ce test, il a été demandé aux enquêteurs, une fois le carnet papier récupéré, de préciser pour chaque plage horaire et en s'appuyant sur toutes ces informations si celle-ci correspondait à (i) une activité faite au domicile, (ii) un trajet domicile-travail, (iii) une activité faite dans un autre lieu ou (iv) si elle correspondait à un autre trajet. Comme nous l'avons mentionné précédemment, ces éléments sont ensuite utilisés par un algorithme d'apprentissage pour codifier l'activité. À partir de ces informations, l'algorithme doit notamment être capable de distinguer les trajets, autres que le trajet domicile-travail, pour se rendre dans un autre lieu (comme ceux pour déposer ses enfants à l'école) qui font partie, dans la nomenclature, de la catégorie « trajets », des promenades hors du domicile (motorisées ou non), des trajets pendant le travail ou des visites chez des amis [31] par exemple qui ne doivent pas être répertoriés comme « trajet » mais associés à d'autres catégories d'activité de la nomenclature. Ainsi, plus précisément, l'hypothèse envisagée est que la collecte avec le questionnaire papier impacterait d'autant plus la durée totale de trajets [32] que la durée des plages horaires répertoriées comme « *autres trajets* » (et donc à coder par l'algorithme) serait élevée.

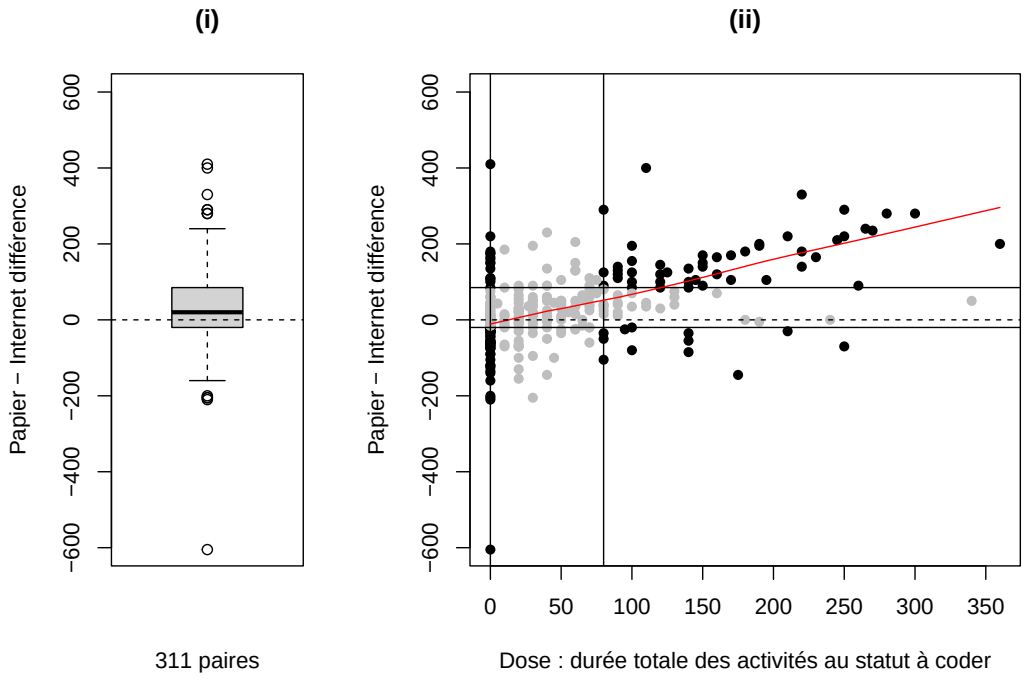
Deux analyses de l'effet de mesure de la collecte papier

La figure 3 (i) représente la différence de durées de trajet entre les carnets papier et

31. qui ne seraient renseignées qu'avec une seule plage horaire sans distinction du temps de trajet associé.
32. calculée par agrégation des plages horaires identifiées comme telles

numérique, au sein de chaque paire, calculées par agrégation de la durée des plages horaires correspondantes après codification du carnet papier. Comme nous l'avons mentionné au début de cette section, en utilisant la statistique du Wilcoxon signé du rang, l'inférence aléatoire, c'est-à-dire en l'absence de biais de sélection, rejette l'hypothèse nulle de l'absence d'effet de la collecte papier sur la durée de trajet. Nous en reprécisons cependant la démarche et les résultats afin que le lecteur puisse par la suite faire le lien avec l'approche qui tient compte de « l'intensité du traitement ».

FIGURE 3 – Différences de durées de trajet (papier - internet) au sein des 311 paires



Note : La figure (i) représente la différence de durée de trajets (papier - internet) au sein des 311 paires. La figure (ii) représente ces différences au sein des paires, Y_i en fonction de la durée totale d_i des plages horaires « autres trajets » dans le carnet du répondant papier, en contrastant en noir, celles supérieures ou inférieures aux 1er et au 3e quartiles de chaque distribution. La ligne rouge correspond à la régression locale (*LOcally Estimated Scatterplot Smoothing*) associée.

En l'absence de biais de sélection, comme explicité dans la section 3.1, l'analyse s'appuie sur l'inférence aléatoire c'est-à-dire en considérant une assignation aléatoire des répondants à un mode de collecte dans chacune des 311 paires, par exemple en tirant à pile ou face 311 fois. Il existe 2^{311} assignations possibles à un mode de collecte, $\boldsymbol{k}_1, \ldots, \boldsymbol{k}_{2^{311}}$, dont la probabilité de distribution est $\boldsymbol{p} = (2^{-311}, \ldots, 2^{-311})$ puisque chaque assignation a la même probabilité de réalisation. Comparer les répondants avec le carnet numérique et le carnet papier, en ignorant la durée totale des plages horaires à coder i.e. les « doses de traitement » avec la statistique de Wilcoxon signé du rang conduit à une p-value pour l'hypothèse nulle d'absence d'effet du mode de collecte de $1{,}4 \times 10^{-8} < 0{,}0001$. Du point de vue de l'inférence aléatoire, cela signifie que moins d'une assignation sur 10 000 pourrait conduire à observer une statistique aussi élevée, sous l'hypothèse d'absence d'effet du mode de collecte et dans le cas d'une assignation aléatoire des modes de collecte.

Cependant le protocole mis en œuvre ne correspondait pas à une assignation aléatoire du mode de collecte aux *répondants* mais aux *enquêtés*, ce qui nous impose de considérer l'impact d'un biais de sélection endogène après appariement sur les inférences obtenues avec une analyse de sensibilité. Lorsque $\Gamma > 1$, l'analyse de sensibilité ne considère plus

une seule probabilité de distribution des 2^{311} traitements possibles $\boldsymbol{k}_1, \ldots, \boldsymbol{k}_{2^{311}}$, mais un ensemble $\mathcal{P}_\Gamma$ de probabilités de distribution $\boldsymbol{p}$. Chaque probabilité de distribution $\boldsymbol{p} \in \mathcal{P}_\Gamma$ possible fournit une p-value pour le test de Wilcoxon, et nous nous intéressons à la p-value la plus élevée, $\overline{P}_\Gamma$, puisque l'on ne connaît pas la vraie probabilité de distribution associée à l'assignation au mode de collecte observée. À $\Gamma = 1{,}7$, la p-value maximale est $\overline{P}_\Gamma = 0{,}049$, c'est-à-dire qu'une caractéristique inobservée qui doublerait les chances (*odds*) d'avoir répondu avec le carnet papier et multiplierait par 8 celles d'observer une différence de durées de trajet positive pourrait expliquer les différences observées.

Nous disposons d'une autre comparaison dans cette étude qui s'intéresse cette fois à la différence de durées de trajet entre le répondant papier et le répondant internet *au sein d'une paire* en fonction de la durée totale des plages horaires à coder dans le carnet papier. Observe-t-on des différences de durée de trajet plus importantes dans les paires où le répondant avec le questionnaire papier a déclaré un volume horaire important de plages horaires à coder ?

La figure 3 (ii) représente la différence de durées de trajet au sein d'une paire en fonction de la durée totale des « autres trajets » déclarée par le répondant au carnet papier. La ligne rouge correspond à la régression locale associée : celle-ci s'éloigne de 0 au fur et à mesure que la durée totale des plages à coder augmente. Ce résultat suggère une tendance des paires à présenter une différence élevée de durées de trajet lorsque le répondant avec le questionnaire papier déclare un temps important dans les plages horaires à coder. Sur cette même figure, les lignes verticales et horizontales correspondent aux 1er et au 3e quartiles des deux distributions considérées. Ainsi, les points (en noir) en haut à droite du graphique représentent des paires au sein desquelles la différence de durée est importante [33] et pour lesquelles la durée totale des plages horaires à coder est élevée [34]. Le test *crosscut* sur l'odds-ratio, dont nous détaillons le calcul en Annexe E, s'intéresse aux sous-populations définies par les quartiles dont le nombre de paires correspondant est donné dans le tableau 10. L'odds-ratio estimé par ce test est de $12{,}8 = (48 \times 48)/(12 \times 15)$, ce qui révèle une association forte entre la durée totale des plages horaires à coder déclarée par le répondant au questionnaire papier et les différences de durée de trajet observées au sein des paires. La p-value associée à ce test obtenue par inférence aléatoire est de $2{,}5 \times 10^{-10}$ (cf. Annexe E).

Tableau 10 – Nombre de paires associées au test *crosscut* de la figure 3 (ii)

Différence de durées	Durée totale, d_i des activités à coder		
Y_i	$d_i \leq 0$	$d_i \geq 80$	Total
$Y_i \geq 85$	12	48	60
$Y_i \leq -20$	48	15	63
Total	60	63	123

Note : Le tableau reporte le nombre de paires des quatre sous-ensembles définis en fonction de la différence (Y_i) de durées de trajet entre le répondant papier et le répondant internet d'une part et la durée totale (d_i) des activités au statut à coder du répondant papier d'autre part. Ainsi, il y a 48 paires pour lesquelles $Y_i \geq 85$ et $d_i \geq 80$, correspondant aux 48 points noirs du coin en haut à droite de la figure (ii) de la figure 3. Les paires associées aux points gris ne sont pas comptabilisés dans les chiffres reportés dans ce tableau (voir annexe E).

Un tel calcul correspond à un protocole fictif d'assignation aléatoire des durées totales des plages horaires à coder entre les différents répondants au carnet papier. Il existe 311 ! assignations différentes, $\boldsymbol{h}_1, \ldots, \boldsymbol{h}_{311!}$ des durées totales de plages horaires à coder entre

33. i.e. supérieure au 3e quartile de la distribution des différences de durées de trajet observées au sein des paires.

34. i.e. supérieure au 3e quartile de la distribution correspondante.

les paires, dont la probabilité de distribution est $\boldsymbol{p}' = (1/311!, \ldots, 1/311!)$ puisque chaque assignation a la même probabilité de réalisation dans le cadre d'une expérience aléatoire. Dit autrement, dans une expérience aléatoire, les « chances » (*odds*) de déclarer un temps total d de plages horaires à coder dans une paire plutôt que d' sont les mêmes pour toutes les paires. La p-value $2{,}5 \times 10^{-10}$ du test nous indique que moins d'une assignation sur 100 000 aurait pu produire une statistique de *crosscut* aussi élevée que celle observée, si l'hypothèse nulle H_0 d'absence d'effet du mode de collecte est vraie dans l'expérience aléatoire considérée.

Il est cependant nécessaire de questionner ce résultat puisque les durées totales de plages horaires à coder n'ont bien sûr pas été affectées aléatoirement entre les paires. Le modèle d'analyse de sensibilité présenté à la section 3.2 peut être adapté en supposant cette fois que les « chances » (*odds*) d'une durée totale d au lieu de d' à coder diffère d'une paire à une autre d'un facteur au plus $\Gamma' \geq 1$ (pour une formalisation explicite, voir Rosenbaum, 2016), où $\Gamma' = 1$ correspond à une expérience aléatoire. Comme précédemment, sous cette hypothèse, il n'existe plus une seule probabilité de distribution associée aux 311 ! assignations possibles, mais un ensemble $\mathcal{P}'_{\Gamma'}$ de probabilités de distributions $\boldsymbol{p}'$. Chaque $\boldsymbol{p}' \in \mathcal{P}'_{\Gamma'}$ donne lieu à une p-value pour le test *crosscut* et nous nous intéressons à la p-value la plus élevée, $\overline{P}'_{\Gamma'}$. Pour $\Gamma' = 5{,}7$, $\overline{P}'_{\Gamma'} = 0{,}049$, c'est-à-dire qu'une caractéristique inobservée qui multiplierait par 10 les chances (*odds*) d'avoir répondu une durée totale à coder élevée et multiplierait par 10 celles d'observer une différence de durées de trajet positive pourrait expliquer les différences observées. .

Facteurs d'évidence

Dans ce qui précède, la première analyse se concentre sur les différences entre un répondant papier et un répondant internet. La seconde analyse, elle, se demande si l'on observe des différences importantes parmi les paires où le répondant papier déclare un temps important dans des plages horaires à coder, c'est-à-dire si les différences de durées de trajet sont corrélées à la durée totale des plages horaires à coder. Chacune semble fournir un résultat probant sur l'existence d'un effet de la collecte avec le questionnaire papier sur la durée de trajets déclarés, mais chacune reste sujette à un possible biais de composition. Néanmoins, il faut noter que sur ce point elles ne sont pas identiques. En effet, la première analyse est sensible à un biais inobservé *entre les répondants à deux modes de collecte* qui détermine qui répond par carnet numérique et qui répond avec le carnet papier : par exemple le fait d'effectuer des trajets plus longs quotidiennement, comme le suggère le résultat sur la durée totale de trajet déclaré à la fin du questionnaire. Elle n'est cependant pas sensible à un biais qui détermine parmi les répondants au carnet numérique qui déclare plus ou moins de durée parmi les « autres trajets ». C'est l'exact contraire pour la seconde analyse. En cela, contrairement par exemple à la comparaison des résultats d'une analyse de sensibilité avec différentes statistiques de test pour lesquels le biais de sélection endogène est de même nature, les deux analyses diffèrent. Peut-on pour autant combiner ces analyses (et leurs résultats) afin de renforcer notre conviction d'un effet de mesure (de la collecte papier par rapport à la collecte numérique), par exemple en obtenant des conclusions moins sensibles à la présence d'un biais de sélection endogène, que celles obtenues en considérant séparément chacune des analyses ?

Y répondre implique de s'interroger sur l'indépendance des deux analyses. Les graphiques de la figure 3 suggèrent des intuitions contradictoires sur cette question. Schématiquement, les différences de la figure 3 (i) ne permettent pas « d'anticiper » la représentation graphique de la figure 3 (ii). Néanmoins, à l'inverse, si l'on projette les points de la figure 3 (ii), il est possible de retrouver la figure 3 (i). En un sens, la première analyse pourrait apparaître comme « indépendante » de la seconde. D'un autre point de vue, les deux analyses

apparaissent fortement corrélées. L'objectif est donc de pouvoir combiner les résultats des analyses de sensibilité de ces deux approches *sans hypothèse supplémentaire* sur leurs relations éventuelles, notamment d'indépendance, dans l'espoir que les conclusions de ces deux comparaisons se renforcent l'une l'autre. En particulier, l'hypothèse supplémentaire que l'on souhaite le plus *éviter* est que la personne la plus à même de répondre avec le carnet papier - par exemple parce qu'elle effectue quotidiennement de long trajets - *n'est pas* aussi la plus à même de déclarer une durée totale élevée d'activités à coder : de fait, il peut tout à fait être envisageable que cela soit le cas.

Sans entrer dans les détails dans ce document de travail [35] Rosenbaum (2017b) démontre que les bornes supérieures des deux p-values obtenues avec les deux analyses précédentes, que Rosenbaum appelle des *facteurs d'évidence*, peuvent être combinées *comme si* elles provenaient d'analyses indépendantes. Il ne s'agit pas de considérer que les deux analyses sont indépendantes, mais de souligner que du point de vue de la p-value, le cas le plus défavorable correspond à la situation d'indépendance, même si les distributions jointes sont fortement corrélées. Plus formellement, il établit la proposition suivante

Proposition. *Si l'hypothèse* H_0 *est vraie, le biais dans la première analyse d'au plus* Γ *et le biais dans la deuxième analyse d'au plus* Γ'*, alors la paire des bornes supérieures de p-values,* $(\bar{P}_\Gamma, \bar{P}_{\Gamma'})$ *est stochastiquement plus grande que la distribution uniforme sur le carré* $[0,1] \times [0,1]$*, ce qui implique que* $\Pr(\bar{P}_\Gamma \leq \alpha, \bar{P}_{\Gamma'} \leq \alpha') \leq \alpha\alpha'$.

Intuitivement, cette proposition établit que les conclusions des deux analyses de sensibilité précédentes peuvent se renforcer l'une l'autre comme nous l'illustrons avec les résultats suivants [36].

35. Nous renvoyons pour cela le lecteur intéressé à l'article correspondant et à Rosenbaum (2020).

36. Précisons que cette proposition ne s'applique pas à la combinaison de tous types d'analyses. En annexe F, nous détaillons succinctement la validité de la démarche mise en œuvre dans notre cas. Cependant, rappelons que comme nous l'avons rapidement évoqué, les résultats obtenus avec différentes statistiques de test ne constituent pas des facteurs d'évidence. À l'inverse, par exemple, la comparaison des résultats d'unités traitées et ceux de deux groupes de contrôle de nature différente permet de constituer trois facteurs d'évidence en considérant les différences de résultats entre une unité traitée appariée avec séparément avec un membre de chaque groupe de contrôle associée à la différence de résultats entre deux unités de contrôle issues chacune de l'un des deux groupes. Pour les *design* d'analyses permettant la constitution de facteurs d'évidence, voir Rosenbaum (2021).

TABLEAU 11 – p-value combinée des facteurs d'évidence d'un effet de la collecte papier sur la durée de trajets

Wilcoxon Γ		Test crosscut Γ′					
		1	4	*5.7*	6	7	∞
	Crosscut →	0,000	0,005	0,049	0,127	0,297	0,906
	Wilcoxon ↓						
1	0,000	0,000	0,000	0,000	0,000	0,000	0,000
1.3	0,000	0,000	0,000	0,000	0,000	0,000	0,001
1.5	0,005	0,000	0,000	0,002	0,002	0,004	**0,025**
1.7	0,049	0,000	0,002	0,012	0,014	**0,023**	0,118
1.9	0,192	0,000	0,008	**0,035**	**0,041**	0,065	0,351
∞	1.000	0,000	**0,028**	0,109	0,127	0,202	1,000

Note : p-values pour deux facteurs d'évidence combinées avec une technique de méta-analyse, le produit tronqué de Zaykin *et al.* (2002), tronqué à 0,2. Zaykin *et al.* (2002) propose de combiner de $\boldsymbol{p} = p_1, \ldots, p_k$, k p-values indépendantes en prenant le produit des p-values qui ne sont pas supérieures à une valeur c. Pour $c = 1$, la méthode est identique à la méthode de Fisher pour combiner des p-values indépendantes. La truncature permet d'accroître la puissance du test. Pour chaque statistique de test, nous avons indiqué en italiques, les valeurs de Γ_{max} et Γ'_{max} à partir desquelles respectivement les conclusions des deux analyses deviennent sensibles à la présence d'un biais de composition caché.

Le tableau 11 reporte les résultats obtenus en combinant les p-values associées à la statistique de Wilcoxon pour différentes valeurs de Γ pour la première analyse, avec celles associées à la statistique de *crosscut* pour différentes valeurs de Γ′ pour la seconde analyse. La combinaison des deux p-values pour chaque couple (Γ, Γ') est réalisée avec des techniques de méta-analyse *comme si* les deux analyses étaient indépendantes et correspondaient aux résultats de deux études différentes, issues de deux échantillons différents, etc. Plus précisément, la méthode pour combiner les deux p-values est leur produit tronqué qui est une généralisation de la méthode de Fisher de combinaison des p-values. La méthode de Fisher consiste à prendre le produit des deux p-values et s'intéresse à sa distribution pour obtenir une nouvelle p-value. Le produit tronqué considère le produit des p-values qui ne sont pas supérieures à une valeur de troncature ici 0,2.

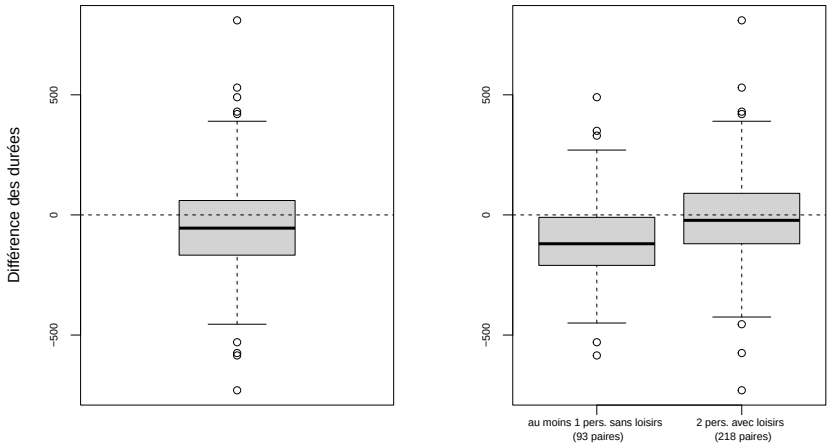
Le premier résultat est que pour des biais d'ampleur supérieurs à ceux considérés pour chacune des analyses prises séparément, respectivement $\Gamma = 1{,}9 > 1{,}7$ et $\Gamma' = 6 > 5{,}7$, la p-value combinée 0,041 nous permet de rejeter l'hypothèse d'une absence d'effet de la collecte papier sur les durées de trajet déclarées. Cela signifie que les conclusions de chaque analyse se renforcent l'une l'autre et que nous disposons en les considérant conjointement une conviction plus forte (et de ce point de vue quantifiable) de l'existence d'un effet de collecte au sens où les biais pour rejeter cette hypothèse doivent être plus importants. Mais plus encore, nous pouvons questionner la pertinence de nos résultats, si par exemple on considère que la seconde analyse est complètement sans intérêt parce qu'elle est complètement biaisée, c'est-à-dire avec une valeur de $\Gamma' \to \infty$. Il s'avère que l'hypothèse d'un effet du mode de collecte serait toujours rejetée si $\Gamma = 1{,}5$ (p-value combinée = 0,025). Cela signifie que dans ce cas de figure nous disposons toujours d'un résultat probant même en ne tenant pas compte de notre seconde analyse. De même, si la première analyse est complètement biaisée, la seconde permet toujours de rejeter l'hypothèse d'absence d'effet pour $\Gamma' = 4$. Il ressort de ces multiples résultats qu'il apparait vraisemblable de considérer que la collecte papier a un effet sur la durée totale de trajets calculée en agrégeant les durées des différents déplacements effectués dans la journée par l'enquêté, et ce même en présence d'une sélection

endogène liée à la plus grande mobilité des répondants avec le carnet papier comparée à celle des répondants avec le carnet numérique. Cet effet de mesure s'expliquerait probablement par la défaillance de l'algorithme de codification pour des activités où le libellé est difficile à interpréter.

4.3 Impact sur la durée des loisirs

Concernant le temps alloué aux loisirs, dans l'hypothèse d'absence de biais de sélection endogène, les répondants déclarent en moyenne 45 min ($[-67; -24]$ à 95 %) de loisirs en moins quand ils répondent sur internet comme l'illustre la figure 4 (gauche). Cet écart refléterait des différences dans la probabilité de déclarer au moins une activité de loisirs (cf. figure 4 droite) : de fait, 25 % des répondants en ligne ne déclarent aucune activité de loisirs alors que cette absence ne représente que 7 % des répondants utilisant le carnet papier.

FIGURE 4 – Différence de durées de loisirs (internet - papier)

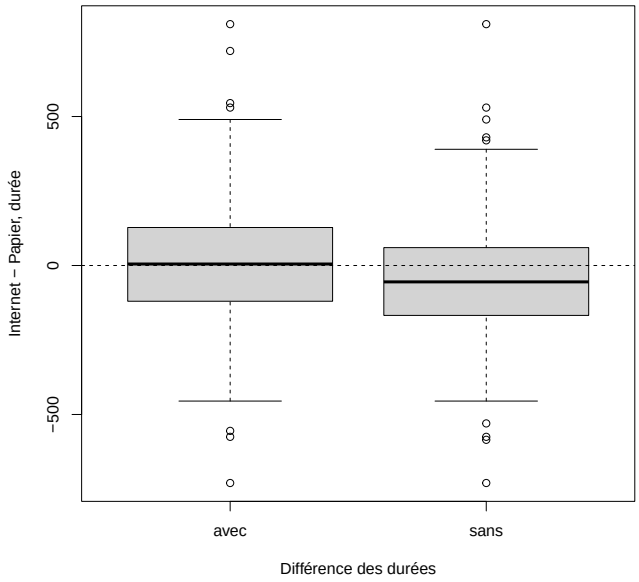


En appliquant une démarche analogue à celle mise en œuvre pour la durée de sommeil (cf. section 4.1), il est possible de tester l'hypothèse d'un effet de la collecte par internet sur la probabilité de ne pas déclarer de temps de loisirs et d'estimer la proportion d'effets attribuables à cet effet. L'analyse de sensibilité révèle que l'hypothèse d'un effet du mode internet est très peu sensible à l'hypothèse d'un biais de sélection endogène puisque $\Gamma_{max} = 2{,}8$. Dit autrement, même en considérant que la probabilité d'assignation au mode de collecte internet au sein d'une paire n'est que de 0,30 au lieu de 0,5, nous pourrions toujours conclure à un effet de la collecte par internet significatif.

Après appariement, 87 paires sur les 311 (soit 28 %) sont discordantes, dont 71 du fait d'une absence de plages horaires de loisirs déclarés par l'enquêté par internet (soit 82 % des paires discordantes). En absence de biais de sélection endogène ($\Gamma = 1$), nous sommes sûrs à 95 % que le mode de collecte internet est à l'origine de 62 % des 71 paires discordantes, soit 14 % des 311 paires au total. Même en considérant l'existence d'un biais de sélection endogène conséquent, par exemple avec $\Gamma = 2$, on peut encore affirmer à 95 % que la collecte par internet conduit à ne pas déclarer de temps de loisirs dans 6 % des 311 paires.

Comment pouvons-nous expliquer un tel effet du mode de collecte internet sur l'absence de déclaration d'activités de loisirs ? Dans l'enquête Emploi du temps, chaque activité est intégrée dans des catégories et sous-catégories plus générales. La catégorie « Loisirs »

FIGURE 5 – Différences de durées de loisirs (internet - papier) au sein des 311 paires - avec ou sans correction du code activité

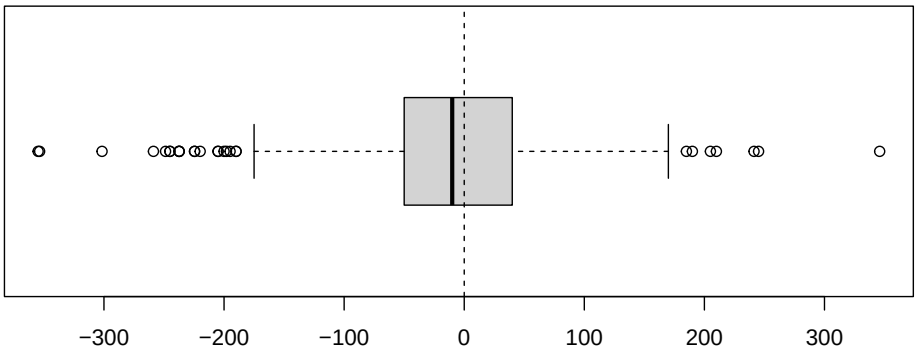


comporte les activités sportives, les activités extérieures (comme les promenades), et les activités culturelles à l'extérieur, ou chez soi comme lire, regarder la télévision, jouer aux jeux vidéos. Ces dernières activités, personnelles et solitaires, ne sont donc pas, dans la nomenclature, intégrées à la catégorie « Temps personnel et physiologique » qui regroupent plutôt les activités de sommeil, de repas ou le temps consacré à l'hygiène personnelle. Celle-ci comporte cependant une sous-catégorie dont la dénomination « autres activités personnelles » peut prêter à confusion. En effet, avec le carnet en ligne (voir figure 11 Annexe G), les enquêtés répondent majoritairement par un *selecter* (80 % des activités) qui affiche en premier les catégories générales (Temps personnel, Travail et études, Tâches domestiques, etc.) et ensuite un détail des sous-catégories. Or la catégorie « Loisirs » apparaît en bas de la page présentant les catégories générales, soit après la catégorie « Temps personnel (et physiologique) ». Une telle ergonomie du sélecteur d'activité pourrait ne pas inciter les répondants à faire défiler les catégories pour atteindre celle des « Loisirs » située plus bas. Dès lors, il serait vraisemblable que les enquêtés choisissent la sous-catégorie des « autres activités personnelles » au sein de la catégorie « Temps personnel et physiologique » affichée en premier, au lieu de parcourir celles proposées dans la catégorie « Loisirs », pour déclarer les plages horaires correspondant au temps passé devant la télévision, à lire, à jouer aux jeux vidéos, etc.

De fait, les répondants par internet déclarent moins d'activités de loisirs comme « regarder la télévision » et « ne rien faire » comparativement aux répondants avec le carnet papier. À l'inverse, les répondants avec le carnet numérique déclarent plus fréquemment des plages horaires dans la sous-catégorie « autres activités personnelles » de la catégorie « Temps personnel et physiologique » (4,5 % des activités reportées) que les répondants avec le carnet papier (moins de 0,1 % des activités reportées).

En suivant cette hypothèse, il est possible de corriger cette potentielle erreur de classification en imputant toutes les « autres activités personnelles » en activités de Loisirs dans les carnets numériques. Comme l'illustre la figure 5, avec cette nouvelle codification, l'effet du mode de collecte estimé précédemment n'est plus significatif, tendant à confirmer

FIGURE 6 – Différences de durées des repas au sein des 311 paires (internet - papier)



Note : le graphique représente la distribution de durées de repas (internet - papier) au sein des 311 paires en utilisant la transformation réciproque proposée par Rosenbaum (2022). Schématiquement, la transformation ne modifie pas 90 % des valeurs, permet une représentation des observations extrêmes distinctes sans qu'elles conduisent à une compression visuelle de la partie centrale de la distribution. Cette transformation n'affecte pas la symétrie de la distribution : formellement dans le cas d'une affectation aléatoire, en l'absence d'effet du traitement la distribution reste symétrique. Ces propriétés assurent que les rangs des valeurs absolues des différences sont inchangées, de telle sorte que les statistiques signées du rang associées, comme la statistique de Wilcoxon, sont inchangées par la transformation.

l'hypothèse d'un effet lié au *design* du questionnaire.

4.4 Impact sur la durée des repas

Comme l'illustre la figure 6, sous l'hypothèse d'absence de biais de sélection endogène, les répondants déclarent en moyenne 16 min (IC à 95 % $[-28; -4]$) de temps de repas en moins quand ils répondent sur internet. Cet effet est cependant sensible à l'hypothèse d'un biais de sélection endogène ($\Gamma_{max} = 1{,}2$).

Si l'on considère le nombre de plages horaires consacrées à des repas, l'hypothèse d'un effet de mesure de la collecte internet est peu sensible à un biais de sélection endogène, puisque, quelle que soit la statistique de test utilisée, Γ_{max} est proche de 2 (cf. tableau 12).

TABLEAU 12 – Analyse de sensibilité Γ_{max} sur la durée totale des repas et leur nombre

Statistique	durée totale	nombre de plages
Moyenne	1,2	2,0
Wilcoxon	1,1	1,8
Stephenson (5,5,5)	1,2	2,2

De fait, 11 % des répondants en ligne déclarent plus de 3 repas par jour (nombre de repas attendu) contre 25 % pour les répondants papier. À l'inverse, une proportion similaire déclarent exactement 3 repas quel que soit le mode de collecte (47 % par internet *versus* 49 % par questionnaire papier). Compléter le carnet par papier a un effet certain sur le nombre de repas déclaré, mais celui-ci est restreint à une proportion faible des enquêtés. Au total, ces résultats suggèrent un impact limité ou nul sur la durée totale consacrée aux

repas, mais une déclaration plus fréquente, avec le carnet papier, de plages horaires courtes de « repas » qui pourraient correspondre à une pause café, *snacking*, etc.

5 Conclusion

Notre étude vise à questionner les difficultés que pose l'utilisation d'un essai croisé dans une enquêté test menée pour étudier l'existence d'effets de mesure avec différents modes de collecte. Comme nous l'avons souligné, la présence d'une possible sélection endogène fragilise les conclusions issues de l'analyse d'un essai croisé puisque celle-ci suppose que la non-participation soit MAR ou MCAR. Il est néanmoins possible de discuter l'existence d'effets de mesure en comparant les réponses données en première interrogation par des personnes similaires enquêtées avec un mode de collecte différent. Pour cela, il convient de mener une analyse de sensibilité (Rosenbaum, 2002b, 2010) sur les conclusions obtenues en supposant l'absence de biais lié à une possible sélection endogène. D'un point de vue méthodologique, notre étude complète la présentation de cette approche par Court et Quantin (2024) avec des données discrètes en considérant le cas de variables d'intérêt continues. Nous détaillons notamment comment déterminer une estimation de l'effet de mesure et l'intervalle de confiance associé, et leur extension dans le cadre d'une analyse de sensibilité. Par ailleurs, cette étude explicite la notion d'effets attribuables au mode de collecte proposée par Rosenbaum (2002a) qui permet une quantification de l'effet de mesure complémentaire au plus traditionnel effet moyen.

Toutes ces méthodes sont appliquées ici à l'enquête test sur l'enquête Emploi du temps, mise en œuvre en 2023, pour étudier l'existence d'un effet de mesure sur les durées déclarées pour différentes activités avec un carnet numérique en lieu et place du carnet papier. Nos résultats mettent en évidence qu'il est difficile de rejeter l'hypothèse d'un effet de mesure du mode de collecte utilisé sur la durée déclarée de certaines activités même en présence de sélection endogène. Cet effet de mesure n'impacterait cependant souvent que les durées d'un nombre restreint d'enquêtés. Plus précisément, la sous-déclaration d'heures de sommeil serait corrélée à la non-complétion de la totalité de l'emploi du temps. En effet, avec le carnet numérique, certains répondants ne déclarent pas leurs dernières activités de la journée, ce qui réduit la période couverte renseignée et probablement le temps de sommeil déclaré. En ce qui concerne la durée de trajets, nos résultats suggère que la difficile codification pour le carnet papier des activités liées à un déplacement autre que ceux du domicile au travail conduirait à sur-estimer la durée totale des trajets déclarés : à titre d'exemple, lorsque l'enquêté déclare « partir faire ses courses », le temps déclaré serait considéré comme intégralement consacré aux déplacements, alors qu'il inclurait aussi le temps passé aux achats. La sous-déclaration en ligne d'activités de loisirs pourrait s'expliquer par l'ergonomie du sélecteur de l'interface qui conduirait les répondants à déclarer leurs activités de loisirs comme des activités associées à une autre catégorie d'activité. Enfin, la sous-déclaration en ligne du nombre de repas pourrait refléter une différence de comportement déclaratif pour les pauses de courte durée comme le temps consacré à prendre un café durant la journée en dehors des heures de repas. L'ampleur, même limitée, de ces effets peut néanmoins être questionnée. En effet, les approches implémentées dans cette étude ne tiennent pas compte d'une possible interférence entre unités enquêtées au sein d'un ménage : par exemple nos estimations n'intègrent pas la possible complétion d'un carnet par ou avec l'aide d'un tiers.

Références

S. Basu et S. Santra : A joint model for incomplete data in crossover trials. *Journal of Statistical Planning and Inference*, 140:2839–2845, 2010.

L. Court et S. Quantin : Discuter l'existence d'un effet de sélection dans un cadre multimode grâce à une analyse de sensibilité : application aux enquêtes annuelles de recensement. *Documents de travail Insee*, M2024/03, 2024.

R. A. Fisher : *The Design of Experiments*. Oliver and Boyd, 1935.

Weang Kee Ho, John N.S. Matthews, Robin Henderson, Daniel Farewell et Lauren R. Rodgers : Dropouts in the AB/BA crossover design. *Statistics in Medicine*, 31:1675–1687, 2012.

J. Hodges et E. Lehmann : Estimates of location based on rank tests. *Annals of Mathematical Statistics*, 34:598–611, 1963.

P.J. Huber : Robust estimation of a location parameter. *Annals of Mathematical Statistics*, 35:73–101, 1964.

E.L. Lehmann : *Nonparametrics : Statistical methods based on ranks*. San Francisco : Holden-Day, 1975.

J. Maritz : *Distribution-Free Statistical Methods*. London : Chapman & Hall, 1981.

J. S. Maritz : Exact robust confidence intervals for location. *Biometrika*, 66:163–166, 1979.

Jerzy Neyman : On the application of probability theory to agricultural experiments. Essay on principles. Section 9. *Roczniki Nauk Roiniczych*, Tom X:1–51, 1923. Réimprimé en anglais dans *Statistical Science*, 1990, 5, 463-480.

M. Pagano et D. Tritchler : Obtaining Permutation Distributions in Polynomial Time. *Journal of the American Association*, 78:435–440, 1983.

H.I. Patel : Analysis of incomplete data in a 2-period crossover design with reference to clinical trials. *Biometrika*, 72:411–418, 1985.

Paul R. Rosenbaum : On permutation tests for hidden biases in observational studies. *Annals of Statistics*, 17:643–653, 1989a.

Paul R. Rosenbaum : The role of known effects in observational studies. *Biometrics*, 45:557–569, 1989b.

Paul R. Rosenbaum : Effects Attributable to Treatment : Inference in Experiments and Observational Studies with a Discrete Pivot. *Biometrika*, 88(1):219–231, March 2001.

Paul R. Rosenbaum : Attributing Effects to Treatment in Matched Observational Studies. *Journal of the American Statistical Association*, 95(457):183–192, 2002a.

Paul R. Rosenbaum : *Observational Studies*. Springer Series in Statistics, 2 édition, 2002b.

Paul R. Rosenbaum : Confidence Intervals for Uncommon but Dramatic Responses to Treatment. *Biometrics*, 63:1164–1171, June 2007a.

Paul R. Rosenbaum : Sensitivity Analysis for m-Estimates, Tests, and Confidence Intervals in Matched Observational Studies. *Biometrics*, 63:456–464, June 2007b.

Paul R. Rosenbaum : *Design of observational studies*. Springer Series in Statistics, 2010.

Paul R. Rosenbaum : A New u-Statistic with Superior Design Sensitivity in Matched Observational Studies. *Biometrics*, 67:1017–1027, 2011.

Paul R. Rosenbaum : Weighted M-statistics With Superior Design Sensitivity in Matched Observational Studies With Multiple Controls. *Journal of the American Statistical Association*, 109(507):1145–1158, september 2014.

Paul R. Rosenbaum : The crosscut statistic and its sensitivity to bias in observational studies with ordered doses of treatment. *Biometrika*, 72:175–183, 2016.

Paul R. Rosenbaum : *Observation and Experiment : An introduction to causal inference*. Harvard University Press, 2017a.

Paul R. Rosenbaum : The General Structure of Evidence Factors in Observational Studies. *Statistical Science*, 32(4):514–530, November 2017b.

Paul R. Rosenbaum : *sensitivitymv : Sensitivity Analysis in Observational Studies*, 2018. URL https://CRAN.R-project.org/package=sensitivitymv. R package version 1.4.3.

Paul R. Rosenbaum : *DOS2 : Design of Observational Studies, Companion to the Second Edition*, 2019. URL https://CRAN.R-project.org/package=DOS2. R package version 0.5.2.

Paul R. Rosenbaum : *Design of Observational Studies*. Springer Series in Statistics, 2 édition, 2020.

Paul R. Rosenbaum : *Replication and Evidence Factors in Observational Studies*. Chapman and Hall/CRC, 2021.

Paul R. Rosenbaum : A New Transformation of Treated-Control Matched-Pairs Differences for Graphical Display. *The American Statistician*, 76(4):346–352, 2022.

Paul R. Rosenbaum : Sensitivity analyses informed by tests for bias in observational studies. *Biometrics*, 79:475–487, 2023.

Paul R. Rosenbaum et Jeffrey H. Silber : Amplification of Sensitivity Analysis in Matched Observational Studies. *Journal of the American Statistical Association*, 104 (488):1398–1405, December 2009.

D.B. Rubin : Estimating causal effects of treatments in randomized and non randomized studies. *Journal of Educational Psychology*, 66(5):688–701, 1974. URL https://doi.org/10.1037/h0037350.

D.B. Rubin : Inference and Missing Data. *Biometrika*, 63:581–592, 1976.

W.R. Stephenson : A general class of one-sample parametrics test statistics based on subsamples. *Journal of the American Association*, 76:960–966, 1981.

Jeffrey M. Wooldridge : *Econometric Analysis of Cross Section and Panel Data*. MIT Press, Cambridge, 2 édition, 2002.

D.V. Zaykin, L.A. Zhivotovsky, P.H. Westfall et B.S. Wier : Truncated product method combining P-values. *Genetic Epidemiology*, 22:170–185, 2002.

Annexes

A Appariement optimal par paires

L'appariement sur le score de propension tend à constituer des groupes traité et de contrôle qui présentent des distributions similaires pour les variables observées. Cependant, si les unités enquêtées au sein de chaque paire ainsi constituée ont un score de propension (estimé) proche, elles peuvent néanmoins différer fortement sur des covariables spécifiques. Afin de constituer des paires d'unités enquêtées plus semblables, nous construisons une distance qui pénalise les différences importantes de scores de propension, puis nous constituons des paires d'unités aussi similaires que possibles, en utilisant un algorithme d'optimisation.

Dans un cadre multivarié, il est important de tenir compte des différences d'unités de mesure des covariables dans le choix de la distance utilisée. La distance de Mahalanobis généralise à plusieurs variables la notion de mesure de la distance en nombre d'écart-types en tenant compte aussi des corrélations entre ces variables. Formellement, si $\hat{\Sigma}$ est la matrice de covariance empirique des variables $\mathbf{x}$ alors la distance de Mahalanobis entre deux unités enquêtées, k et l associées à ces covariables est $(\mathbf{x}_k - \mathbf{x}_l)^T \hat{\Sigma}^{-1} (\mathbf{x}_k - \mathbf{x}_l)$. Cependant, si une covariable présente des observations aberrantes ou une distribution très étirée, son écart-type peut être important et la distance de Mahalanobis aura tendance à ignorer cette covariable dans l'appariement. De plus, avec des variables binaires, la variance est maximale lorsque la probabilité de l'évènement est de $\frac{1}{2}$, et minimale lorsqu'elle est proche de 0 ou de 1. Dès lors la distance de Mahalanobis accorde un poids plus important aux variables binaires mesurant des évènements rares, c'est-à-dire avec des probabilités proches de 0 ou de 1. Afin de contourner ces difficultés, nous privilégions la distance de Mahalanobis du rang. Concrètement, cette distance est calculée (i) en remplaçant les valeurs des covariables (chaque covariable étant considérée séparément) par leurs rangs, en appliquant un rang moyen en cas de valeurs identiques, et (ii) en ajustant la matrice de covariance des rangs des covariables avec une matrice diagonale dont les éléments sont les ratios des écarts-types des rangs des valeurs uniques par les écarts-types des rangs des valeurs y compris identiques. La première étape (i) permet de limiter l'influence des valeurs aberrantes et des longues queues de distribution. La présence de valeurs identiques réduit la variance des rangs : la seconde étape (ii) réajuste la matrice de variance covariance de telle sorte que chaque covariable a sa variance hors valeurs identiques. Une telle correction réduit l'influence usuelle avec la distance de Mahalanobis des covariables présentant de nombreuses valeurs identiques, comme les variables binaires mesurant un évènement rare.

Cette distance est usuellement calculée entre toutes les unités enquêtées dont la différence de scores de propension estimés $|\hat{e}(\mathbf{x}_k) - \hat{e}(\mathbf{x}_l)|$ n'excède pas, en valeur absolue, un caliper w, dont la largeur w est ici prise égale à la moitié de l'écart-type du score de propension estimé $\hat{e}(\mathbf{x})$. Cependant, à cause de la taille réduite de l'échantillon, il n'existe pas nécessairement d'appariement par paires pour lequel la condition $|\hat{e}(\mathbf{x}_k) - \hat{e}(\mathbf{x}_l)| \leq w$ entre deux unités k et l appariées soit respectée pour toutes les unités traitées. Dès lors, pour assurer un appariement de chaque unité enquêtée traitée, nous calculons la distance de Mahalanobis du rang entre toutes les unités, à laquelle nous ajoutons une fonction de pénalité lorsque la contrainte imposée par le caliper n'est pas respectée. Formellement, la pénalité utilisée dans cette étude est $1000 \times \max(0, |\hat{e}(\mathbf{x}_k) - \hat{e}(\mathbf{x}_l)| - w)$. Cette pénalité est nulle si $|\hat{e}(\mathbf{x}_k) - \hat{e}(\mathbf{x}_l)| \leq w$, mais elle est égale à $1000 \times (|\hat{e}(\mathbf{x}_k) - \hat{e}(\mathbf{x}_l)| - w)$ si $|\hat{e}(\mathbf{x}_k) - \hat{e}(\mathbf{x}_l)| > w$.

Les durées des différentes activités déclarées par une unité enquêtée dépendent du jour de la semaine où le questionnaire est rempli (week-end ou non) et de sa situation principale

(en emploi, au chômage[37], retraité ou préretraité, en incapacité de travailler, en études, au foyer ou autre situation). Il est donc important de pouvoir s'assurer qu'au sein de chaque paire, les unités enquêtées soient identiques sur ces caractéristiques observées. En ce qui concerne, le jour de la semaine, un appariement exact est imposé. Pour tenir compte de la situation principale, une deuxième pénalité est ajoutée à la distance évoquée précédemment. Si deux unités enquêtées k et l n'ont pas la même situation principale, la fonction de pénalité utilisée ajoute à la distance 10 fois la distance maximale observée (sur la distance préalablement calculée) si les deux unités diffèrent sur leur situation principale.

Une fois la matrice de distance constituée, un appariement optimal par paires est réalisé. Un tel appariement constitue des paires composées d'une unité enquêtée traitée et d'une unité de contrôle de telle sorte que la somme des distances au sein des paires soit minimale[38]. Dès lors, un appariement optimal cherchera à éviter les distances pénalisées en respectant autant que possible la contrainte de caliper ; lorsque cela n'est pas possible, il privilégiera un appariement où la violation de la contrainte du caliper est peu fréquente et la plus réduite possible. En ce qui concerne la situation principale, si un appariement exact est possible, il sera considéré ; sinon, un appariement aussi proche que possible d'un appariement exact sera effectué.

B Que veut dire juger de la qualité de l'appariement ?

Comme l'a explicité la partie méthodologique consacrée à l'analyse de sensibilité, l'inférence aléatoire pour mettre en évidence un effet causal ne repose pas sur la comparaison de personnes identiques[39], mais sur la distribution de la probabilité d'assignation au mode de collecte. De ce point de vue, si l'appariement vise à réduire les « différences » entre les deux échantillons, la question de sa qualité se pose en d'autres termes : ces écarts ont-ils été réduits suffisamment, c'est-à-dire par rapport à ceux que l'on aurait observés si nous avions fait une expérience complètement aléatoire. Une expérience aléatoire aurait « équilibré » les covariables inobservées comme observées. Dans le cas de l'appariement, bien évidemment, on ne peut que s'intéresser aux seules covariables observées.

Dans notre appariement, nous considérons 63 covariables. Si, dans le cadre d'une expérience aléatoire, on réalise un test de randomisation sur ces 63 covariables, alors la seule raison pour rejeter l'hypothèse nulle d'absence d'effet est le hasard. Ainsi on s'attend donc à obtenir $63 \times 0{,}05 = 3{,}15$ p-values $\leq 0{,}05$.

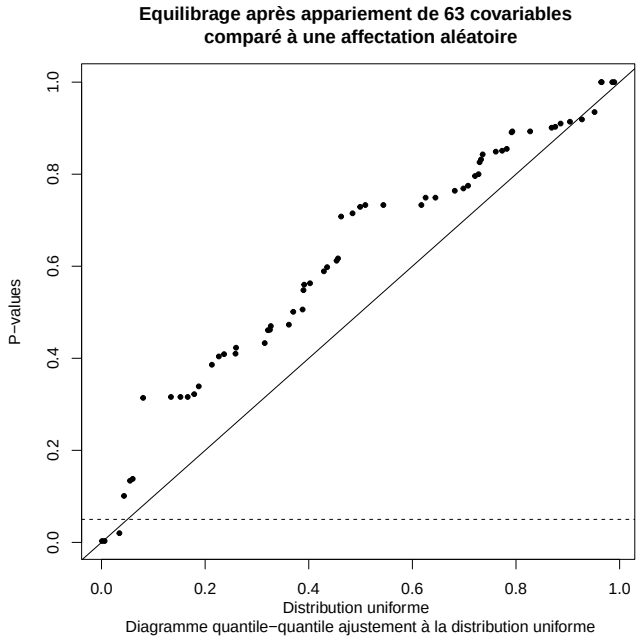
Comment la qualité de notre appariement se compare-t-elle à ce que nous aurions obtenu avec une randomisation complète ?

37. inscrit ou non à Pôle Emploi.

38. Ce problème de minimisation n'est pas immédiat parce que la plus « proche »unité de contrôle d'une unité traitée peut aussi être la plus proche unité de contrôle d'une autre unité traitée. Un choix *best-first* ou algorithme glouton (*greedy algorithm*) n'aboutira pas généralement à trouver l'appariement par paires optimal.

39. Comme le souligne Paul R. Rosenbaum dans sa présentation lors de la cérémonie de remise des récompenses du comité des présidents de sociétés statistiques en 2019, si l'on considère 63 covariables binaires, nous disposons $2^{63} = 9.2 \times 10^{18}$ catégories possibles, c'est-à-dire d'autant de types de personnes. Or, il n'y a que 8.0 milliards de personnes sur Terre. Cela signifie que, chaque fois que nous trouvons une personne dans une catégorie, il y a en moyenne 1.2×10^{9} catégories sans aucune personne. Il s'en déduit que l'on sait, de manière certaine, que nous ne comparerons pas des personnes identiques sur 63 covariables.

FIGURE 7 – Différences entre traités et contrôles sur les 63 covariables après appariement comparées à la situation qui aurait été observée après une expérience aléatoire



Comme illustré sur le graphique 7, l'équilibre obtenu grâce à l'appariement effectué pour cette étude sur les covariables observées est meilleur que celui qui aurait été obtenu par l'assignation aléatoire. Mais, encore une fois, cependant, il ne s'agit que des covariables observables.

C U-statistiques de Rosenbaum

On considère trois entiers $m, \underline{m}, \overline{m}$ avec $1 \leq \underline{m} \leq \overline{m} \leq m < I$; le triplet $m, \underline{m}, \overline{m}$ définit la U-statistique considérée. Son calcul est le suivant :

- On considère les sous-ensembles de m paires. Pour chaque sous-ensemble, les différences $V_i Y_i$ sont ordonnées selon leurs valeurs absolues $|Y_i|$,
- Une fois l'ordonnancement effectué, on comptabilise le nombre de $V_i Y_i$ positive, c'est-à-dire le nombre de différences (paires) où la réponse apportée par l'unité traitée est supérieure à celle de l'unité de contrôle, parmi celles numérotées $\underline{m}, \underline{m}+1, \ldots, \overline{m}$. Puis l'on agrège ce résultat sur l'ensemble des $\binom{I}{m}$ sous-ensembles.

Pour clarifier les choses, nous détaillons maintenant quelques exemples de U-statistique. La statistique (1,1,1) est la statistique du signe : elle totalise le nombre de paires où la réponse apportée par l'unité traitée est supérieure à celle de l'unité de contrôle. La statistique (m,m,m), qui correspond à la statistique de Stephenson (Stephenson, 1981), s'intéresse aux $\binom{I}{m}$ sous-ensembles $(Y_{i1}, \ldots, Y_{im})$ de m paires et compte le nombre de différences positives les plus élevées (puisque $\overline{m} = m$). La statistique (8,7,8) considère les $\binom{I}{8}$ sous-ensembles $(Y_{i1}, \ldots, Y_{i8})$ de 8 paires et compte le nombre de différences positives parmi les deux $|Y_{ik}|$ les plus élevées. Enfin, la statistique (20,16,19) examine tous les sous-ensembles de paires de taille 20, ignore le signe de la valeur $|Y_{ik}|$ la plus élevée, et comptabilise le nombre de différences positives Y_{ik} parmi les 4 plus grandes valeurs $|Y_{ik}|$ suivantes. Notons que, comme le remarque Rosenbaum (2011), la statistique (2,2,2) est analogue à la statistique de Wilcoxon signée du rang.

Formellement, les U-statistiques sont des statistiques signées du rang. Si l'on note a_i le rang de $|Y_i|$ pour $i = 1, \ldots, I$, la différence $V_i Y_i$ avec le rang a_i a la l^{e} plus grande valeur $|Y_i|$ dans $\binom{a_i-1}{l-1}\binom{I-a_i}{m-l}$ ensembles de taille m (Rosenbaum, 2011). Ainsi la statistique $(m,\underline{m},\overline{m})$ est :

$$T = \sum_{i=1}^{I} \text{sign}(V_i Y_i) q_i$$

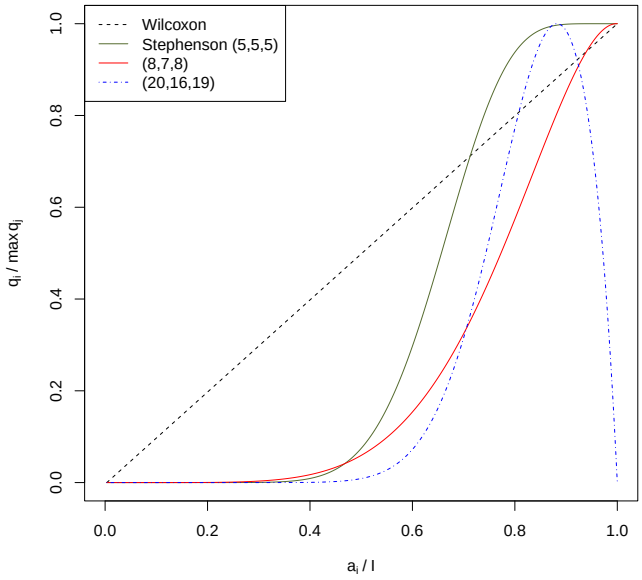
où

$$q_i = \binom{I}{m}^{-1} \sum_{l=\underline{m}}^{\overline{m}} \binom{a_i - 1}{l - 1}\binom{I - a_i}{m - l}$$

On peut remarquer que si $m = \underline{m} = \overline{m} = 2$, $q_i = \binom{a_i-1}{m-1} = \binom{a_i-1}{1} = a_i - 1$, et donc $\widetilde{T}$ correspond à la statistique de Wilcoxon signée du rang à une constante près.

En faisant varier $m, \underline{m}$ et $\overline{m}$, il est possible de contrôler le degré d'influence d'observations de différentes magnitudes. Pour l'illustrer, nous représentons sur la figure 8, les valeurs de $q_i / \max_{1 \leq j \leq I}(q_j)$ en fonction de a_i/I pour différentes valeurs de $m, \underline{m}$ et $\overline{m}$ et $I = 300$. Schématiquement, nous représentons donc pour différentes valeurs $m, \underline{m}$ et $\overline{m}$, la contribution de chaque paire à la statistique considérée en fonction de son rang. La statistique de Stephenson (5,5,5), les U-statistiques (8,7,8) et (20,16,19) accordent relativement plus de poids aux valeurs élevées de $|Y_i|$ et moins de poids aux valeurs faibles de $|Y_i|$ que la statistique de Wilcoxon signée du rang. La statistique (20,16,19) permet, elle, d'accorder un poids plus important aux valeurs élevées de $|Y_i|$, tout en diminuant, contrairement à la statistique (8,7,8) l'influence des valeurs extrêmes qui peuvent être des valeurs aberrantes.

FIGURE 8 – Représentation de q_i, normalisé pour avoir un maximum de 1, en fonction de a_i/I où a_i est le rang de $|Y_i|$



D Estimation d'un effet du traitement additif et constant

Dans cette annexe, nous présentons dans un premier temps l'estimateur de Hodges-Lehmann d'un effet additif constant associé à la statistique de Wilcoxon signée du rang. Parce que

cet estimateur s'appuie sur la distribution de la statistique du test, il n'est pas possible de le calculer en présence d'un biais de composition caché. Dès lors, nous détaillons comment l'encadrement (5) de la distribution de la statistique de test peut être utilisé pour déterminer un intervalle des valeurs possibles de l'estimateur d'Hodges-Lehmann. Cet intervalle des valeurs possibles de l'effet estimé traduit l'incertitude liée à l'existence d'une caractéristique inobservée, et son amplitude dépend de la valeur du paramètre Γ considéré dans l'analyse de sensibilité.

D.1 L'estimateur d'Hodges-Lehmann d'un effet additif constant du traitement

Il existe de nombreux estimateurs d'un effet τ si l'on suppose l'effet du traitement additif et constant, c'est-à-dire si l'on suppose que $\mathbf{R} = \mathbf{r}_C + \tau\mathbf{Z}$. L'estimateur proposé par Hodges et Lehmann (1963); Lehmann (1975) se déduit de la statistique de test retenue pour tester l'hypothèse nulle H_0 d'une absence d'effet du traitement. Comme nous le verrons dans la section D.2 suivante, c'est donc naturellement que l'on pourra utiliser l'encadrement (5) de la distribution de la statistique de test pour déduire un intervalle des valeurs possibles de l'effet, constant et additif, du traitement.

Pour tester l'hypothèse nulle d'un effet additif et constant $H_0 : \boldsymbol{\tau} = \tau_0$, il suffit de calculer la statistique de test ajustée, $t(\mathbf{Z},\mathbf{R} - \tau_0\mathbf{Z})$, puis de tester à partir de celle-ci l'hypothèse nulle d'absence d'effet du traitement. Hodges et Lehmann se proposent de déterminer d'estimer un effet additif et constant du traitement par la valeur τ_0 pour laquelle la statistique de test T ajustée a la distribution « attendue ». Schématiquement [40], l'estimateur de Hodges-Lehmann est la solution $\hat{\tau}$ de l'équation $\bar{\bar{t}} = t(\mathbf{Z},\mathbf{R} - \tau_0\mathbf{Z})$ où $\bar{\bar{t}}$ est l'espérance « attendue » de la statistique de test en l'absence d'effet du traitement, c'est-à-dire après ajustement. Ce calcul est possible car en l'absence de biais de composition caché et d'effet du traitement, l'espérance de la statistique de Wilcoxon signée du rang est connue.

Pour s'en rendre compte, considérons - dans cette section et dans la suivante - et uniquement pour simplifier l'expression des résultats qu'il n'existe pas de valeurs identiques de $V_i Y_i$. En l'absence d'effet du traitement, on « s'attend » donc à sommer la moitié des rangs des I différences, puisqu'il y a une chance sur deux que l'unité traitée ait une valeur supérieure à l'unité de contrôle, en l'absence de biais de composition caché. Comme la somme totale des I rangs est $\frac{I(I+1)}{2}$, il vient que $\bar{\bar{t}} = \frac{I(I+1)}{4}$.

Formellement, Hodges et Lehmann (1963) définissent leur estimateur comme

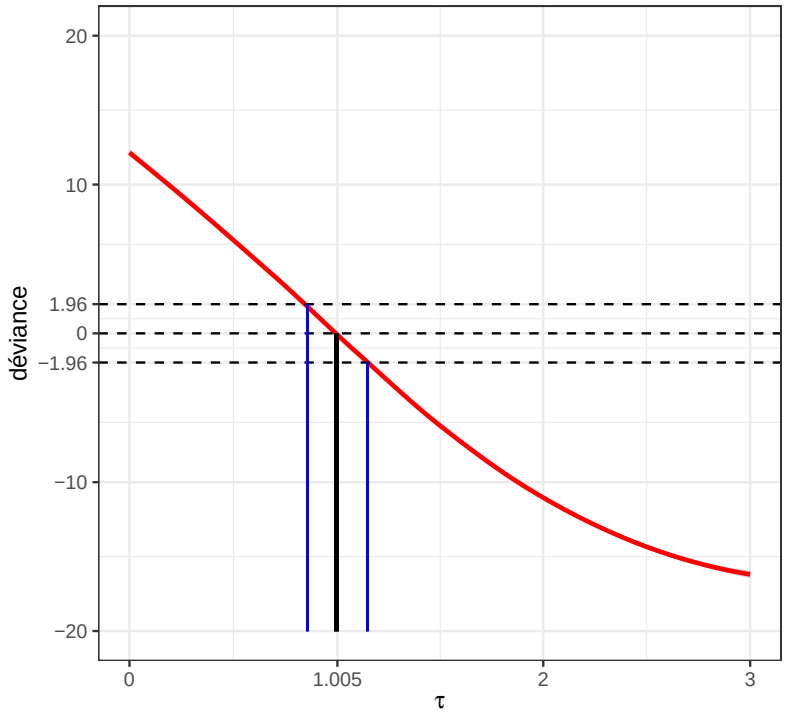
$$\begin{aligned}\hat{\tau} &= \text{SOLUTION}\{\bar{\bar{t}} = t(\mathbf{Z},\mathbf{R} - \hat{\tau}\mathbf{Z})\} \\ &= \frac{\inf\{\tau : \bar{\bar{t}} > t(\mathbf{Z},\mathbf{R} - \hat{\tau}\mathbf{Z})\} + \sup\{\tau : \bar{\bar{t}} < t(\mathbf{Z},\mathbf{R} - \hat{\tau}\mathbf{Z})\}}{2}\end{aligned}$$

afin de tenir compte de la possibilité que la statistique de test ne prenne que des valeurs discrètes [41]. Ainsi, en l'absence de biais de composition caché, il est possible « d'inverser » la statistique de test afin d'estimer un effet additif et constant du traitement. Comme le montre aussi le graphique 9, la démarche peut être étendue pour déterminer un intervalle de confiance à 95 % de l'effet estimé.

40. Pour une explication détaillée, voir (Rosenbaum, 2002b, section 2.7.2)

41. L'estimateur de Hodges-Lehmann hérite des propriétés associés à la statistique de test $t(.,.)$ utilisé, au sens où, par exemple, dès que le test est convergent, l'estimateur de Hodges-Lehmann est convergent (voir Maritz, 1981, pour des explications détaillées). En effet, un test est convergent, comme c'est le cas du test de Wilcoxon signé du rang, si la probabilité de rejeter une hypothèse fausse, quel qu'elle soit, tend vers 1 si la taille de l'échantillon s'accroît. Intuitivement, cette notion est donc reliée à la convergence d'un estimateur dont la probabilité qu'il soit proche de la vraie valeur tend vers 1 si la taille de l'échantillon s'accroît.

FIGURE 9 – Une illustration de l'estimateur d'Hodges-Lehmann de l'effet du traitement et de l'intervalle de confiance correspondant



Note : Pour chaque valeur de τ, effet additif constant du traitement considéré sous H_0, la courbe (en rouge) représente l'écart entre la statistique (centrée et réduite) de test de Wilcoxon signée du rang correspondante avec la valeur nulle (correspondant à l'espérance de la loi $\mathcal{N}(0,1)$). On détermine ainsi la valeur τ qui annule cet écart (segment vertical noir) et les bornes de l'intervalle de confiance à 95 % correspondant (segments verticaux bleus).

D.2 L'analyse de sensibilité de l'estimateur de Hodges-Lehmann

En présence d'un biais de composition caché, la distribution de la statistique de test n'est pas connue et n'est plus égale à $I(I+1)/4$, de telle sorte qu'il n'est plus possible de déterminer l'estimateur de Hodges-Lehmann d'un effet additif et constant. Cependant, en considérant le modèle d'analyse de sensibilité (3), l'espérance de la statistique de test ajusté $\bar{\bar{t}}$ peut être encadrée par les espérances de T^- et de T^+, tels que définis dans la section 3.2, c'est-à-dire par deux grandeurs connues $\bar{\bar{t}}_{\min}$ et $\bar{\bar{t}}_{\max}$. Formellement, si l'on applique les formules de la section 3.2 (et toujours en considérant l'absence de valeurs identiques pour d_s), il vient que :

$$\bar{\bar{t}}_{\min} = \frac{p^- I(I+1)}{2} \quad \text{et} \quad \bar{\bar{t}}_{\max} = \frac{p^+ I(I+1)}{2}$$

avec $p^- = 1/(1+\Gamma)$ et $p^+ = \Gamma/(1+\Gamma)$.

Si l'on calcule $\hat{\tau} = \text{SOLUTION}\{\bar{\bar{t}} = t(\mathbf{Z}, \mathbf{R} - \hat{\tau}\mathbf{Z})\}$ pour toutes les valeurs $\bar{\bar{t}}$ de l'intervalle $[\bar{\bar{t}}_{\min}; \bar{\bar{t}}_{\max}]$, on détermine ainsi l'intervalle des valeurs possibles des estimateurs de Hodges-Lehmann associés à un effet additif et constant. En pratique le minimum et le maximum de $\hat{\tau}$ correspondent aux estimateurs de Hodges-Lehmann associés à $\bar{\bar{t}}_{\min}$ et $\bar{\bar{t}}_{\max}$.

Au final, l'analyse de sensibilité de l'estimateur de Hodges-Lehmann consiste ainsi à déterminer l'intervalle des valeurs possibles de cet estimateur pour différentes grandeurs de Γ.

D.3 L'analyse de sensibilité de l'intervalle de confiance d'un estimateur de Hodges-Lehmann

Comme pour tout estimateur, il est possible, en l'absence de biais de composition caché, de déterminer un intervalle de confiance à l'estimateur de Hodges-Lehmann. De même, en présence d'un biais de composition caché, il est possible de déterminer un *intervalle de confiance de l'intervalle des valeurs possibles* de l'estimateur de Hodges-Lehmann.

Formellement, celui-ci est obtenu à partir de l'approximation normale de la distribution de T_τ^+ et de T_τ^-. Ses bornes correspondent à :

$$\inf\left\{\tau : \frac{T_\tau - \mathrm{E}(T_\tau^+)}{\sqrt{\mathrm{var}(T_\tau^+)}} \leq 1{,}96\right\} \quad \text{et} \quad \sup\left\{\tau : \frac{T_\tau - \mathrm{E}(T_\tau^-)}{\sqrt{\mathrm{var}(T_\tau^-)}} \geq -1{,}96\right\}$$

E Le test *crosscut* : définition et analyse de sensibilité

Dans cette partie, nous présentons sommairement la statistique de test *crosscut* proposée par Rosenbaum (2016). Cette statistique permet de tester l'hypothèse d'effets importants pour de fortes doses du traitement considéré. En cela, il convient d'adapter les notations « d'effet du traitement » et de « réponses potentielles » au cas d'un ensemble discret de doses du traitement.

Notations, doses, réponses potentielles On considère, comme dans la partie 3, I paires, $i = 1, \ldots, I$, comportant chacune deux personnes enquêtées $j = 1{,}2$ telles que $\mathbf{x}_{i1} = \mathbf{x}_{i2}$. Il y a un ensemble $\mathcal{D}$ de doses possibles[42]. Pour toute dose d, l'enquêté j dans la paire i a une réponse potentielle r_{ijd}, de telle sorte que l'effet de la dose d au lieu de la dose d' est $r_{ijd} - r_{ijd'}$. L'hypothèse nulle d'absence d'effet du traitement de Fisher (1935) ("*sharp*

42. $\mathcal{D}$ est un ensemble à valeurs discrètes, l'extension à intervalles de valeurs ne nécessite que des ajustements mineurs à l'exposé présenté (voir Rosenbaum, 2016)

null hypothesis of no treatment effet") correspond alors à $H_0 : r_{ijd} - r_{ijd'} = 0$ pour tout $d, d' \in \mathcal{D}$, pour tout i et j. Plus précisément, si H_0 est vraie, il existe une valeur pour tout i et j, que l'on peut noter r_{0ij}, telle que $r_{ijd} = r_{0ij}$ pour tout $d \in \mathcal{D}$.

On note $\mathbf{R} = (R_{11}, \ldots, R_{2I})^T$ le vecteur de dimension $2I$ associé aux réponses des enquêtés. De même, on note $\mathbf{D} = (D_{11}, \ldots, D_{2I})^T$ le vecteur des doses de chaque enquêté et $\mathbf{u} = (u_{11}, \ldots, u_{2I})$ celui d'une caractéristique inobservée telle que $u_{i1} \neq u_{i2}$ pour certaines paires i. Enfin, si H_0 est vraie, on note $\mathbf{r}_0 = (r_{011}, r_{012}, \ldots, r_{0I2})^T$.

On note $\mathcal{F} = (y_{ijd}, \mathbf{x}_{ij}, u_{ij}, d \in \mathcal{D}, i = 1, \ldots, I, j = 1,2)$ pour les covariables (observées ou non) et les réponses potentielles. Comme dans la partie 3, $\mathcal{F}$ ne précisait pas si l'enquêté répondait par internet ou par papier, i.e. s'il était traité ou non. De même, la définition de $\mathcal{F}$ ne précise pas la dose D_{ij} associée à chaque enquêté. Cependant si H_0 est vraie, alors $\mathbf{R} = \mathbf{r}_0$ et ne dépend pas de $\mathbf{D}$, de telle sorte que $\mathcal{F}$ caractérise $\mathbf{Y}$, puisque chaque enquêté a la même réponse quelle que soit la dose qui lui est associée.

En présence d'un biais de composition caché, la probabilité que l'enquêté j de la paire i reçoive la dose D_{ij}, $\theta_{ijd} = \Pr(D_{ij} = d \mid \mathcal{F})$ n'est pas connue et peut être différente pour les enquêtés de la même paire i, si par exemple $u_{i1} \neq u_{i2}$.

La statistique de *crosscut*

La statistique de *crosscut* s'intéresse au sous-groupe des enquêtés qui présentent des réponses élevées ou faibles (dans notre étude, durées de trajet) et de fortes ou faibles doses de traitement (dans notre étude, durée totale des plages horaires du carnet papier dont le statut - trajet ou non - est à déterminer). Le test associé à cette statistique questionne si, dans ce sous-groupe d'enquêtés, des réponses élevées coïncident avec des doses de traitement élevées. La définition des notions de « réponses/doses élevées » ou « faibles » s'appuie sur une partition des distributions des grandeurs observées $\mathbf{R}$ et $\mathbf{D}$. Dans son article, Rosenbaum discute l'impact du choix de la partition, définie par les quantiles utilisées pour la constituer, sur la puissance de l'analyse de sensibilité associée et présente plusieurs recommandations. Remarquons néanmoins que si la partition est définie par les valeurs médianes de chaque distribution, l'analyse porte sur l'ensemble des unités enquêtées et non sur un sous-groupe.

Pour définir la statistique de *crosscut*, nous considérons les réponses R_{ij} apportées par les enquêtés : si la réponse R_{ij} est élevée ou faible, on pose $h_{ij} = 1$ et $h_{ij} = 0$ sinon. En cela, h_{ij} permet de distinguer les unités enquêtées appartenant au sous-groupe considéré. Par ailleurs, on pose que $y_{ij} = 1$ si la réponse est élevée. et 0 sinon [43].

Du point de vue des doses de traitement, il existe deux sous-ensembles disjoints, $\mathcal{D}_1$,$\mathcal{D}_2$ des doses avec $\mathcal{D}_1 \cup \mathcal{D}_2 \subseteq \mathcal{D}$. Les sous-ensembles $\mathcal{D}_1$ et $\mathcal{D}_2$ sont fixés *a priori* au sens de l'inférence aléatoire. Comme pour les réponses des enquêtés, deux variables permettent de déterminer si les doses sont élevées ou non : $W_{ij} = 1$ renseigne si $D_{ij} \in \mathcal{D}_1 \cup \mathcal{D}_2$ et $Z_{ij} = 1$ si $D_{ij} \in \mathcal{D}_1$.

43. Par construction, $h_{ij} = 0 \Rightarrow y_{ij} = 0$.

Tableau 13 – Notation pour une paire i

	Réponses élevées ou faibles		
	$y_{ij}=1$	$y_{ij}=0$	Total
Doses élevées ou faibles			
$Z_{ij}=1$	$T_i=\sum_{j=1,2} Z_{ij}y_{ij}$		$n_i=\sum_{j=1,2} Z_{ij}h_{ij}$
$Z_{ij}=0$			N_i-n_i
Total	$m_i=\sum_{j=1,2} W_{ij}y_{ij}$	N_i-m_i	$N_i=\sum_{j=1,2} W_{ij}h_{ij}$

Note : les N_i personnes ont $h_{ij}=1$ et $W_{ij}=1$, i.e. qu'elles présentent des doses de traitement élevées ou faibles et des réponses élevées ou faibles.

La statistique de *crosscut* correspond à $T=\sum_{i=1}^{I} T_i$ où $T_i=\sum_{j=1,2} Z_{ij}r_{ij}$. Plus précisément, comme l'illustre le tableau 13, au sein d'une paire i, T_i correspond au nombre d'enquêtés avec une réponse élevée et une dose de traitement élevée, m_i au nombre d'enquêtés avec une réponse élevée ($y_{ij}=1$) et ayant une dose élevée ou faible ($D_{ij}\in\mathcal{D}_1\cup\mathcal{D}_2$), n_i au nombre d'enquêtés avec une dose élevée $D_{ij}\in\mathcal{D}_1$) et ayant une réponse élevée ou faible ($h_{ij}=1$) et N_i au nombre d'enquêtés ayant une réponse élevée ou faible ($h_{ij}=1$) et une dose élevée ou faible ($D_{ij}\in\mathcal{D}_1\cup\mathcal{D}_2$).

F Facteurs d'évidence

Dans cette partie, nous justifions succinctement l'approche par facteurs d'évidence mis en œuvre dans la section 4.2. Pour illustrer notre propos, nous nous appuierons sur un exemple fictif constitué de deux paires dont les unités enquêtées, numérotées de 1 à 4, sont présentées dans le tableau 14 qui précisent pour chacune le mode de collecte et la durée totale à coder déclarée. Dans ce qui suit, nous désignerons par « traitement » la collecte par carnet papier et par « dose », i.e. la durée totale des plages horaires à coder.

Tableau 14 – Quatre positions vis-à-vis du mode de collecte dans deux paires

Structure		Position de traitement		Durée de trajet
Identifiant	Paire	Collecte papier	Durée à coder	
1	1	1	150	240
2	1	0	0	130
3	2	1	30	70
4	2	0	0	40

Il existe deux traitements différents et deux doses de traitement différentes soit, pour reprendre les termes de Rosenbaum (2017b), 4 « positions de traitements » : être le répondant avec le carnet papier qui déclare 150 minutes « d'autres trajets », être le répondant avec le carnet numérique associé à ce répondant, être le répondant avec le carnet papier qui déclare 30 minutes « d'autres trajets » et être le répondant avec le carnet numérique associé à celui-ci.

Structure générale

Considérons l'expérience qui affecterait chacun des 4 enquêtés à l'une de ces positions de traitement tout en conservant la composition des paires intacte car elles ont été constituées de telle sorte que le répondant avec le carnet papier et celui par internet soient similaires

sur les caractéristiques sociodémographiques observées.

Formellement, il est utile de représenter une assignation à une position de traitement par une matrice de permutation, qui affecte aux 4 personnes enquêtées appariées ($(1,2,3,4)^T$) dans $n = 2$ paires une position de traitement.

$$\boldsymbol{gn} = \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{bmatrix} \cdot \begin{bmatrix} 1 \\ 2 \\ 3 \\ 4 \end{bmatrix} = \begin{bmatrix} 3 \\ 4 \\ 2 \\ 1 \end{bmatrix}$$

Dans l'exemple ci-dessus, l'assignation au traitement $\boldsymbol{g}$ affecte le premier enquêté au mode de collecte papier avec 30 min. de durée à coder, tout en conservant au sein de la même paire l'autre enquêté auquel il est apparié et en lui allouant le mode de collecte par internet. L'ensemble des assignations possibles dans l'expérience que nous considérons est un sous-groupe $\mathcal{G}$ fini des $(2n)!$ matrices de permutations possibles. En effet, il existe 2 façons d'affecter les paires à une durée à coder, 150 min. à une paire et 30 min. à l'autre. Par ailleurs, il existe 2 façons au sein des 2 paires d'allouer le mode de collecte papier à l'un des enquêtés, soit 2^2. Ainsi, il n'existe pas $4! = 24$ assignations possibles du traitement, car on souhaite conserver la constitution des paires mais $2 \times 2^2 = 8$ assignations possibles. En notant $|\mathcal{S}|$ le nombre d'éléments d'un ensemble fini $|\mathcal{S}|$, il existe $|\mathcal{G}| = (n)! \times 2^n$ assignations possibles avec n paires.

Une distribution de probabilité sur $\mathcal{G}$ affecte à toute matrice de permutation $\boldsymbol{g} \in \mathcal{G}$ une probabilité $p(\boldsymbol{g})$. Plus précisément, une distribution de probabilité p sur $\mathcal{G}$ vérifie (i) $p(\boldsymbol{g}) \geq 0$ pour tout $\boldsymbol{g} \in \mathcal{G}$ et (ii) $1 = \sum_{\boldsymbol{g} \in \mathcal{G}} p(\boldsymbol{g})$ et une assignation $\boldsymbol{G}$ à des positions de traitement issue de cette distribution a $\Pr(\boldsymbol{G} = \boldsymbol{g}) = p(\boldsymbol{g})$. Dans le cas d'une expérience aléatoire, $\boldsymbol{g}$ est choisie au hasard parmi $\mathcal{G}$; ainsi $p(\boldsymbol{g}) = |\mathcal{G}|^{-1}$ pour tout $\boldsymbol{g} \in \mathcal{G}$ et $p = (p_{\boldsymbol{g}_1}, p_{\boldsymbol{g}_2}, \ldots, p_{\boldsymbol{g}_{|\mathcal{G}|}}) = (\frac{1}{|\mathcal{G}|}, \ldots, \frac{1}{|\mathcal{G}|})^T$ [44]. Si les différentes assignations possibles modifient la position de traitement allouée à chaque unité, il est important de se rappeler que sous l'hypothèse H_0 d'une absence d'effet du mode de collecte (*Fisher's sharp null hypothesis*), les durées de trajet sont fixes ; ainsi les différentes permutations ne changent pas les durées de trajet déclarées. L'hypothèse H_0 pose que la personne 1 déclarerait toujours 240 minutes de trajet sous les 8 positions de traitement possibles ; il en est de même pour les 3 autres personnes de notre exemple.

Comme le suggère l'exemple précédent, cette assignation des unités à une position de traitement peut être reliée aux deux analyses avec inférence aléatoire menées dans la section 4.2 en considérant deux sous-groupes de permutations.

Le premier sous-groupe de permutations, $\mathcal{K}$, de $\mathcal{G}$ caractérise l'analyse où l'on ignore les doses de traitement en ne s'intéressant qu'à la comparaison des durées de trajet du répondant avec le carnet numérique et le carnet papier. En cela, l'expérience aléatoire associée assigne des positions de traitement en conservant les doses de traitement allouées à chaque paire et en modifiant uniquement le choix de l'enquêté qui répond avec le carnet papier [45]. Dans notre exemple, il existe 4 permutations possibles, et si nous considérons n paires, $\mathcal{K}$ a $|\mathcal{K}| = 2^n$ éléments. De fait, la statistique de Wilcoxon signée du rang utilisée dans cette première analyse est invariante aux $n!$ permutations de toutes les paires et ne considèrent que les 2^n permutations du mode de collecte au sein des paires.

44. Comme nous le verrons par la suite, dans le cas d'une expérience non aléatoire, on considérera un *ensemble* $\mathcal{P}$ de distributions de probabilité p, cet ensemble reflétant notre incertitude, calibrée dans le cadre d'une analyse de sensibilité, sur la vraie distribution de probabilité.

45. On rappelle que sous H_0 les durées de trajet déclarées sont fixes au sens de Fisher.

Il existe un deuxième sous-groupe de permutations, $\mathcal{H}$, de $\mathcal{G}$ qui correspond à l'inférence aléatoire réalisée dans la deuxième analyse. Celui-ci permute uniquement les « doses de traitement » entre les paires tout en conservant au sein de chaque paire l'enquêté qui répond avec le carnet papier. Dans notre exemple, $\mathcal{H}$ n'a que 2 éléments et avec n paires, $|\mathcal{H}| = n!$ éléments. La statistique de *crosscut* utilisée dans la deuxième analyse est en effet conditionnée par les modes de collecte affectés au sein de chaque paire au sens où est fixée l'une des 2^n permutations possibles du mode de collecte au sein de chaque paire et utilise les $n!$ permutations de doses de traitement entre les paires.

Comme le suggère cette présentation des deux sous-groupes, il est possible de montrer que tout assignation $\boldsymbol{g} \in \mathcal{G}$ peut s'écrire comme le produit d'une assignation issue de $\mathcal{K}$ et d'une assignation issue de $\mathcal{G}$ et surtout que cette représentation est *unique* : $\boldsymbol{g} = \boldsymbol{hk}$ où $\boldsymbol{h} \in \mathcal{H}$ et $\boldsymbol{k} \in \mathcal{K}$. La loi de probabilité de $\boldsymbol{g}$ correspond ainsi à la loi de probabilité *jointe* de $\boldsymbol{g} = \boldsymbol{hk}$.

Analyse de sensibilité pour le test d'absence d'effet de mesure
Dans notre étude, le mode de collecte ou les « doses de traitement » ne sont pas affectés aléatoirement aux *répondants*. Le modèle d'analyse de sensibilité conduit à considérer non plus *une* distribution de probabilité mais un ensemble de distributions de probabilités possibles. Ainsi pour la première analyse, avec une valeur Γ de paramètre de sensibilité, on considère un ensemble $\mathcal{P}_\Gamma$ dont les éléments sont des distributions de probabilité $\boldsymbol{p} \in \mathcal{K}$. De même, on considérera $\mathcal{P}'_{\Gamma'}$ l'ensemble de distributions de probabilité $p' \in \mathcal{H}$ pour une valeur Γ' du paramètre de sensibilité.

Combiner deux analyses de sensibilité
Le point de départ consiste donc à considérer deux facteurs d'évidence tels que (i) chaque assignation à une position de traitement $\boldsymbol{g} \in \mathcal{G} = \mathcal{HK}$ a une unique représentation $\boldsymbol{g} = \boldsymbol{hk}$ avec $\boldsymbol{h} \in \mathcal{H}$ et $\boldsymbol{k} \in \mathcal{K}$, où $\mathcal{H}$ est un sous-groupe de $\mathcal{G}$ et (ii) la statistique de test du premier facteur $t(\boldsymbol{g}) = t(\boldsymbol{hk})$ est $\mathcal{H}$-invariant, soit $t(\boldsymbol{hk}) = t(\boldsymbol{k})$.

Dans notre cas, les deux facteurs d'évidence sont les deux analyses menées séparément. Par ailleurs la statistique de Wilcoxon signée du rang prend la même valeur quelle que soit les doses de traitement, puisque cette statistique n'est pas fonction des doses de traitement. Aucune autre hypothèse n'est nécessaire, notamment sur l'indépendance entre $\boldsymbol{H}$ et $\boldsymbol{K}$.

Il s'agit ensuite de tester, deux fois, l'hypothèse nulle H_0 d'absence d'effet, une fois avec un test marginal en utilisant $t(\boldsymbol{k})$ (ici la statistique de Wilcoxon signée du rang) et une fois avec un test conditionnel sachant $\boldsymbol{K} = \boldsymbol{k}$ utilisant $t'(\boldsymbol{hk})$ (ici la statistique de *crosscut*), afin d'obtenir une paire de bornes supérieures de p-values $(\overline{P}_\Gamma, \overline{P}'_{\Gamma'})$ des vraies p-values, (P, P'). Formellement, on suppose $\boldsymbol{p} \in \mathcal{P}_\Gamma$ et l'on teste H_0 avec la statistique $t(\boldsymbol{g}) = t(\boldsymbol{hk})$ qui est invariant sur $\mathcal{H}$ au sens où $t(\boldsymbol{hk}) = t(\boldsymbol{k})$ pour tout $\boldsymbol{h}$ et $\boldsymbol{k}$ afin d'obtenir la p-value maximale $\overline{P}_\Gamma$. On teste à nouveau H_0 à partir des distributions conditionnelles $p' \in \mathcal{P}'_{\Gamma'}$ de $t'(\boldsymbol{g}) = t'(\boldsymbol{hk})$ de $\boldsymbol{H}$ sachant $\boldsymbol{K} = \boldsymbol{k}$ pour obtenir la p-value maximale $\overline{P}'_{\Gamma'}$ des p-values conditionnelles.

Pour combiner les deux facteurs d'évidence, i.e les deux analyses de sensibilité ainsi menées, il est nécessaire de pouvoir borner celle issue de la distribution jointe. Pour cela Rosenbaum (2017b) établit la proposition suivante sur les p-values issues de la distribution jointe.

Proposition. *Si l'hypothèse H_0 est vraie, si $\Pr(\boldsymbol{K} = \boldsymbol{k})$ est une des distributions de probabilité $\boldsymbol{p} \in \mathcal{P}_\Gamma$, si $\Pr(\boldsymbol{H} = \boldsymbol{h} | \boldsymbol{K} = \boldsymbol{k})$ est une des distributions de probabilité $\boldsymbol{p}' \in \mathcal{P}'_{\Gamma'}$, alors pour toute distribution jointe du produit knit, la paire des bornes supérieures de p-values, $(\bar{P}_\Gamma, \bar{P}'_{\Gamma'})$ est stochastiquement plus grande que la distribution uniforme sur le carré $[0,1] \times [0,1]$, ce qui implique que $\Pr(\bar{P}_\Gamma \leq \alpha, \bar{P}'_{\Gamma'} \leq \alpha') \leq \alpha\alpha'$ pour tout $0 \leq \alpha \leq 1$ et $0 \leq \alpha' \leq 1$*

Cette proposition implique que du point de vue pratique il est possible de combiner ces deux analyses ($\overline{P}_{\Gamma}$,$\overline{P}'_{\Gamma'}$) comme s'il s'agissait de deux p-values indépendantes issues de deux analyses non corrélées.

G Une illustration des carnets papiers et numériques

FIGURE 10 – Une illustration du carnet papier

Exemple de remplissage de ce carnet

➤ **Notez vos occupations de manière détaillée :**

Tâches domestiques. Subdivisez en lessive, vaisselle, raccommodage, etc.

Lecture (sauf études). Précisez ce que vous lisez (journal, roman, etc.).

Trajets. Distinguez les trajets des autres activités.

Travail. Inutile de détailler vos activités durant le travail, inscrivez simplement « je travaille ». N'oubliez pas de noter les pauses entre les périodes de travail.

		Décrivez vos différentes occupations de la journée : Indiquez les heures de début et de fin (plage horaire de l'activité) grâce à une accolade. Décrivez votre occupation.	Faites-vous autre chose en même temps ? (lecture, conversation, radio, TV...)	Avez-vous utilisé un ordinateur ou un smartphone pendant votre activité ? (ou tout autre objet connecté à internet)
18 h	10	Je rentre du travail avec un collègue	Conversation	☐
	20			☐
	30			☐
	40	Je me repose sur le canapé	J'écoute de la musique	☒
	50	Je poste des messages sur mon Facebook		☒
19 h		Je pars pour le supermarché	Radio	☐
	10	Je fais des courses pour dîner		☐
	20			☐
	30	Trajet de retour du supermarché	Radio	☐
	40			☐
	50	Je surveille ma nièce qui fait ses devoirs		☐
20 h		Je prépare le dîner	Je garde ma nièce	☐
	10			☐
	20			☐
	30	Je mange avec ma femme, mon frère et ma nièce	Conversation	☐
	40			☐
	50	Je range la cuisine	J'écoute la radio	☐
		Je regarde les actualités sur mon téléphone	Je grignote des biscuits	☒

6

(a) Saisie des activités

Lieu ou Moyen de transport		En présence de qui (plusieurs réponses possibles)						Votre activité est dans un but (une seule réponse possible)			
		Personnes du ménage					Autres personnes que vous connaissez				
Cochez cette case s'il s'agit d'un trajet entre votre domicile et votre travail		Seul (1)	Votre conjoint (2)	Votre père, votre mère (3)	Enfant(s) du ménage (4)	Autres personnes du ménage (5)	(6)	Personnel ou pour mon ménage (1)	Profes-sionnel (2)	Aide à un autre ménage (3)	Bénévole, pour une association (4)
	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
En voiture	☒	☒	☐	☐	☐	☐	☒	☐	☒	☐	☐
	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
Chez moi	☐	☐	☐	☐	☐	☐	☐	☒	☐	☐	☐
"	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
À pied	☐	☒	☐	☐	☐	☐	☐	☒	☐	☐	☐
Au supermarché	☐	☒	☐	☐	☐	☐	☐	☒	☐	☐	☐
	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
À pied	☐	☒	☐	☐	☐	☐	☐	☒	☐	☐	☐
	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
Chez moi	☐	☐	☐	☐	☐	☐	☒	☐	☐	☒	☐
"	☐	☐	☐	☐	☐	☐	☒	☒	☐	☐	☐
	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
"	☐	☐	☒	☐	☐	☐	☒	☒	☐	☐	☐
	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
"	☐	☐	☒	☐	☐	☐	☐	☒	☐	☐	☐
"	☐	☒	☐	☐	☐	☐	☐	☒	☐	☐	☐

7

(b) Saisie du lieu/moyen de transport

FIGURE 11 – Une illustration de l'affichage du sélecteur du carnet numérique

Série des Documents de Travail « Méthodologie Statistique »

9601 : Une méthode synthétique, robuste et efficace pour réaliser des estimations locales de population.
G. DECAUDIN, J.-C. LABAT

9602 : Estimation de la précision d'un solde dans les enquêtes de conjoncture auprès des entreprises.
N. CARON, P. RAVALET, O. SAUTORY

9603 : La procédure FREQ de SAS - Tests d'indépendance et mesures d'association dans un tableau de contingence.
J. CONFAIS, Y. GRELET, M. LE GUEN

9604 : Les principales techniques de correction de la non-réponse et les modèles associés.
N. CARON

9605 : L'estimation du taux d'évolution des dépenses d'équipement dans l'enquête de conjoncture : analyse et voies d'amélioration.
P. RAVALET

9606 : L'économétrie et l'étude des comportements. Présentation et mise en œuvre de modèles de régression qualitatifs. Les modèles univariés à résidus logistiques ou normaux (LOGIT, PROBIT).
S. LOLLIVIER, M. MARPSAT, D. VERGER

9607 : Enquêtes régionales sur les déplacements des ménages : l'expérience de Rhône-Alpes.
N. CARON, D. LE BLANC

9701 : Une bonne petite enquête vaut-elle mieux qu'un mauvais recensement ?
J.-C. DEVILLE

9702 : Modèles univariés et modèles de durée sur données individuelles.
S. LOLLIVIER

9703 : Comparaison de deux estimateurs par le ratio stratifiés et application aux enquêtes auprès des entreprises.
N. CARON, J.-C. DEVILLE

9704 : La faisabilité d'une enquête auprès des ménages.
1. au mois d'août.
2. à un rythme hebdomadaire

C. LAGARENNE, C. THIESSET

9705 : Méthodologie de l'enquête sur les déplacements dans l'agglomération toulousaine.
P. GIRARD.

9801 : Les logiciels de désaisonnalisation TRAMO & SEATS : philosophie, principes et mise en œuvre sous SAS.
K. ATTAL-TOUBERT, D. LADIRAY

9802 : Estimation de variance pour des statistiques complexes : technique des résidus et de linéarisation.
J.-C. DEVILLE

9803 : Pour essayer d'en finir avec l'individu Kish.
J.-C. DEVILLE

9804 : Une nouvelle (encore une !) méthode de tirage à probabilités inégales.
J.-C. DEVILLE

9805 : Variance et estimation de variance en cas d'erreurs de mesure non corrélées ou de l'intrusion d'un individu Kish.
J.-C. DEVILLE

9806 : Estimation de précision de données issues d'enquêtes : document méthodologique sur le logiciel POULPE.
N. CARON, J.-C. DEVILLE, O. SAUTORY

9807 : Estimation de données régionales à l'aide de techniques d'analyse multidimentionnelle.
K. ATTAL-TOUBERT, O. SAUTORY

9808 : Matrices de mobilité et calcul de la précision associée.
N. CARON, C. CHAMBAZ

9809 : Échantillonnage et stratification : une étude empirique des gains de précision.
J. LE GUENNEC

9810 : Le Kish : les problèmes de réalisation du tirage et de son extrapolation.
C. BERTHIER, N. CARON, B. NEROS

9901 : Perte de précision liée au tirage d'un ou plusieurs individus Kish.
N. CARON

9902 : Estimation de variance en présence de données imputées : un exemple à partir de l'enquête Panel Européen.
N. CARON

0001 : L'économétrie et l'étude des comportements. Présentation et mise en œuvre de modèles de régression qualitatifs. Les modèles univariés à résidus logistiques ou normaux (LOGIT, PROBIT) (version actualisée).
S. LOLLIVIER, M. MARPSAT, D. VERGER

0002 : Modèles structurels et variables explicatives endogènes.
J.-M. ROBIN

0003 : L'enquête 1997-1998 sur le devenir des personnes sorties du RMI - Une présentation de son déroulement.
D. ENEAU, D. GUILLEMOT

0004 : Plus d'amis, plus proches ? Essai de comparaison de deux enquêtes peu comparables.
O. GODECHOT

0005 : Estimation dans les enquêtes répétées : application à l'Enquête Emploi en Continu.
N. CARON, P. RAVALET

0006 : Non-parametric approach to the cost-of-living index.
F. MAGNIEN, J. POUGNARD

0101 : Diverses macros SAS : Analyse exploratoire des données, Analyse des séries temporelles.
D. LADIRAY

0102 : Économétrie linéaire des panels : une introduction.
T. MAGNAC

0201 : Application des méthodes de calages à l'enquête EAE-Commerce.
N. CARON

C 0201 : Comportement face au risque et à l'avenir et accumulation patrimoniale - Bilan d'une expérimentation.
L. ARRONDEL, A. MASSON, D. VERGER

C 0202 : Enquête Méthodologique Information et Vie Quotidienne - Tome 1 : bilan du test 1, novembre 2002.
J.-A. VALLET, G. BONNET, J.-C. EMIN, J. LEVASSEUR, T. ROCHER, P. VRIGNAUD, X. D'HAULTFOEUILLE, F. MURAT, D. VERGER, P. ZAMORA

0203 : General principles for data editing in business surveys and how to optimise it.
P. RIVIERE

0301 : Les modèles logit polytomiques non ordonnés : théories et applications.
C. AFSA ESSAFI

0401 : Enquête sur le patrimoine des ménages - Synthèse des entretiens monographiques.
V. COHEN, C. DEMMER

0402 : La macro SAS CUBE d'échantillonnage équilibré
S. ROUSSEAU, F. TARDIEU

0501 : Correction de la non-réponse et calage de l'enquêtes Santé 2002
N. CARON, S. ROUSSEAU

0502 : Correction de la non-réponse par répondération et par imputation
N. CARON

0503 : Introduction à la pratique des indices statistiques - notes de cours
J-P BERTHIER

0601 : La difficile mesure des pratiques dans le domaine du sport et de la culture - bilan d'une opération méthodologique
C. LANDRE, D. VERGER

0801 : Rapport du groupe de réflexion sur la qualité des enquêtes auprès des ménages
D. VERGER

M2013/01 : La régression quantile en pratique
P. GIVORD,
X. D'HAULTFOEUILLE

M2014/01 : La microsimulation dynamique : principes généraux et exemples en langage R
D. BLANCHET

M2015/01 : la collecte multimode et le paradigme de l'erreur d'enquête totale
T. RAZAFINDROVONA

M2015/02 : Les méthodes de Pseudo-Panel
M. GUILLERM

M2015/03 : Les méthodes d'estimation de la précision pour les enquêtes ménages de l'Insee tirées dans Octopusse
E. GROS
K. MOUSSALAM

M2016/01 : Le modèle Logit Théorie et application.
C. AFSA

M2016/02 : Les méthodes d'estimation de la précision de l'Enquête Emploi en Continu
E. GROS
K.MOUSSALAM

M2016/03 : Exploitation de l'enquête expérimentale Vols, violence et sécurité.
T. RAZAFINDROVONA

M2016/04 : Savoir compter, savoir coder. Bonnes pratiques du statisticien en programmation.
E. L'HOUR
R. LE SAOUT
B. ROUPPERT

M2016/05 : Les modèles multiniveaux
P. GIVORD
M. GUILLERM

M2016/06 : Econométrie spatiale : une introduction pratique
P. GIVORD
R. LE SAOUT

M2016/07 : La gestion de la confidentialité pour les données individuelles
M. BERGEAT

M2016/08 : Exploitation de l'enquête expérimentale Logement internet-papier
T. RAZAFINDROVONA

M2017/01: Exploitation de l'enquête expérimentale Qualité de vie au travail
T. RAZAFINDROVONA

M2018/01: Estimation avec le score de propension sous R
S. QUANTIN

M2018/02: Modèles semi-paramétriques de survie en temps continu sous R
S. QUANTIN

M2019/01 : Les méthodes de décomposition appliquées à l'analyse des inégalités
B. BOUTCHENIK
E. COUDIN
S. MAILLARD

M2020/01 : L'économétrie en grande dimension
J. L'HOUR

M2021/01 : R Tools for JDemetra+ - Seasonal adjustment made easier
A. SMYK
A. TCHANG

M2021/02 : Le traitement du biais de sélection endogène dans les enquêtes auprès des ménages par modèle de Heckman
L. CASTELL
P. SILLARD

M2021/03 : Conception de questionnaires auto-administrés
H. KOUMARIANOS
A. SCHREIBER

M2022/01 : Introduction à la géomatique pour le statisticien : quelques concepts et outils innovants de gestion, traitement et diffusion de l'information spatiale
F. SEMECURBE
E. COUDIN

M2022/02 : Le zonage en unites urbaines 2020
V. COSTEMALLE
S. OUJIA
C. GUILLO
A. CHAUVET

M2023/01 : Les réseaux de neurones appliqués à la statistique publique : méthodes et cas d'usages
D. BABET
Q. DELTOUR
T. FARIA
S. HIMPENS

M2023/02 : Redressements de la première vague de l'enquête epicov : un exemple de correction des effets de sélection dans les enquêtes multimodes
L.CASTELL
C. FAVRE-MARTINOZ
N. PALIOD
P. SILLARD

M2023/03 : Appariements de données individuelles : concepts, méthodes, conseils
L.MALHERBE

M2023/04 : Victimations déclarées et effets de mode : enseignements de l'expérimentation panel multimode de l'enquête cadre de vie et sécurité
L. CASTELL
M. CLERC
D. CROZE
S. LEGLEYE
A. NOUGARET

M2024/01 : Estimation en temps réel de la tendance-cycle : apport de l'utilisation des filtres asymétriques dans la détection des points de retournement
A.QUARTIER-LA-TENTE

M2024/02 : La disponibilité des coordonnées de contact dans fidéli-nautile - quels enseignements pour les protocoles de collecte ?
G. CHARRANCE (INED)

M2024/03 : Discuter l'existence d'un effet de sélection dans un cadre multimode grâce à une analyse de sensibilité - Application aux enquêtes annuelles de recensement
L. COURT
S. QUANTIN

M2024/04 : Vers une désaisonnalisation des séries temporelles infra-mensuelles avec JDemetra+
A. SMYK
K. WEBEL

M2025/01 : Les estimations par capture-recapture ou par système multiple : quelques éléments théoriques
P. ARDILLY
H. KOUMARIANOS

M2025/02 : Tests cognitifs pour les enquêtes auto-administrées : quelques éléments de méthode
D. GUILLEMOT
J. DIRAND
C. FLUXA

M2025/03 : Statistiques fondées sur des données administratives - esquisse d'un cadre général
H. KOUMARIANOS
P. RIVIÈRE

M2025/04 : Peut-on estimer un effet de mesure sur une enquête à partir d'un essai croisé ab/ba : la question de la non-réponse non ignorable dans l'enquête test emploi du temps
Loreline COURT
Simon QUANTIN