Composante du produit nº 11-522-X au catalogue de Statistique Canada
La série des symposiums internationaux de Statistique Canada : recueil

Article

Symposium 2008 :
Collecte des données : défis, réalisations et nouvelles orientations

Exactitude des échantillons de données sur les voyages : utiliser les méthodes en ligne ou les envois postaux?

par Nandini Nadkarni, Ph.D. et George Harmon

2009

Statistique Canada Statistics Canada

Canada

Recueil du Symposium 2008 de Statistique Canada
Collecte des données : défis, réalisations et nouvelles orientations

Exactitude des échantillons de données sur les voyages : utiliser les méthodes en ligne ou les envois postaux?

Nandini Nadkarni, Ph.D. et George Harmon[1]

Résumé

La collecte des données en ligne a commencé en 1995. Il s'agissait alors d'une solution de rechange pour mener certains types de recherche auprès des consommateurs, mais elle a pris de l'ampleur en 2008. Cette croissance a surtout été observée dans les études qui utilisent des méthodes d'échantillonnage non probabiliste. Bien que l'échantillonnage en ligne ait acquis de la crédibilité pour ce qui est de certaines applications de recherche, de sérieuses questions demeurent concernant le bien-fondé des échantillons prélevés en ligne dans le cas des recherches exigeant des mesures volumétriques précises du comportement de la population des États-Unis, notamment en ce qui a trait aux voyages. Dans le présent exposé, nous passons en revue la documentation et comparons les résultats d'études fondées sur des échantillons probabilistes et des échantillons prélevés en ligne pour comprendre les différences entre ces deux méthodes d'échantillonnage. Nous montrons aussi que les échantillons prélevés en ligne sous-estiment d'importants types de voyages, même après pondération en fonction de données démographiques et géographiques.

Mots clés : Collecte de données en ligne, études sur les voyages, exactitude, uniformité.

1. Introduction

1.1 Contexte

Au cours des dernières années, bon nombre d'organismes de recherche ont envisagé d'effectuer en ligne leurs études de suivi et leurs propres recherches afin d'accélérer les temps de réponse et de réduire les coûts du travail sur le terrain ou les coûts opérationnels. Si la rapidité et la rentabilité constituent des objectifs louables, elles sont cependant secondaires par rapport à l'objectif primordial d'obtenir des mesures valides et représentatives. Pour les dirigeants d'organismes de voyage privés et publics qui ont des ressources à gérer et des comptes à rendre à leurs actionnaires ou au public, la principale exigence consiste à établir des rapports exacts et crédibles sur les habitudes de voyage et sur les visiteurs.

D'autres questions se posent en ce qui concerne les différences entre les méthodes d'échantillonnage traditionnelles et les panels en ligne. L'échantillonnage probabiliste traditionnel consiste à prélever des échantillons représentatifs de la population. On constitue ces échantillons de manière ascendante en les prélevant dans des sections géographiques et en équilibrant le panel, à chaque niveau, selon l'âge, le sexe, le revenu et d'autres caractéristiques connues. Les échantillons prélevés en ligne sont différents puisqu'ils sont constitués de volontaires; ces derniers ne sont pas recrutés dans des régions géographiques, mais plutôt sur le Web. Après avoir dénombré les volontaires, le gestionnaire de panel en ligne les répartit par bloc géographique et tente de constituer un échantillon représentatif. Il est difficile d'établir et de maintenir l'équilibre démographique et proportionnel des échantillons de volontaires à l'intérieur des niveaux géographiques.

Un échantillon prélevé en ligne peut être représentatif d'une population en ligne, mais cette dernière n'est pas nécessairement représentative de la population totale des États-Unis. D'après l'enquête sur l'accès à Internet menée en décembre 2008 par le Pew Internet & American Life Project, 73 % des adultes américains utilisent Internet ou le courrier électronique (Madden et Jones, 2008). Une autre enquête du même organisme révèle que 55 % des adultes américains disposent chez eux d'une liaison Internet à large bande, en hausse par rapport à l'année précédente alors que 47 % disposaient chez eux de l'accès à grande vitesse. Les deux groupes chez lesquels on n'observe aucune

[1] D.K. Shifflet and Associates Ltd., 1750 Old Meadow Rd, Suite 620, McLean, VA 22102 (nnadkarni@dksa.com).

variation de la croissance sont les Afro-Américains et les ménages à faible revenu. Les non-utilisateurs d'Internet représentent un pourcentage élevé d'utilisateurs éventuels du service à large bande, mais bon nombre d'entre eux ne sont pas intéressés à communiquer en ligne. Environ le quart (27 %) des adultes américains n'utilisent pas Internet; ils sont habituellement âgés (l'âge médian étant de 61 ans) et disposent d'un revenu inférieur à celui des utilisateurs en ligne. Toutes proportions gardées, ils sont plus de deux fois plus nombreux que les utilisateurs d'Internet à faire partie d'un ménage à faible revenu (Horrigan, 2008).

Pour souligner les difficultés liées à l'échantillonnage en ligne, les principales associations professionnelles de recherche ont émis de rigoureuses lignes directrices à l'intention des chercheurs qui utilisent des échantillons non probabilistes composés de volontaires en ligne, et leur déconseillent même de tenter d'estimer l'erreur d'échantillonnage liée à ces études. La American Association for Public Opinion Research (AAPOR) est le principal organisme professionnel de recherche par enquêtes et sondages de l'opinion publique aux États-Unis; ses membres sont issus du milieu universitaire, des médias, de l'administration publique, du secteur sans but lucratif et du secteur privé.

Selon les directives de l'AAPOR, les chercheurs et les auteurs de rapports de recherche qui effectuent des enquêtes et des sondages à participation volontaire auprès d'échantillons prélevés en ligne sont tenus de divulguer que les répondants n'ont pas été sélectionnés au hasard parmi la population totale, mais plutôt parmi les personnes ayant pris l'initiative ou ayant accepté de répondre volontairement. L'AAPOR recommande en outre à ses membres d'utiliser, dans les enquêtes en ligne et autres enquêtes menées auprès de personnes autosélectionnées, la formulation suivante : « Les répondants à cette enquête ont été sélectionnés parmi les personnes qui [se sont portées volontaires/se sont inscrites pour participer aux enquêtes et aux sondages en ligne de (nom de l'entreprise)]. Les données (ont été/n'ont pas été) pondérées en fonction de la composition démographique de (la population cible). L'échantillon étant fondé sur les personnes qui ont décidé volontairement de participer [au panel], contrairement à un échantillon probabiliste, il n'est pas possible de calculer une estimation de l'erreur d'échantillonnage. Toutes les enquêtes et tous les sondages sur échantillon peuvent faire l'objet de plusieurs sources d'erreur, dont l'erreur d'échantillonnage, l'erreur de couverture et l'erreur de mesure » (American Association for Public Opinion Research).

L'Association de la recherche et de l'intelligence marketing (organisme né de la fusion de l'Association canadienne des organisations de recherche en marketing (ACORM), du Conseil canadien de la recherche par sondage (CCRS) et de l'Association professionnelle de recherche en marketing (APRM)) fait également état des limites de l'échantillonnage non probabiliste. Dans la section de son Code de déontologie (2007) consacrée à l'intégrité de l'information, l'ARIM stipule que les chercheurs doivent « éviter de faire des déclarations sur les marges d'erreur d'échantillonnage par rapport aux estimations démographiques lorsqu'on n'a pas utilisé des échantillons aléatoires » (Association de la recherche et de l'intelligence marketing, 2007).

Pour évaluer une nouvelle méthode, comme l'échantillonnage en ligne, on effectue des études de comparabilité afin de déterminer si elle produit des résultats aussi valides que ceux de la méthode établie. Si ses résultats valent ceux de la méthode traditionnelle, la nouvelle méthode pourrait alors constituer une solution de remplacement viable. Depuis 25 ans, D.K. Shifflet & Associates mène de vastes enquêtes postales par panel auprès d'échantillons représentatifs de la population des États-Unis (50 000 par mois au cours des 15 dernières années). On utilise régulièrement ces enquêtes mensuelles pour étudier les habitudes de voyage, les voyageurs, les visiteurs et les dépenses de voyage des résidents des États-Unis. Les résultats de ces études servent à estimer le total des voyages, la part de marché des marques commerciales et le volume des voyages pour l'ensemble des États-Unis, par État et par ville. Les systèmes et les méthodes de recherche de D.K. Shifflet & Associates offrent une exactitude très proche de celle des principales mesures calculées dans divers secteurs verticaux du domaine des voyages. Établie depuis des décennies, notre riche base de données sur les habitudes de voyage permet d'établir des comparaisons avec des études gouvernementales périodiques à grande échelle et de valider l'exactitude de nos systèmes.

En 2006 et 2007, D.K. Shifflet & Associates a mené en parallèle des essais portant sur plus de 200 000 ménages en ligne selon la même périodicité (mensuelle) et en posant les mêmes questions que dans ses études postales par panel courantes, ce qui permettait de comparer directement des échantillons prélevés en ligne à un échantillon probabiliste traditionnel éprouvé. En outre, nous avons étudié la possibilité de repondérer les échantillons prélevés en ligne pour qu'ils paraissent démographiquement représentatifs de la population des États-Unis afin de déterminer si leurs questions relatives aux attitudes et aux comportements étaient représentatives.

D.K. Shifflet & Associates est manifestement bien placé pour comparer les résultats obtenus à partir d'échantillons prélevés en ligne à ceux d'échantillons probabilistes traditionnels. En prenant comme référence notre méthodologie validée, les résultats des essais menés en parallèle peuvent révéler si les échantillons prélevés en ligne représentent exactement les caractéristiques démographiques, les habitudes de voyage et les attitudes des résidents des États-Unis.

2. Analyses

Les cadres supérieurs de D.K. Shifflet & Associates possèdent des décennies d'expérience en collecte de données à l'aide des panels NPD, NFO et Market Facts/Synovate ainsi qu'une expérience approfondie auprès de plusieurs fournisseurs commerciaux d'échantillons en ligne bien établis. Pour les analyses courantes, nous avons d'abord comparé la composition de l'échantillon final, c'est-à-dire les personnes ayant répondu à l'enquête en tant que voyageurs et non-voyageurs, à partir des échantillons postaux et des échantillons prélevés en ligne. Les deux échantillons ont été conçus pour être géographiquement représentatifs et équilibrés selon l'âge, le sexe, le revenu, le niveau de scolarité et la composition du ménage.

Voici quelques exemples des principales différences entre les échantillons parallèles prélevés en ligne et les études postales de référence (tableau 4-1). De façon générale, dans la composition finale des échantillons prélevés en ligne :

- les personnes peu instruites, soit celles qui n'ont pas terminé leurs études secondaires, sont sous-représentées;
- les personnes à faible revenu, soit les ménages qui gagnent moins de 25 000 $ par année, sont sous-représentées;
- les ménages qui gagnent plus de 75 000 $ par année sont surreprésentés;
- les personnes de 45 à 65 ans sont surreprésentées et celles de 65 ans et plus sont sous-représentées.

La repondération des répondants peut compenser bon nombre de différences démographiques, mais la question essentielle demeure : les membres des panels en ligne ont-ils des attitudes différentes ou se comportent-ils différemment lorsqu'on fait abstraction des différences démographiques? Après avoir repondéré les deux échantillons pour les rendre comparables en fonction des caractéristiques démographiques (tableau 4-2), on observe dans les échantillons prélevés en ligne les différences suivantes :

- les répondants déclarent beaucoup moins de voyages d'affaires;
- les voyages effectués pour assister à des réunions d'affaires sont sous-représentés;
- les répondants font plus de voyages de même jour;
- ils voyagent beaucoup plus pour visiter des parents et des amis;
- ils déclarent des séjours de plus courte durée;
- les dépenses des visiteurs sont moins élevées.

3. Conclusion et constatations

Cette analyse révèle que les études sur les voyages qui utilisent uniquement des données en ligne n'offrent pas une représentation uniforme et exacte de la population totale des États-Unis ni de ses habitudes de voyage, même après repondération des données en fonction de caractéristiques géographiques et démographiques connues. On obtient des résultats en ligne inexacts et très variables dans le cas des mesures des attitudes et des comportements, ainsi que de la part de marché. En outre, à moins d'être fortement pondérés, les échantillons prélevés en ligne produisent une sous-estimation des habitudes de voyage, dont le nombre de visiteurs, le volume des voyages d'affaires et des congrès, les séjours de plus de 24 heures et les dépenses des visiteurs.

Sans proportions connues, il est impossible d'ajuster les échantillons prélevés en ligne. Pour obtenir des proportions connues, il faut mener de coûteuses études en parallèle portant sur des échantillons représentatifs obtenus à l'aide de méthodes hors ligne traditionnelles. Compte tenu des variations mensuelles et saisonnières historiques des voyages, on a besoin chaque mois d'un échantillon représentatif à grande échelle. À l'heure actuelle, les échantillons prélevés en ligne ne produisent pas de mesures uniformes et exactes des voyages.

On peut utiliser des échantillons prélevés en ligne, mais uniquement en tant que complément d'échantillons représentatifs prélevés fréquemment (chaque mois) qui produisent des données de référence permettant de repondérer les réponses en ligne. Ces analyses montrent que les échantillons prélevés en ligne sont moins représentatifs de la population des États-Unis que les méthodes d'échantillonnage traditionnelles comme les études postales par panel. Ce constat n'a rien d'étonnant puisqu'une bonne partie de la population des États-Unis n'utilise pas Internet.

À l'avenir, les échantillons prélevés en ligne pourraient servir à estimer avec exactitude les principales habitudes de voyage mais, dans l'immédiat, les études fondées uniquement sur l'échantillonnage en ligne risquent d'être incohérentes, inexactes et trompeuses. Pour les dirigeants d'organismes de voyage privés et publics qui ont des ressources à gérer et des comptes à rendre à leurs actionnaires ou au public, des rapports crédibles sur le volume et le profil des voyages doivent être fondés sur des échantillons probabilistes représentatifs afin que les chiffres soient valides et fiables.

4. Tableaux

Les tableaux ci-dessous montrent les différences de composition entre les échantillons finaux de répondants (échantillons de personnes qui répondent à l'enquête en tant que voyageurs et non-voyageurs) aux enquêtes en ligne, comparativement aux échantillons probabilistes.
Une différence nulle (0) indique que les résultats sont comparables à ceux de l'échantillon probabiliste.
Une différence négative (-) indique que les répondants en ligne sont moins représentés que ceux de l'échantillon postal.
L'absence de signe négatif indique que les répondants en ligne sont davantage représentés.

Tableau 4-1
Profil des échantillons finaux

	Différence en pourcentage	
	2006	2007
Âge		
18-34 ans	2	-16
35-44 ans	7	13
45-54 ans	14	28
55-64 ans	21	31
65 ans et plus	-33	-43
Éducation		
Études secondaires	-17	-22
Études collégiales partielles	13	9
Au moins un baccalauréat	9	15
Études de deuxième et troisième cycles	-12	1
Profession		
Gestion, professionnel	2	17
Technique, ventes, administration	3	3
Services	-30	-26
Agricole, forestière, pêche	-58	-57
Artisan, réparateur	-43	-19
Opérateur, ouvrier	-13	-16
Retraité, étudiant, forces armées	7	-2
Sexe		
Homme	-14	-8
Femme	8	4
Revenue du ménage		
Moins de 15 000 $	-31	-46
15 000 $ à 24 999 $	-18	-29
25 000 $ à 34 999 $	-12	-20
35 000 $ à 49 999 $	-3	-2
50 000 $ à 74 999 $	19	26
75 000 $ à 99 999 $	36	14
100 000 $ à 149 999 $	12	53
150 000 $ et plus	-7	43

Tableau 4-2
Différence comportementale dans les échantillons finaux avec neutralisation des caractéristiques démographiques

	Différence en pourcentage	
	2006	2007
Raison		
Affaires	-26	-20
Groupe d'entreprises	-30	-20
Entreprise de passage	-24	-20
Loisir	11	8
Vacances	-14	-15
Visiter des amis/membres de la parenté	49	40
Événement spécial/autre raison	8	11
Durée du séjour		
La journée	23	34
1 nuit	0	2
2 nuits	-1	-6
3-5 nuits	-12	-16
6-7 nuits	-29	-34
8-14 nuits	-34	-44
15 nuits et plus	-35	-45
Moyen de transport		
Avion	-20	-17
Train	110	58
Autobus	-34	-38
Automobile	10	11
Fougonnette de camping/Véhicule récréatif	120	146
Camion ou semi-remorque	-100	-100
Autre	-46	-59
Type de logement		
Domicile/appartement/condominium (appartient à une autre personne)	118	115
Ma résidence secondaire	-99	-98
Hôtel	-34	-31
Maison en multipropriété	-59	-60
Gîte du passant	-60	-52
Terrain de camping	-51	-57
Bateau	-95	-97
Autre (pas un hôtel)	74	111
Appartement meublé luxueux	71	-27
Dépenses de voyage		
Dépenses totales	-16	-16
Aliments	-15	-17
Transport	-6	-4
Divertissement	-15	-18
Autres dépenses	-9	-16
Magasinage	-10	-12

Bibliographie

American Association for Public Opinion Research. Opt-In Surveys and Margin of Error. Rapport inédit.

Association de la recherche et de l'intelligence marketing (2007). Code de déontologie et règles de pratique de l'Association de la recherche et de l'intelligence marketing. 10.

Horrigan, J.B. (2008), Home Broadband Adoption 2008, Pew Internet & American Life Project, Washington D.C., É.-U.

Madden, M. et Jones, S. (2008), Networked Workers. Pew Internet & American Life Project, Washington D.C., É.-U.