TIRAGE D'UN ECHANTILLON EQUILIBRE A PARTIR D'UN ECHANTILLON A PROBABILITES INEGALES, APPLICATION AU PANEL MEDIAMAT

Lorie DUDOIGNON et Aurélie VANHEUVERZWYN
Direction des Études et Méthodes Scientifiques
MÉDIAMÉTRIE

1. Introduction

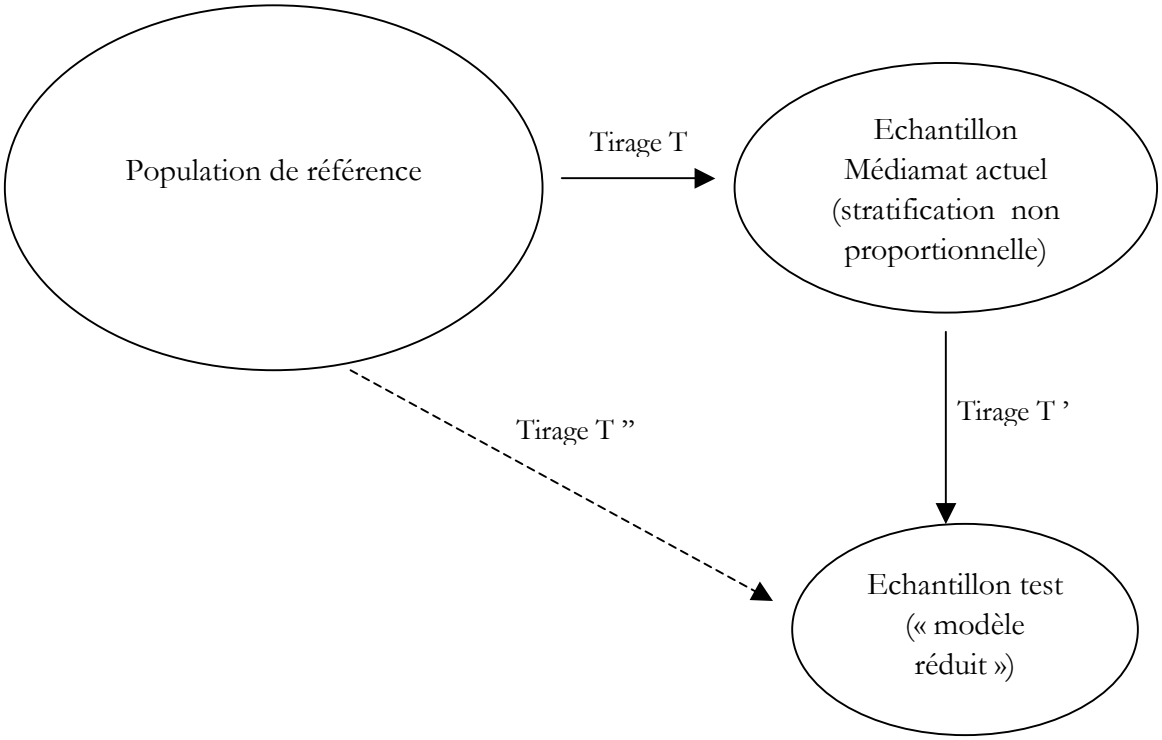
Les enquêtes par sondage doivent aujourd'hui répondre à des besoins d'analyse sur des populations rares. Ainsi, on est souvent amené à s'éloigner de la stratification avec allocation proportionnelle : l'échantillon n'est plus un « modèle réduit » de la population de référence ; les probabilités égales sont souvent considérées comme le « Graal » par les praticiens du marketing.

Dans ce contexte, Médiamétrie doit démontrer aux utilisateurs que s'éloigner volontairement et scientifiquement d'un modèle réduit n'introduit pas de déviation, et que les déformations volontaires sont corrigées sans problème par le redressement.

L'objectif de la communication est de présenter une méthode mise en place à Médiamétrie - et son application au panel Médiamat de mesure d'audience de la télévision - pour convaincre un public d'utilisateurs non avertis de l'efficacité d'un redressement. Pour cela, nous comparons les résultats obtenus sur deux échantillons : l'échantillon Mediamat (construit avec plusieurs stratifications …) et un échantillon test (sous-échantillon du premier et construit comme un modèle réduit de la population de référence.)

2. Méthodologie

2.1. Description



La méthode consiste à tirer (tirage T ’), au sein d’un échantillon existant obtenu par un tirage T (stratification non proportionnelle), un sous-échantillon représentatif en « modèle réduit » de la population de référence au regard d’un certain nombre de critères, échantillon qu’on aurait pu constituer directement par le tirage T ”. Les résultats issus de cet échantillon test sont ensuite comparés à ceux de l’échantillon Médiamétrie. On peut vérifier, en comparant les deux séries de résultats, que les déformations initiales imposées par le tirage T n’impactent pas les résultats obtenus dès lors que le redressement est adapté.

2.2. Algorithme de tirage

L’algorithme de tirage doit permettre de répondre, de manière exacte, aux contraintes de quotas marginaux. On est donc dans le cas d’un tirage équilibré.

On dispose, dans une base de sondage, de Q variables qualitatives utilisées comme critères de quotas. On désire tirer un échantillon de taille fixe n, selon un procédé aussi proche que possible du sondage aléatoire simple mais respectant des contraintes de quotas marginaux sur ces Q critères.

L’algorithme utilisé doit être capable de gérer au moins six critères de quotas.

On notera par la suite n la taille de l’échantillon à tirer et m le nombre d’unités dans la base de sondage.

Une macro développée à Médiamétrie

L'algorithme utilisé initialement a été développé en 2001. Cet algorithme permet d'effectuer des tirages avec ou sans remise sur un nombre non limité de contraintes de distribution sur variables qualitatives. Le principe consiste à répéter les étapes ci-dessous jusqu'à ce que le nombre d'unités sélectionnées soit égal à n :

1. On sélectionne une unité aléatoirement parmi celles ayant une probabilité de sélection non nulle,
2. On décrémente les quotas sur les modalités qu'elle présente,
3. Dans le cas d'un tirage sans remise, on annule la probabilité de sélection de l'unité sélectionnée,
4. Si le quota sur une des modalités de l'unité sélectionnée est saturé, on annule les probabilités de sélection de toutes les unités présentant la modalité en question,

   On peut laisser une marge de manœuvre à l'algorithme, en lui permettant une légère sortie de quota fixée à k unités, dans ce cas on annule les probabilités de sélection dès lors que le nombre d'unités sélectionnées atteint la valeur de quota + k.
5. On compte le nombre d'unités ayant une probabilité de sélection non nulle :

   Si ce nombre est égal à zéro, on arrête le tirage,

   Sinon, on continue le tirage en repartant à l'étape 1.

Dans le cas du tirage sans remise, l'ordre de sélection des unités peut conduire à ne pas atteindre la taille d'échantillon désirée : recherche du mouton à cinq pattes. Dans ce cas, on réitère la procédure jusqu'à atteindre la taille d'échantillon désirée. Le nombre d'itérations augmente en fonction du taux de sondage et de la distance entre les deux structures (celle de la base de sondage et celle de l'échantillon.)

Cet algorithme peut donc s'avérer, dans les cas complexes, coûteux en temps. C'est pourquoi il nous a semblé avantageux d'utiliser la méthode du cube mise au point par Deville et Tillé (2000.)

La macro CUBE

La macro CUBE a été développée à l'INSEE et mise gratuitement à disposition en avril 2004 sur le site Internet de l'INSEE. Initialement écrite par des élèves de l'ENSAI, elle a été finalisée par Frédéric Tardieu et Bernard Weytens.

La problématique consiste à tirer n unités parmi les m de la base de sondage en imposant à l'échantillon final une structure différente de l'échantillon initial.

L'étape préalable consiste à affecter une probabilité d'inclusion à chaque unité de la base de sondage. Cette probabilité d'inclusion doit prendre en compte les contraintes de quotas marginaux que l'on souhaite imposer. Pour calculer ces probabilités d'inclusion, on utilise la macro CALMAR développée à l'INSEE par Olivier Sautory (1993.)

La fonction de distance utilisée doit être bornée afin de garantir un jeu de poids comparable à des probabilités (c'est-à-dire compris entre 0 et 1, à un facteur multiplicatif près.) On a choisi la méthode « logit » en imposant comme borne supérieure m/n (on aurait aussi pu prendre la méthode linéaire tronquée qui présente les mêmes caractéristiques que la méthode logit.)

On calcule ensuite la probabilité d'inclusion comme le produit du poids de redressement par le taux de sondage n/m. Enfin, on lance le tirage avec la probabilité d'inclusion comme seule variable de contrôle. L'option d'atterrissage choisie est l'option A.

3. Application au panel Médiamat de mesure d'audience de la télévision

3.1. Le panel Médiamat

Médiamat est le dispositif permettant de mesurer l'audience des chaînes de télévision en France métropolitaine (en dehors de la Corse.) Il résulte d'une étude permanente d'audience de la télévision, réalisée par Médiamétrie auprès d'un panel national d'environ 8000 personnes âgées de 4 ans et plus possédant au moins un téléviseur dans leur résidence principale, obtenu par grappage de 3150 foyers.

Pour observer les comportements individuels de vision de la télévision, Médiamétrie installe dans chaque foyer faisant partie du panel Médiamat un ou plusieurs audimètres, munis de télécommande à touches individuelles, qui enregistrent à la seconde toutes les utilisations du (ou des) téléviseur(s) du foyer comme la marche et l'arrêt du téléviseur, l'écoute des différentes chaînes, l'utilisation du magnétoscope ou l'utilisation du téléviseur pour des jeux vidéo ou comme moniteur, et la présence de chacun des membres du foyer.

Composition du panel Médiamat

En 1989, le panel Médiamat était représentatif des foyers français équipés d'un téléviseur, représentatif au sens proportionnel à la population de référence sur les critères de région, du nombre de personnes dans le foyer, de l'âge du chef de ménage, de la catégorie socioprofessionnelle du chef de ménage et de la présence d'enfants dans le foyer.

En 1997, une première « déformation » a été introduite dans le panel Médiamat pour pallier le creux des naissances et garantir une précision statistique des résultats sur les cibles à effectifs faibles (enfants, adolescents ….) Cette déformation consiste en une stratification des foyers selon l'âge du chef de ménage :

- strate 1 : foyers dont le chef a moins de 50 ans,
- strate 2 : foyers dont le chef a 50 ans ou plus.

A mesure du remplacement des foyers, ceux de la strate 1 ont été sur-représentés ; l'objectif, progressivement atteint, était que la strate 1 contienne 60% des foyers du panel et la strate 2 40% alors que, dans la population de référence, chaque strate contient une proportion proche de 50% des foyers.

Fin 1999, à la demande de France 3, le panel est enrichi sur 6 régions INSEE de façon à permettre l'élaboration de résultats statistiquement très fiables au niveau de chaque région France 3 (qui sont toutes des unions logiques de régions INSEE.)

Cumulé avec le précédent, cet aménagement revient à une stratification croisant les 21 régions INSEE métropolitaines avec le critère « foyers dont le chef a moins de 50 ans / foyers dont le chef a 50 ans ou plus », soit 42 strates en tout, pour chacune desquelles le nombre total réel de foyers est connu.

Début 2001 est introduit le sur-échantillon MédiaCabSat composé de foyers abonnés à une offre télévisuelle élargie[1]. Le sur-échantillon MédiaCabSat comprend lui-même des « déformations » sur trois critères : présence de ménagère avec enfant de moins de 15 ans, présence d'enfant de 4 à 10 ans et présence de jeune de 15 à 24 ans, pour permettre l'élaboration de résultats fiables sur les cibles stratégiques des chaînes thématiques.

[1] Offre élargie : abonnement Canal Satellite, TPS, câble numérique ou câble analogique plus de 15 chaînes.

Le redressement dans le panel Médiamat

Depuis ses débuts le panel Mediamat est redressé par la méthode RAS tronqué (bornage des poids de redressement.) et ceci en 2 étapes. La première étape consiste à redresser au niveau des foyers, la deuxième à redresser au niveau des individus (le poids initial d'un individu étant celui du foyer auquel il appartient.)

Les critères de redressement ont été modifiés au fil des évolutions du panel (introduction des sur-échantillons …) et de celles de la population de référence (apparition de nouveaux équipements audiovisuels …)

Il existe aujourd'hui 17 critères de redressement foyer (la région, l'habitat, le nombre de personnes au foyer … et des critères d'équipement audiovisuel) et 2 critères individu (l'âge et le sexe.)

3.2. Description des tests réalisés

Objectif

L'objectif est de « reproduire » le panel Médiamat tel qu'il était en 1997, c'est-à-dire sans les sur-échantillons France 3 et MédiaCabSat et de comparer les résultats d'audience obtenus sur cet échantillon avec ceux du Médiamat. Ainsi, nous voulons démontrer que les évolutions de structure du panel n'ont pas d'impact sur les résultats d'audience.

Méthode

Un échantillon de 2200 foyers (taille du panel de 1997) a été tiré de manière aléatoire, avec quotas, parmi les 3150 foyers constituant le Panel Médiamat. Les quotas utilisés permettent de garantir la représentativité de cet échantillon par rapport à la population française équipée TV (l'offre élargie et les « petites » régions France 3 ne sont donc pas sur-représentées.) La stratification sur l'âge du chef a été maintenue (sur-représentation des foyers dont le chef a moins de 50 ans.)

Les quotas utilisés pour ce tirage sont les suivants :

- Région UDA
- Age du chef de ménage
- CSP du chef de ménage
- Nombre de personnes au foyer
- Présence d'enfant
- Type d'offre (restreinte ou élargie)

Afin de valider le tirage ainsi obtenu, la structure de l'échantillon a été comparée à la structure théorique de la population équipée TV. La chaîne de production du Médiamat a ensuite été appliquée à l'échantillon de 2200 foyers (même redressement, même mode de calcul des indicateurs d'audience …)

3.3. Résultats

Deux tests ont été réalisés. Pour chacun de ces tests, un échantillon a été tiré et les résultats d'audience de cet échantillon ont été produits sur trois cibles, cinq tranches horaires, huit chaînes et sept jours. Au total sur les deux tests, on a 1680 cases de comparaison.

Structure de l'échantillon

Pour chacun des tests, les deux algorithmes ont été appliqués pour effectuer le tirage. Pour pouvoir satisfaire les contraintes de taille de l'échantillon et de vitesse de convergence, un écart absolu de 3,5% avec la distribution théorique a du être accordé au tirage fait avec l'algorithme développé à Médiamétrie.

		STRUCTURE THEORIQUE	STRUCTURE ECHANTILLON TEST (macro Médiamétrie)	STRUCTURE ECHANTILLON TEST (macro Cube)	ECARTS = THEORIQUE - TEST(macro Médiamétrie)	ECARTS = THEORIQUE - TEST(macro Cube)
Région UDA	IDF	18,7	15,3	18,5	3,5	0,2
	Nord	6,6	6,0	6,5	0,6	0,1
	Est	9,3	10,5	10,1	-1,2	-0,8
	BP Est	7,8	9,0	7,7	-1,2	0,2
	BP Ouest	9,4	10,6	9,7	-1,2	-0,3
	Ouest	13,2	14,5	13,3	-1,2	-0,1
	Sud ouest	11,4	12,7	11,1	-1,2	0,3
	Sud est	11,5	10,1	11,1	1,4	0,4
	Mediterrannée	12,0	11,2	12,0	0,7	0,0
Catégorie d'agglo	Rurale	24,8	24,5	23,7	0,3	1,1
	<20,000 hab	16,9	17,9	16,6	-1,0	0,2
	de 20 à 100,000 hab	12,9	14,2	14,0	-1,3	-1,1
	>=100,000 hab	28,5	29,9	29,5	-1,4	-1,0
	agglo parisienne	16,9	13,5	16,1	3,4	0,8
NPF	1	27,7	25,6	26,4	2,1	1,3
	2	29,3	30,2	30,7	-0,9	-1,3
	3	17,2	15,9	16,5	1,3	0,6
	4	16,9	18,0	17,0	-1,2	-0,2
	5 ou +	8,9	10,3	9,4	-1,3	-0,4
CSP du chef	Agriculteurs	2,1	1,8	1,8	0,2	0,2
	Artisans, commerçants	4,4	3,0	3,0	1,4	1,4
	Cadres	11,3	12,1	11,9	-0,8	-0,6
	Prof. Interm	15,1	16,8	14,9	-1,7	0,2
	Employés	12,3	15,2	15,2	-3,0	-3,0
	Ouvriers	22,9	20,4	21,3	2,5	1,7
	Inactifs	32,0	30,7	31,9	1,3	0,1
Age du chef	<35 ans	22,6	18,5	19,0	4,1	3,6
	de 35 à 49 ans	37,4	40,2	41,2	-2,9	-3,8
	de 50 à 64 ans	18,8	20,5	18,2	-1,7	0,6
	>= 65 ans	21,2	20,8	21,6	0,5	-0,4
Age de la ménagère	<35 ans	23,0	20,7	20,1	2,3	2,9
	de 35 à 49 ans	30,8	33,8	34,0	-2,9	-3,2
	de 50 à 64 ans	15,9	17,2	15,9	-1,4	-0,1
	>= 65 ans	18,2	16,1	16,7	2,0	1,4
	RDA homme	12,1	12,2	13,3	-0,1	-1,1
Présence d'enfant	Présence	32,4	33,6	32,4	-1,2	0,1
	Absence	67,6	66,4	67,6	1,2	-0,1
Equipement TV actif	1 poste	58,2	59,3	60,8	-1,1	-2,7
	2 postes ou +	41,8	40,7	39,2	1,1	2,7
Equipement magnétoscope	Non équipé	17,7	15,3	16,0	2,4	1,7
	Equipé	82,3	84,7	84,0	-2,4	-1,7
Type d'offre	Offre Elargie	22,1	23,4	22,1	-1,2	0,0
	Offre Restreinte	77,9	76,6	77,9	1,2	0,0

Tableau 1 : Comparaison des structures obtenues lors du dernier test avec les deux macros de tirage.

Les deux méthodes nous permettent de sélectionner l'échantillon souhaité. La macro Cube permet d'obtenir une meilleure adéquation entre la structure de l'échantillon et celle souhaitée et ceci dans des temps beaucoup plus courts.

Analyse des écarts de taux moyen d'audience.

L'indicateur de référence dans le cadre du Mediamat est le Taux Moyen d'audience. C'est donc sur cet indicateur que nous avons effectué nos comparaisons.

Définition du Taux Moyen d'audience

On considère la variable de base :

X(i, c, s) = 1 si l'individu (ou le foyer) i regarde la chaîne c à la seconde s

X(i, c, s) = 0 si non.

Le taux moyen d'audience de l'émission E de durée S, diffusée sur la chaîne c, est définie comme :

(1) $TM(E, c) = \Sigma_i [\pi(i) \Sigma_s X(i, c, s)/S] / N$

ou encore :

(2) $TM(E, c) = \Sigma_s [\Sigma_i \pi(i) X(i, c, s)/N] / S$

où $\pi(i)$ est le poids de l'individu i après redressement et N est la somme des poids des individus de la cible, $\Sigma_i \pi(i) = N$.

Le Taux Moyen d'Audience peut donc être interprété comme :

(1) la moyenne, calculée sur la cible, des proportions de vision de E : l'individu moyen de la population étudiée a regardé TM (en %) de la durée de l'émission E

(2) la moyenne, calculée sur toute la durée de E exprimée en secondes, de la proportion de personnes regardant l'émission E à la seconde s

Dans les tableaux 2 et 3 sont présentés des extraits des résultats obtenus lors des deux tests.

		Test n°1		Test n°2	
Jour nommé	**Cible**	**Echantillon**	**Médiamat**	**Echantillon**	**Médiamat**
Lundi	**Foyer**	25,5	25,4	25,4	25,4
	4+	14,7	14,8	15,1	15,1
	15+	15,9	16,0	16,1	16,0
Mardi	**Foyer**	24,4	24,2	24,4	24,5
	4+	13,8	13,8	14,5	14,5
	15+	14,9	14,8	15,2	15,3
Mercredi	**Foyer**	25,1	25,4	24,8	25,0
	4+	14,7	14,7	14,9	14,6
	15+	15,0	15,2	15,3	15,2
Jeudi	**Foyer**	25,7	25,7	25,0	25,0
	4+	15,3	15,4	14,6	14,6
	15+	15,2	15,4	15,5	15,6
Vendredi	**Foyer**	25,5	25,0	25,5	25,7
	4+	15,2	14,9	15,4	15,5
	15+	15,4	15,2	16,2	16,1
Samedi	**Foyer**	25,2	25,2	26,2	26,4
	4+	15,4	15,4	15,6	16,0
	15+	16,0	15,9	16,5	16,8
Dimanche	**Foyer**	28,0	27,7	29,2	29,4
	4+	17,5	17,1	18,0	17,8
	15+	18,4	18,0	19,3	19,1

Tableau 2 : Taux moyen d'audience entre 3h et 27h sur le Total TV par jour nommé et par cible

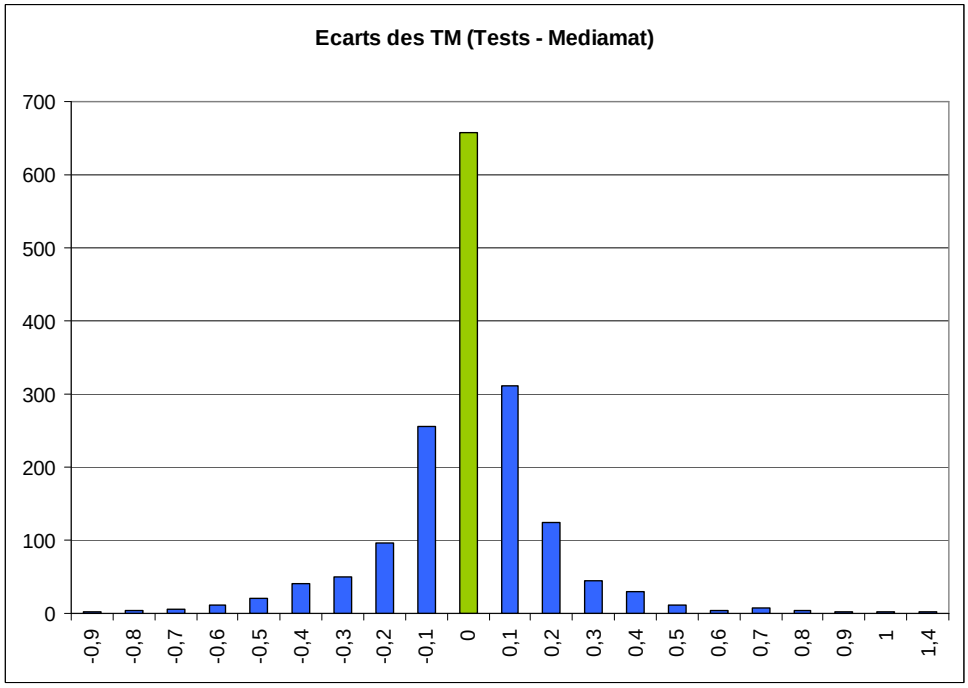
		Test n°1		Test n°2	
Cible	**Tranche horaire**	**Echantillon**	**Médiamat**	**Echantillon**	**Médiamat**
Foyer	**03:00 - 27:00**	25,6	25,5	25,8	25,9
	12:00 - 14:00	38,6	38,3	37,8	38,1
	19:00 - 20:00	55,5	55,3	56,8	56,9
	20:00 - 21:00	66,7	67,1	67,7	67,4
	21:00 - 22:30	68,4	69,0	68,0	68,0
4+	**03:00 - 27:00**	15,2	15,2	15,4	15,4
	12:00 - 14:00	23,5	23,1	23,1	23,2
	19:00 - 20:00	35,0	34,7	35,9	35,9
	20:00 - 21:00	44,2	44,4	44,8	44,5
	21:00 - 22:30	44,9	45,2	44,5	44,5
15+	**03:00 - 27:00**	15,8	15,8	16,3	16,3
	12:00 - 14:00	24,3	24,0	24,3	24,4
	19:00 - 20:00	36,3	36,1	37,4	37,4
	20:00 - 21:00	46,6	46,9	47,8	47,5
	21:00 - 22:30	48,4	48,7	48,7	48,9

Tableau 3 : Taux moyen d'audience en moyenne Lundi-Dimanche sur le Total TV par cible et par tranche horaire

Que ce soit en total journée ou sur des tranches horaires plus fines, les écarts sont faibles et ceci sur l'ensemble des jours et des cibles étudiées.

Des tests de Student ont été réalisés sur l'ensemble des résultats et seulement un écart est significatif sur les 1680 cases étudiées.

Dans l'ensemble, comme on peut le voir sur le graphique 1, la distribution des écarts est bien concentrée autour de 0 (la moyenne.)



Graphique 1 : Distribution des écarts.

Nous avons ensuite étudié plus en détail le signe des écarts afin de s'assurer que celui ci n'est pas systématique pour une chaîne, une cible ou encore une tranche horaire ...

		Jour1	Jour2	Jour3	Jour4	Jour5	Jour6	Jour7	Jour8	Jour9	Jour10	Jour11	Jour12	Jour13	Jour14
Foyer	03:00 - 27:00	=	=	=	=	=	=	=	=	=	=	=	=	=	=
Foyer	12:00 - 14:00	+	=	=	-	+	=	=	+	=	+	+	=	+	=
Foyer	19:00 - 20:00	=	+	=	=	+	+	=	+	=	+	+	-	-	+
Foyer	20:00 - 21:00	=	=	-	-	=	=	=	=	=	+	=	=	=	+
Foyer	21:00 - 22:30	=	-	=	+	+	+	+	=	-	-	+	=	=	-
4+	03:00 - 27:00	=	=	=	=	=	=	=	=	=	=	=	=	=	=
4+	12:00 - 14:00	=	=	+	=	=	=	+	=	=	=	=	=	=	=
4+	19:00 - 20:00	=	=	-	=	+	+	=	+	=	+	=	=	-	+
4+	20:00 - 21:00	=	=	-	-	=	=	=	=	+	+	=	=	=	+
4+	21:00 - 22:30	=	-	=	+	=	+	+	=	=	-	=	+	-	-
15+	03:00 - 27:00	=	=	=	=	=	=	=	=	=	=	=	=	=	=
15+	12:00 - 14:00	=	=	=	-	=	=	+	=	=	=	=	=	=	=
15+	19:00 - 20:00	=	+	-	=	+	+	-	+	=	+	=	-	-	+
15+	20:00 - 21:00	=	=	=	-	=	=	=	=	=	+	=	=	=	+
15+	21:00 - 22:30	=	-	=	+	=	+	+	=	=	-	=	+	-	-

Tableau 4 : Exemple de tableau étudié pour une des chaînes Hertziennes.

Sur l'ensemble des tableaux étudiés, aucun écart systématique n'apparaît.

4. Conclusion

Les résultats de ces tests nous permettent de conclure que les déformations introduites volontairement dans le plan de sondage du panel Médiamat n'engendrent pas de biais systématique que le redressement ne serait capable de corriger. Les résultats d'audience ne sont donc pas impactés par les introductions successives de sur-échantillons.

Cette approche pragmatique permet de convaincre les utilisateurs des résultats de Médiamétrie n'ayant pas nécessairement de culture des sondages, que « représentativité » ne veut pas forcément dire « modèle réduit. »

Cette méthodologie, pour l'instant appliquée à Mediamat, devrait être étendue à d'autres études de Médiamétrie.

5. Bibliographie

ARDILLY, P. (1991), *Échantillonnage représentatif optimum à probabilités inégales*, Annales d'Économie et de Statistique, 23, 91-113.

ARDILLY, P. (1994), *Les techniques de sondage*, Paris, Technip.

BROSSIER, G. et DUSSAIX A.-M. (éds.), *Enquêtes et sondages : Méthodes, modèles, applications, nouvelles approches*, Paris, Dunod.

BOUSABAA, A., LIEBER J. et SIROLLI R. (1999)**,** *La macro Cube*, Technical report, INSEE, Rennes.

COCHRAN, W.-G. (1977), *Sampling Techniques*, 3ème édition, New-York, Wiley.

DEVILLE, J.-C. et TILLÉ Y. (2001), Échantillonnage par la méthode du Cube, variance et estimation de variance, *in* DROESBEKE, J.-J. et LEBART L. *Enquêtes, modèles et applications*, 344-363, Paris, Dunod.

DUSSAIX, A.-M. et GROSBRAS J.-M. (1993), *Les sondages : principes et méthodes*, Paris, PUF.

DEVILLE, J.-C., SÄRNDAL C.-E. et SAUTORY O. (1993), Generalized raking procedure in survey sampling, *Journal of the American Statistical Association*, 88, 1013-1020.

ROY, G. et VANHEUVERZWYN A. (2001), Un algorithme de tirage équilibré, *in* DROESBEKE, J.-J. et LEBART L. (éds.), *Enquêtes, modèles et applications*, 344-363, Paris, Dunod.

ROY, G. et VANHEUVERZWYN A. (2001), Redressement par la macro CALMAR : applications et pistes d'amélioration, *in* LEJEUNE M. (éd), *Traitement des fichiers d'enquêtes*, Grenoble, PUG.

SAUTORY, O. (1993), La macro CALMAR Redressement d'un échantillon par calage sur marges, *Série des documents de travail de la Direction des Statistiques Démographiques et Sociales*, F 9310.

TASSI, P. (2005), *Modèles statistiques de la mesure d'audience des médias audiovisuels*, Paris, Economica.

TILLÉ, Y. (2001), *Théorie des sondages*, Paris, Dunod.