

Journées de Méthodologie Statistique de l'INSEE



*Tirage d'un échantillon équilibré
à partir d'un échantillon à
probabilités inégales
Application au panel Médiamat*

15 mars 2005



Introduction





Introduction

- **Les enquêtes par sondage doivent aujourd'hui répondre à de nouveaux besoins**
 - Besoins d'analyse sur des populations rares
 - ⇒ Les cibles rares sont sur-représentées dans les échantillons pour pouvoir être analysées
 - ⇒ L'échantillon n'est plus un modèle réduit de la population de référence
- **Comme l'a démontré Neyman dans les années 30**
 - il est souvent utile de déformer intentionnellement l'échantillon par rapport à la structure de la population



Introduction

➤ **Quel impact sur les résultats ?**

- Il convient de s'assurer que la déformation de l'échantillon n'engendre pas de biais systématique dans les résultats

➤ **Pour cela...**

- Le redressement est la réponse **théorique** :
 - mais dans le cas de stratifications multiples, il convient d'aller plus loin pour s'assurer que la complexité du plan de sondage ne crée pas de biais que le redressement ne peut corriger
- La réponse **pratique** :
 - tirage d'un échantillon « modèle réduit » de la population de référence à partir de l'échantillon stratifié initial



Plan de la présentation

- **Description de la méthodologie**
 - Principe
 - Algorithme de tirage
- **Application au panel Médiamat**
 - Description du panel Médiamat
 - Description des tests réalisés
 - Résultats des tests
- **Conclusion**



1 - Description de la méthodologie



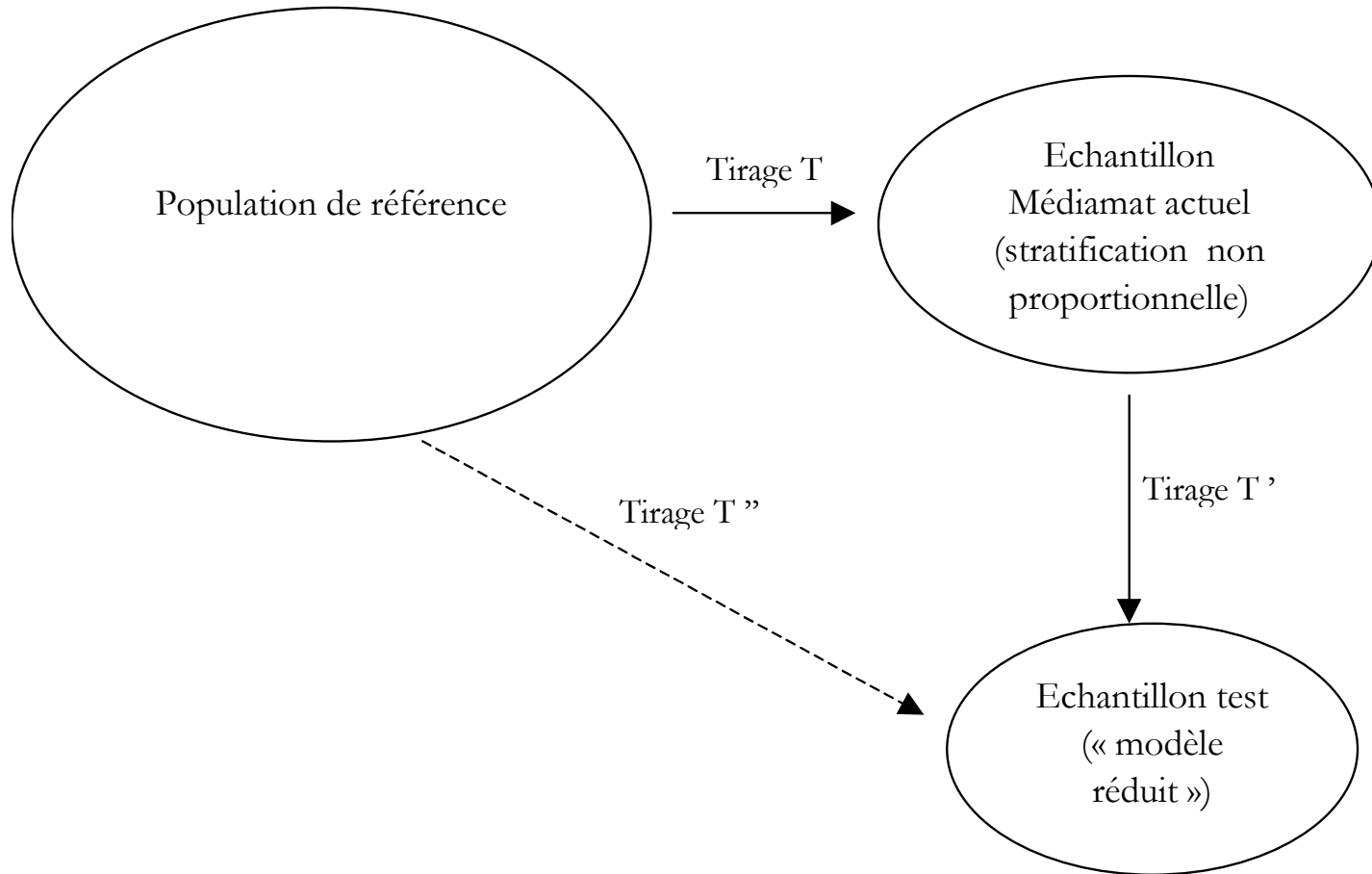
1.1 - Principe

➤ Les étapes

- Tirage, au sein d'un échantillon existant obtenu par stratification non proportionnelle, d'un sous-échantillon représentatif en « modèle réduit » de la population de référence au regard d'un certain nombre de critères
- Production des résultats d'audience sur les deux échantillons
- Comparaison des deux séries de résultats



1.1 - Principe





1.2 – Algorithme de tirage

➤ Description de l'algorithme

- On dispose de :
 - Une base de sondage de m observations
 - Q variables qualitatives utilisées comme critères de quotas
- Objectif :
 - Tirer un échantillon de taille fixe n respectant des contraintes de quotas marginaux sur ces Q critères



1.2 – Algorithme de tirage

➤ Une macro SAS développée à Médiamétrie

– Principe :

- Répéter les étapes ci-dessous jusqu'à ce que le nombre d'unités sélectionnées soit égal à n :
 - On sélectionne une unité aléatoirement parmi celles ayant une probabilité de sélection non nulle,
 - On décrémente les quotas sur les modalités qu'elle présente,
 - Dans le cas d'un tirage sans remise, on annule la probabilité de sélection de l'unité sélectionnée,
 - 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 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.

– Inconvénient :

- Algorithme qui peut s'avérer, dans les cas complexes, coûteux en temps



1.2 – Algorithme de tirage

➤ La macro CUBE

- Contrainte :
 - Tirage de n unités parmi les m de la base de sondage en imposant à l'échantillon final une structure différente de l'échantillon initial
- Étape préalable :
 - Affecter à chaque unité une probabilité d'inclusion prenant en compte les contraintes de quotas marginaux
 - Pour cela, utilisation de la macro CALMAR avec fonction de distance bornée
 - Borne inférieure = 0
 - Borne supérieure = m / n
 - Probabilité d'inclusion = poids de redressement $\times n / m$
- Tirage :
 - Utilisation de la macro CUBE avec la probabilité d'inclusion comme seule variable de contrôle
 - L'option d'atterrissage choisie est l'option A



2 – Application au panel Médiamat



2.1. Description du panel Médiamat

➤ Médiamat

- Dispositif de référence pour la mesure de l'audience des chaînes de télévision en France métropolitaine (en dehors de la Corse)
 - Panel d'environ 8000 individus âgés de 4 ans et plus possédant au moins un téléviseur dans leur résidence principale, obtenu par grappage de 3150 foyers
 - Installation dans chaque foyer d'un ou plusieurs audimètres, munis de télécommande à touches individuelles, qui enregistrent à la seconde toutes les utilisations du ou des téléviseurs du foyer



2.1. Description du panel Médiamat

➤ Médiamat

– Historique :

- 1989 : le panel Médiamat est un modèle réduit des foyers français équipés d'un téléviseur au regard de critères régionaux et socio-démographiques
- 1997 : introduction d'une première sur-représentation des foyers dont le chef a moins de 50 ans
- 1999 : introduction, à la demande de France 3, de sur-représentations régionales
- 2001 : introduction du sur-échantillon MédiaCabSat composé de foyers abonnés à une offre télévisuelle élargie



2.1. Description du panel Médiamat

➤ Le redressement du panel Médiamat

- RAS tronqué en deux étapes :
 - Étape 1 : Redressement au niveau des foyers
 - Étape 2 : Redressement au niveau des individus, le poids initial d'un individu étant celui du foyer auquel il appartient
- Critères de redressement modifiés au fil des évolutions du panel et de celles de la population de référence
 - Aujourd'hui, 17 critères foyers et 2 critères individus



2.2. Description des tests réalisés

➤ **Objectif :**

- Reproduire le panel Médiamat tel qu'il était en 1997

➤ **Méthode :**

- Tirage d'un échantillon de 2200 foyers parmi les 3150 du panel actuel
- Quotas :
 - Région UDA
 - Age du chef de ménage
 - PCS du chef de ménage
 - Nombre de personnes dans le ménage
 - Présence d'enfants
 - Type d'offre télévisuelle (restreinte ou élargie)



2.3. Résultats des tests

➤ Test de l'algorithme :

- Les deux algorithmes ont été appliqués pour effectuer le tirage
- Macro Médiamétrie :
 - Un écart absolu de 3,5% par quota a du être accordé pour que le tirage soit réalisable en moins de 1000 itérations
- Macro CUBE :
 - Meilleure adéquation entre la structure de l'échantillon sélectionné et celle souhaitée
 - Temps de traitement beaucoup plus courts

2.3. Résultats des tests

➤ Comparaison des structures :

		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
	Méditerranée	12,0	11,2	12,0	0,7	0,0
Catégorie d'agglomération	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
	agglomération parisienne	16,9	13,5	16,1	3,4	0,8
Niveau de qualification (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



2.3. Résultats des tests

➤ Description des analyses des résultats d'audience

- Tirage de deux échantillons tests
 - Test n°1 : novembre 2003
 - Test n°2 : mars 2004
- Indicateur analysé
 - Taux moyen d'audience par jour nommé, par cible, par tranche horaire et par chaîne
 - 1680 points de comparaison

2.3. Résultats des tests

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

Jour nommé	Cible	Test n°1		Test n°2	
		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

2.3. Résultats des tests

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

Cible	Tranche horaire	Test n°1		Test n°2	
		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



2.3. Résultats des tests

➤ Bilan

- Des écarts très faibles
 - Seulement un écart significatif sur les 1680 points étudiés
- Signe des écarts
 - Distribution des écarts autour de 0
 - Pas d'écart systématique pour une chaîne, une cible ou une tranche horaire



Conclusion



Conclusion

➤ Résultats :

- Les déformations introduites volontairement dans le plan de sondage n'engendrent pas de biais systématique que le redressement ne serait capable de corriger

➤ Méthode :

- Approche pragmatique ayant permis de convaincre les utilisateurs des résultats de Médiamétrie que « représentativité » ne veut pas forcément dire « modèle réduit »



Merci pour votre attention

Lorie DUDOIGNON

ldudoignon@mediametrie.fr

Aurélie VANHEUVERZWYN

avanheuverzwyn@mediametrie.fr

Médiamétrie

Direction des Études et Méthodes Scientifiques

Imaginez un monde sans mesure




mediametrie