Colloque francophone sur les sondages 2005

PRECISION DES INDICATEURS D'AUDIENCE DE LA TELEVISION

Lorie Dudoignon et Aurélie Vanheuverzwyn[1]

RÉSUMÉ

Médiamat est le dispositif de référence de la mesure d'audience de la télévision en France au niveau national. Il résulte d'une étude permanente d'audience de la télévision, réalisée par Médiamétrie auprès d'un panel national d'environ 3150 foyers. Ce dispositif permet à Médiamétrie de fournir au marché des résultats d'audience au quotidien. Le principal indicateur d'audience est le Taux Moyen. L'objectif des travaux présentés dans cet article est de calculer la précision statistique de cet indicateur d'audience. Dans un premier temps, nous étudierons cette précision statistique de manière analytique et dans un deuxième temps nous utiliserons des méthodes de réplication (Bootstrap) pour estimer numériquement cette précision.

1. INTRODUCTION

Médiamat est le système 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 3150 foyers dont l'ensemble des individus âgés de 4 ans et plus participent à l'étude.
Pour observer les comportements individuels de consommation 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, la vision des différentes chaînes, l'utilisation du magnétoscope ou l'utilisation du téléviseur pour des jeux vidéo ou comme moniteur, ainsi que la présence de chacun des membres du foyer.

Cet outil permet à Médiamétrie de fournir à ses clients, médias ou publicitaires, des résultats d'audience au quotidien. Le marché a des besoins en analyse toujours plus fins et est de ce fait de plus en plus exigeant sur la qualité des études et l'interprétation qu'ils peuvent faire des chiffres qu'ils utilisent. Aussi, Médiamétrie a souhaité entamer un programme de recherche dont la finalité est de fournir aux clients de Médiamat la précision des indicateurs qu'ils utilisent soit par des abaques, soit par un calcul à la volée dans les logiciels de consultation.

L'un des problèmes posés dans l'estimation de la précision des indicateurs d'audience est l'effet de grappe qui survient dans les tirages à plusieurs degrés. En effet, dans le Médiamat, on sélectionne des foyers et on interroge l'ensemble des individus des foyers sélectionnés. Cela induit une corrélation entre individus. Ainsi, l'audience d'un individu d'un foyer peut conditionner fortement celle des autres individus du foyer ce qui implique généralement une perte de précision du fait de l'ajout d'un coefficient de corrélation.

Dans cet article seul l'indicateur principal du Médiamat est étudié ; il s'agit du Taux Moyen d'audience. Les travaux de Médiamétrie ont toutefois porté sur tous les indicateurs.

La première partie reprend l'expression analytique de la variance du Taux Moyen et la seconde partie décrit la méthode empirique utilisée pour les estimations d'intervalle de confiance ainsi que les principaux résultats obtenus.

[1] Lorie Dudoignon, Responsable Scientifique, Médiamétrie, 55/63 rue Anatole France 92532 Levallois-Perret Cedex, France, ldudoignon@mediametrie.fr, Aurélie Vanheuverzwyn, Directrice Scientifique, Médiamétrie, 55/63 rue Anatole France 92532 Levallois-Perret Cedex, France, avanheuverzwyn@mediametrie.fr.

2. METHODE ANALYTIQUE

2.1 Notations

- E est une tranche horaire ou une émission ; on note Dd(c,E), la durée de diffusion de la chaîne c durant la tranche horaire E.
- i désigne un individu de la cible étudiée, i = 1 à n, n étant l'effectif brut de cette cible dans le panel (brut = avant redressement.)
- $\pi(i)$ est le poids (coefficient de pondération) de l'individu i après redressement ; N est la somme des coefficients de pondérations des membres permanents du panel de la cible, $\Sigma_i\ \pi(i) = N$, i variant de 1 à n.
- C est le nombre de chaînes ; c est l'indice courant d'une chaîne de C, c = 1 à C ; le média TV, noté TTV, est l'union logique de toutes les chaînes : TTV = $\cup_{c\in C}(c)$.
- La variable de base est :

  X(i, c, s) = 1 si i regarde c à la seconde s

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

2.2 Définition du Taux Moyen

Pour l'individu i, le temps passé sur la chaîne c pendant la tranche E est alors (en pondéré) :

$$D(i,c,E) = \pi(i)\times \sum_{s\in E} X(i,c,s)$$

Ce calcul étant fait pour chaque individu i et chaque chaîne c, la durée de vision totale de E sur c est :

$$D(c,E) = \sum_i D(i,c,E)$$

Par définition, le taux moyen de la chaîne c pendant la tranche horaire ou l'émission E est donné par :

$$TM(c,E) = \frac{D(c,E)}{N\times Dd(c,E)}$$

On peut aussi l'écrire :

$$(1)\ \ TM(c,E) = \sum_i \frac{D(i,c,E)}{N\times Dd(c,E)} = \frac{1}{N}\sum_i \left[\pi(i)\sum_{s\in E}\frac{X(i,c,s)}{Dd(c,E)}\right]$$

ou encore :

$$(2)\ \ TM(c,E) = \frac{1}{Dd(c,E)}\sum_{s\in E}\left[\sum_i \frac{\pi(i)\times X(i,c,s)}{N}\right]$$

Le Taux Moyen d'audience s'interprète donc de deux façons :

- sous la forme (1), TM(c, E) est la moyenne, sur l'ensemble des individus, des proportions de vision de E : l'individu moyen de la population étudiée a regardé TM (en %) de la durée de l'émission E,
- sous la forme (2), TM(c, E) est 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.

2.3 Précision du Taux Moyen

On note $Z(i) = Z(i,c,E) = \dfrac{D(i,c,E)}{Dd(c,E)}$ la proportion de vision de l'individu i sur la chaîne c pendant E.

Remarque : *Z désigne la variable aléatoire à valeurs dans [0, 1] représentant la part de vision de E ; Z est caractérisée par une loi de probabilité qui décrit la façon dont Z varie sur l'ensemble de la population étudiée, ainsi que par sa moyenne m(Z) = E(Z) et sa variance $\sigma^2(Z)$ = V(Z). On appelle m(Z) le taux d'audience théorique de E dont TM est un estimateur (sans biais).*

On a, bien sûr, puisque $Z = \dfrac{D}{Dd}$: $V(D) = \sigma^2(D) = Dd^2 \times \sigma^2(Z)$.

En prenant la variance de TM(c, E) et en développant :

$$(3)\quad V(TM(c,E)) = \frac{1}{N^2}\left[\sum_i \pi^2(i)V(Z(i)) + \sum_i \sum_{j \neq i} \pi(i)\pi(j)Cov(Z(i),Z(j))\right]$$

Toutes les variances V(Z(i)) étant égales à $\sigma^2(Z)$, et en notant $\rho(i, j)$ le coefficient de corrélation linéaire entre Z(i) et Z(j), on obtient :

$$(4)\quad V(TM(c,E)) = \frac{\sigma^2(Z)}{N^2}\left[\sum_i \pi^2(i) + \sum_i \sum_{j \neq i} \pi(i)\pi(j)\rho(i,j)\right]$$

Notons par R le terme $\sum_i \pi^2(i) + \sum_i \sum_{j \neq i} \pi(i)\pi(j)\rho(i,j)$: $V(TM(c,E)) = R\dfrac{\sigma^2(Z)}{N^2}$

En faisant certaines hypothèses (légèrement restrictives) comme celle selon laquelle les corrélations entre deux individus sont constantes, on peut trouver un estimateur de V(TM(c,E)) (Tassi, 2005.)

Remarque *: Pour certaines cibles, ce terme de corrélation est même nul : c'est le cas des cibles composées d'au plus un individu par foyer comme celles des ménagères ou des chefs de ménage …*

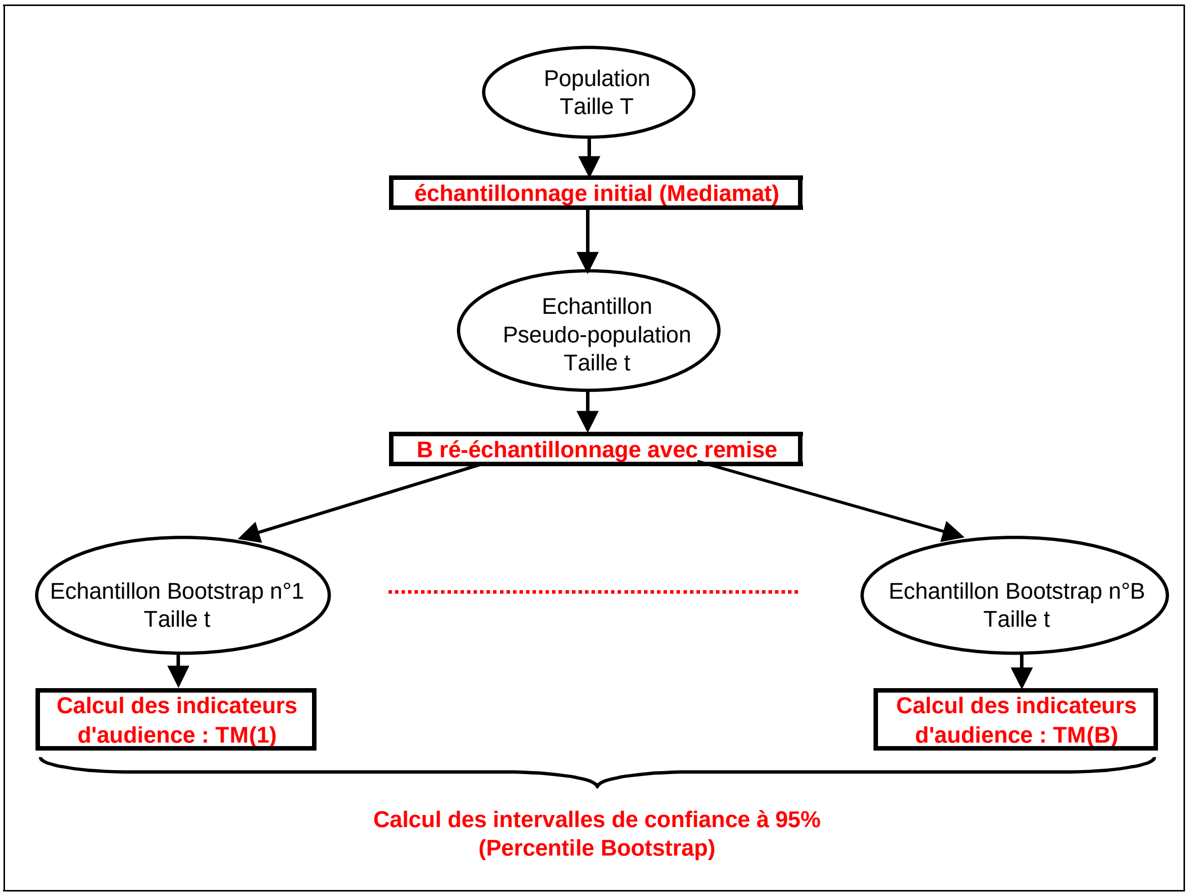
L'application numérique n'en reste pas moins très lourde et difficile à mettre en œuvre (notamment en raison du calculs de toutes les corrélations.) De plus, nous ne pouvons pas implémenter ce type de calcul dans les logiciels de restitutions des résultats car ceux-ci seraient beaucoup trop coûteux en temps.

3. METHODE EMPIRIQUE

3.1 Description de la méthode

La méthode du Bootstrap, technique de ré-échantillonnage introduite par Efron (Efron, 1979), permet de pallier les défauts de la méthode précédente. On s'intéresse à une caractéristique numérique $T(F)$ qu'on estime par $\hat{T}_n(X_1,...,X_n)$ sur un échantillon de taille n et on désire avoir une idée des performances de $\hat{T}_n$ en tant qu'estimateur de T. L'idée de base du Bootstrap part du principe suivant : si c'était possible, il suffirait de tirer un grand nombre d'échantillons de taille n dans la loi de F. En considérant que notre échantillon initial est une bonne représentation de la loi de F, il suffit de répliquer l'échantillon primitif un grand nombre de fois, en effectuant B tirages avec remise parmi les observations qui le constituent, afin d'utiliser le fait que la répartition empirique générée converge uniformément vers la « vraie » répartition. On obtient alors la meilleure approximation non paramétrique possible de la statistique étudiée.
La technique du Bootstrap est relativement récente et fut conçue à l'origine dans le cas d'observations i.i.d. Son adaptation à des plans de sondage complexes, et en particulier les enquêtes par quotas, est encore mal résolue et peu documentée.

Dans le cadre de l'application du Bootstrap à notre panel, nous avons privilégié la simplicité de la mise en œuvre en utilisant une méthode de ré-échantillonnage non conforme au tirage de l'échantillon de base, à savoir un tirage aléatoire simple avec remise (cf. Graphique n°1.) Ce tirage est effectué au niveau foyer afin de conserver le grappage initial.
Dans un deuxième temps, on tient compte, pour le calcul des indicateurs d'audience, des quotas établis en intégrant dans le calcul les poids du redressement. Il est à noter cependant que les poids du redressement initial sont appliqués à chaque boucle du Bootstrap. On ne refait pas de nouveau redressement à chaque boucle du Bootstrap car ceci s'avérerait trop coûteux en temps de calcul.



Graphique n°1 : Schéma de mise en œuvre du Bootstrap

3.2 Résultats

Les calculs ont été faits sur 97 plages horaires, 45 cibles, 9 chaînes (dont le TTV) et ce pour 7 jours (1 semaine.)
On a donc 97 × 45 × 9 × 7 = 274 995 valeurs de TM et les intervalles de confiance obtenus par Bootstrap associés.

Toutes ces données nous ont permis d'étudier les facteurs explicatifs des valeurs des bornes de l'intervalle de confiance. Les conclusions sont les suivantes :

- les bornes ne dépendent pas du jour étudié,
- les bornes ne dépendent pas de la chaîne étudiée,
- les bornes dépendent de la cible étudiée mais uniquement au travers de la taille de cette cible,
- les bornes dépendent naturellement de la valeur du TM.

Nous avons donc ensuite essayé de modéliser les bornes de l'intervalle de confiance (notées Borne sup et Borne inf) à partir des deux variables explicatives Taux Moyen (noté TM) et taille de la cible (notée n_{cible}.) Nous avons testé plusieurs modèles pour finalement retenir le modèle suivant :

$$(5)\quad \text{Borne sup} = TM + A \quad \text{avec} \quad \log A = a_1 \times \log \frac{TM}{\sqrt{n_{cible}}} + a_2$$

$$(6)\quad \text{Borne inf} = TM - B \quad \text{avec} \quad \log B = b_1 \times \log \frac{TM}{\sqrt{n_{cible}}} + b_2$$

Les estimations de a_1, a_2, b_1 et b_2 ont été obtenues en faisant une régression linéaire qui donne des résultats très satisfaisants puisque dans les deux cas (pour la borne supérieure et pour la borne inférieure) on obtient un coefficient de détermination important ($R^2 > 0{,}95$.)

Finalement on obtient donc :

$$(7)\ \text{Borne sup} = TM + A \ \text{ avec } \log A = 0{,}75 \times \log \frac{TM}{\sqrt{n_{cible}}} + 2{,}13$$

$$(8)\ \text{Borne inf} = TM - B \ \text{ avec } \log B = 0{,}83 \times \log \frac{TM}{\sqrt{n_{cible}}} + 2{,}11$$

Les expressions obtenues par cette méthode pour les bornes de l'intervalle de confiance du Taux Moyen sont simples et peuvent donc être aisément implémentées dans les logiciels de restitution des résultats mis à disposition des clients.

RÉFÉRENCES

Bertail, P., et P. Combris (1997), « Bootstrap généralisé d'un sondage », *Annales d'économie et de statistique*, 46.

Bickel, P.J., and D.A. Freedman (1981), « Some asymptotic theory for the bootstrap », *The Annals of Statistics*, 9, p. 1196-1217.

Bickel, P.J., and D.A. Freedman (1984), « Asymptotic normality and the bootstrap in stratified sampling », *The Annals of Statistics*, 12, p. 470-482.

Deville, J.-C. (1987), « Réplication d'échantillons : demi-échantillons, jackknife et bootstrap », in J.-J. Droesbeke, B. Fichet et P. Tassi (éds), *Les sondages*, Paris : Economica, p.147-171.

Diaconis, P., and B. Efron (1983), « Computer-Intensive Methods in Statistics », *Scientific American*, May, p. 116-130.

Efron, B. (1979), « Bootstrap methods : another look at the Jackknife », *The Annals of Statistics*, 7, p. 1-26.

Efron, B., and R. Tibshirani (1986), « Bootstrap Methods for Standard Errors, Confidence Intervals, and Other Measures of Statistical Accuracy », *Statistical Science*, 1, p. 54-77.

Hauvespre, C. (2004), « Estimation d'intervalles de confiance des indicateurs d'audience par Bootstrap », rapport de stage non publié, Paris, France : Médiamétrie.

Shao, J., and D. Tu (1996), *The jackknife and bootstrap*, New York : Springer.

Simonart, S., P. Tassi, et N. Verdurmen (1998), « Précision du taux d'audience d'une émission et du GRP d'une campagne publicitaire télévisée », *Recherches et Applications en Marketing*, 13-4, p. 41-52.

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

Vanheuverzwyn, A. (1999), « Estimation des intervalles de confiance des indicateurs d'audience de la radio », rapport de stage non publié, Paris, France : Médiamétrie.