Statistique et Médias: la partie immergée de l'iceberg

Philippe Tassi

Journées ENS - ENSAI

Rennes, 28 septembre 2007

mediametrie

Les médias

A priori, un objectif simple en terme de mesure :

Compter et qualifier

- Combien de téléspectateurs, d'auditeurs, d'internautes, de cinéphiles, …. ?

- Qui sont – ils ? (profil socio-démographique ou autre)

mediametrie

Les médias : pas seulement des paillettes et du strass

Des populations à étudier en évolution rapide et non homogène :

- offre
- équipements
- réception/initialisation, …

L'importance des enjeux entraîne une exigence forte en qualité de la part des clients (échantillonnage, précision des résultats)

Des interrogations quotidiennes

Une pédagogie permanente

- Il existe de nombreuses questions pas très bien posées ou résolues (exemple : précision d'indicateurs non linéaires construits sur des panels avec grappage)

mediametrie

Les médias : pas seulement des paillettes et du strass

Un secteur en constante complexification : des voies d'accès qui se mélangent, du mono-média au cross-médias, des consommateurs et des écrans mobiles ….

Des comportements individuels ou collectifs qui deviennent compliqués

Des indicateurs d'audience différents

Et, même si cela n'est pas du domaine d'aujourd'hui, des instruments automatiques de mesure ou d'identification des médias/supports regardés/écoutés qui reposent sur des technologies et de l'électronique en évolution permanente

mediametrie

Une vraie R & D

- Publications (colloques à comité scientifique, revues à comité de lecture)

1987 – 89 : 4

1990 – 94 : 21

1995 – 99 : 19

2000 – 04 : 23

2005 et 06 : 13

- 5 brevets déposés et acceptés, 6 dépôts en cours

mediametrie

Les centres d'intérêt

Méthodes de sondage et applications

sondages complexes
quotas, probabilités égales ou inégales,
tirage d'un échantillon équilibré à partir d'un échantillon à probabilités inégales

Redressements

étude chronologique des coefficients de pondération dans un panel,
effet de l'ordre des variables de redressement
Procédure Calmar, applications et améliorations

Recueil

effet « colonne » en déclaratif
mélange de modes de recueil

Enquêtes téléphoniques

listes rouges,
exclusifs du téléphone mobile,
les abandons en cours d'enquête (modèles de survie)
l'échantillon des répondants

mediametrie

Les centres d'intérêt

Précision d'indicateurs

indicateurs d'audience radio, TV
indicateurs publicitaires

Analyse des comportements du public (description et modélisation)

audience des films à la télévision
pré-programmation des stations de radio et comportement d'écoute
auditeurs mono-session ou multi-sessions
les jeunes, les cadres, …

Suivi Qualité (ISO) et processus de production

Mémorisation de la publicité TV et contexte d'insertion

mediametrie

Les centres d'intérêt

Modèles de probabilisation

Prévision

estimation des audiences TV à partir d'intentions de vision du public
processus autorégressifs hilbertiens

Rapprochement et Fusion de fichiers

Données user et site sur Internet
Presse et Internet, TV et Internet
Pluri-médias
« Unité de compte » commune

Signal

Marquage

mediametrie

Quatre illustrations

1 – Sondages

2 – Précision

3 – Modèles de Probabilisation

4 - Comportement

mediametrie

1 - Téléphone mobile et enquêtes par téléphone

Usage du téléphone dans les enquêtes par sondage :

- existence d'un annuaire « base de sondage »
- génération aléatoire de numéros pour reconstituer les ménages en liste rouge
- simplicité des contrôles,
- homogénéité de la passation du questionnaire,

peu de difficulté d'accès aux personnes ou foyers enquêtés,

- dispersion géographique maximale
- maîtrise des coûts et donc possibilité d'un échantillon plus important

mediametrie

Téléphone mobile et enquêtes par téléphone

Mais :

- augmentation constante des foyers « exclusifs du téléphone mobile »

- ces foyers ou individus ont un profil socio-démographique très atypique de ces foyers (jeunes, CSP-, foyers de 1 ou 2 personnes)

- leur comportement d'audience est très différent (même quand on corrige les effets de structure socio-démographique de cette population)

⇒ Impérieuse nécessité d'intégrer cette strate de population dans les enquêtes d'audience

mediametrie

Téléphone mobile et enquêtes par téléphone

Les difficultés:

- Mobilité individuelle et qualité de la communication
- Disponibilité de l'interlocuteur
- Pas d'annuaire, mais connaissance des plages 06XY octroyées par l'ARCEP aux opérateurs

⇒ Possibilité de faire du sondage aléatoire simple dans un ensemble de numéros pas tous attribués par un opérateur

mediametrie

Téléphone mobile et enquêtes par téléphone

Programme de test réalisé à partir de 1998:

- faisabilité
- acceptation par les possesseurs
- acceptation de rappel ou de rdv
- analyse des abandons en cours d'interview ⇒ estimation de la durée maximale d'entretien
- premières estimations des comportements d'audience TV et radio
- constitution préalable d'une base de sondage qualifiée (identification des exclusifs du mobile, données socio-démographiques, accord pour participer à une étude)

mediametrie

Téléphone mobile et enquêtes par téléphone

Conclusions:

- Intégration des exclusifs du téléphone mobile dans les enquêtes de référence radio dès 2003

- Intégration dans le panel TV Médiamat en 2004, avec remontée des informations d'audience par un modem externe de type GSM (mais rotation plus forte car population moins « fixée », moins sédentaire)

mediametrie

2 - Précision d'un indicateur : part d'audience

Exemple de la TV :

- Emission E, de durée S, échantillon de taille n, D(i) = durée de vision de E, Z(i) = D(i)/S

- DEI(E) = Σ_i D(i)/n
- Taux moyen d'audience de E : A(E) = DEI(E)/S = Σ_i Z(i)/n
- Part d'audience de E : PA(E) = DEI(E) / DEI(TV)
= A(E)/A(TV)

- Définis sur un segment C de population (« cible »)

mediametrie

Précision d'un indicateur : part d'audience

- Panel d'individus par grappage au sein d'un échantillon de ménages, avec probabilités inégales
- ⇒ corrélations entre individus au sein d'un même ménage et entre deux dates t et t' pour un même individu

- Etablissement d'une formule théorique :

$$V(PA(E)) = V(A(E))/E^2(A(TV)) + E^2(A(E))V(A(TV))/E^4(A(TV)) - 2E(A(E))Cov(A(E), A(TV))/E^3(A(TV))$$

- Calculs « à la volée » complexes en approche analytique
- ⇒ détermination d'abaques par une procédure en 2 étapes

mediametrie

Précision d'un indicateur : part d'audience

Etape 1 : bootstrap adapté au plan de sondage du panel Médiamat (Efron, Deville, Cho & Lo, …)

Etape 2 : modélisation des bornes de l'intervalle de confiance de la part d'audience

Ainsi, pour une émission E de part d'audience PA(E), sur une cible C de taille n(C), ayant un taux moyen d'audience A(E) alors que le TotalTV a une audience A(TV), on établit pour les bornes de l'intervalle :

$$B = k\, A(TV)^{\alpha}\, A(E)^{\beta}\, n(C)^{\gamma}$$

Le R^2 de ces modèles est de l'ordre de 92 – 95 %

L'estimation de ces modèles permet de construire des abaques facilement utilisables pas des non spécialistes

mediametrie

3 - Probabilisation

Radio

Cohabitation de deux dispositifs d'audience

- ***Etude 126000*** : recueil par sondage téléphonique de l'écoute au cours des 24 dernières heures ; fournit les audiences par station, par cible, par quart d'heure, pour un jour moyen d'une période (trimestre ou bimestre) : référence en connaissance du comportement moyen mais pas de « profondeur » chronologique

- ***Panel sur 23 jours*** : carnet d'écoute auto-administré sur 23 jours (15 jours « de semaine » lundi-vendredi, 4 week-end), en deux vagues (janvier, septembre) ; permet d'introduire la dimension temporelle et donc d'approcher les « habitudes » de contact (accumulation et duplications d'audience dans le temps)

mediametrie

Probabilisation

Médiaplanning

Campagne publicitaire de K insertions d'un spot

X est la v.a. dénombrant les contacts avec la campagne

Loi de X : *distribution des contacts*

Indicateurs de performance d'un plan de communication:

E(X) = GRP

P(X > 0) = Couverture

E(X/X>0) = Répétition

GRP = Couverture . Répétition

mediametrie

Probabilisation

Médiaplanning

L'estimation des performances du plan est faite a priori (prévision), à partir de probabilités de contact d'un individu avec chaque station de radio et pour chaque quart d'heure.

La 126000 et le Panel ne permettent pas une bonne estimation de ces probabilités (pas de recul temporel pour la 126000, effet discrétisation pour le panel)

⇒ Nécessité de disposer d'une méthode d'estimation mêlant les deux sources d'information

mediametrie

Probabilisation

Probabilisation

- Qu'est-ce qu'un fichier d'audiences probabilisées ?

C'est un échantillon d'individus caractérisés, d'une part, par des informations sociodémographiques et, d'autre part, par un vecteur de probabilités de contact avec un support (on appelle support un quart d'heure particulier pour une station radio donnée)

- L'objectif de la probabilisation est d'affecter ces probabilités individuelles de contact avec chaque support de façon à être le plus proche possible des données effectivement mesurées.

Individu i → P(i écoute s en q), ∀ la station s, ∀ le quart d'heure q

mediametrie

Probabilisation

Objectif

- Prendre en compte de façon conjointe les informations issues des deux sources de données
- Redresser les probabilités en cas de nouveaux résultats 126000 entre deux panels (en particulier s'il y a eu des modifications significatives du paysage radio)

Principe

- La répartition effective des probabilités d'audience sur une population donnée est modélisée par un modèle de probabilités dit « bêta-binomial » (Chandon, 1976), avec loi bêta pour le paramètre sur]0, 1[et segments en 0 et 1.
- Cette modélisation supprime l'effet de discrétisation et permet de lisser les fluctuations d'échantillonnage, car elle ne nécessite que l'estimation d'un petit nombre de paramètres.
- Elle permet un ajustement sur les niveaux moyens 126000 (GRP) tout au long de la probabilisation et fournit une méthode de réajustement lors des nouveaux résultats 126000.

Probabilisation

On fait l'hypothèse que, pour un support donné, sur une cible donnée, les probabilités individuelles sont réparties continûment et sont modélisées par une loi Bêta avec masses en 0 et 1:

$$z\delta_0 + u\delta_1 + \frac{(1 - z - u)\, t^{a-1}(1-t)^{b-1}}{\int_0^1 \xi^{a-1}(1-\xi)^{b-1} d\xi}$$

On re-paramétre la loi bêta (Stehlé, 1998) en posant :

$$\tau = 1/(a + b) \qquad p = a\,\tau$$

mediametrie

Probabilisation

Cela revient, pour chaque station et pour chaque quart d'heure, à découper la cible en trois catégories d'auditeurs :

Les « jamais »:

- probabilité d'écoute = 0, en proportion **z.**

Ce segment correspond aux individus qui n'auront aucun contact avec le support.

La quantité 1 – z représente l'audience cumulée totale du support, c'est-à-dire la proportion d'individus de la cible que le support est susceptible de toucher, au moins une fois, au cours du temps (asymptote de la courbe de montée en audience).

Les « toujours » :

- probabilité d'écoute = 1, en proportion **u**.

Les auditeurs « probables » (ou « occasionnels »)

Leur probabilité d'écoute suit une loi Bêta de paramètres p et τ, d'espérance p et de variance $p(1 – p)\tau/(1 + \tau)$; ils sont en proportion (1 - z - u).

Sens de p et τ : p est l'audience moyenne des auditeurs probables, et la dispersion des probabilités individuelles autour de p est une fonction croissante de τ.

Probabilisation

Ajustement des paramètres

On cherche à estimer les quatre paramètres (z, u, p, τ) de façon que le modèle théorique soit le plus proche possible des données observées sur le panel.

La distance entre modèle théorique et résultats observés du panel est donnée par :

$$\text{Distance} = \sum_{j=1}^{t} j\,(MeaTh[j] - MeaObs[j])^2 \;+\; k\frac{(DdcObs[0]-z)^2}{DdcObs[0]} \;+\; k\frac{(DdcObs[t]-u)^2}{DdcObs[t]}$$

Où :

MeaTh[j] désigne, dans le modèle théorique, la valeur de la courbe de montée en audience après *j* jours.

DdcObs[j] désigne la proportion, observée dans le panel, de panélistes touchés *j* fois par le support.

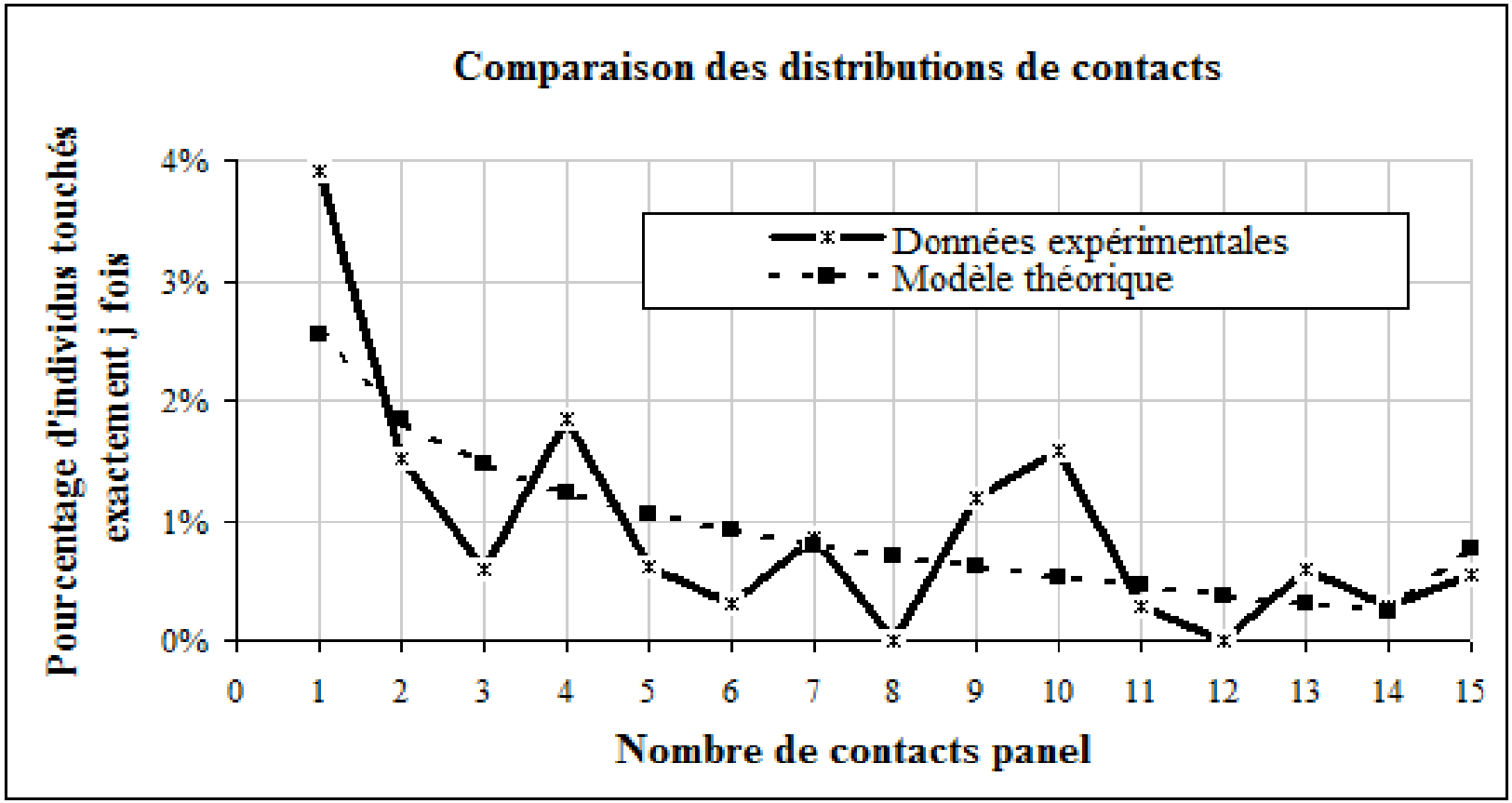
MeaObs[j] désigne la valeur de la courbe de montée en audience, observée panel, après *j* jours.

t est le nombre d'observations, 15 pour l'univers Lundi–Vendredi, (3 semaines de 5 jours) et 4 pour le samedi et pour le dimanche (4 week-end dans le panel).

k est un coefficient de régularisation (fixé à 0.1).

mediametrie

Probabilisation



mediametrie

Probabilisation

En pratique :

Phase 1 : détermination des coefficients (z, u, p, τ) par minimisation de la distance modèle/observations pour chaque support, univers de temps, (lundi-vendredi, samedi, dimanche) et cellule de population (selon des critères socio-démographiques explicatifs du comportement radio)

Phase 2 : hiérarchisation des individus

A qui attribuer la probabilité 0 %, 1 %, etc … ?

Cela nécessite de « séparer » les individus du panel et les classer selon une probabilité d'audience croissante

Critères : nombre de contacts observés dans le panel, déclarations d'habitudes dans la tranche horaire, utilisation d'une distance socio-démographique entre individus

Phase 3 : affectation des probabilités individuelles

mediametrie

Probabilisation

Remarque sur la prise en compte des GRP 126000

Les GRP donnés par le modèle précédent sont ceux du panel.

En notant g l'audience d'un support, on a la relation :

$$g = u + (1 - z - u)\, p$$

L'ajustement du modèle pour prendre en compte les GRP 126000 revient tout simplement à modifier p (donc l'audience moyenne des occasionnels) sans toucher z et u, de façon à obtenir l'audience désirée.

mediametrie

4 - Mémorisation et contexte d'insertion de la publicité TV

La problématique

- Contexte : travail de thèse

- Thème : insertion d'écrans publicitaires au sein des émissions et programme de recherche sur l'impact de l'implication dans un programme sur la mémorisation

- Mesure de variables d'implication, d'intensité d'attention, des émotions, de leur intensité et de leur qualité

(⇒ problématique des échelles de mesure)

- Données de deux types
    - en laboratoire
    - en situation réelle

mediametrie

Mémorisation et contexte d'insertion de la publicité TV

Analyse de la littérature ⇒ Formulation d'hypothèses sur les liens entre les variables

Exemples:

- niveau d'attention ↑, mémorisation ↓
- intensité des émotions ↑, mémorisation ↓
- qualité des émotions ↑, mémorisation ↑
- implication dans le genre ↑, mémorisation ↓
- implication dans le genre ↑, attention au programme ↑

mediametrie

Mémorisation et contexte d'insertion de la publicité TV

- Modèle d'équations structurelles

(traitement par la méthode LISREL – cf Stan et Saporta sur la comparaison des approches PLS et LISREL)

Résultats majeurs :

- effet généralement négatif de l'attention sur la mémorisation publicitaire
- effet généralement négatif de l'intensité des émotions sur la mémorisation, en particulier pour les genres séries, films policiers et émissions sportives

mediametrie

En conclusion

Un secteur passionnant,

ancré dans les comportements quotidiens,

riche et diversifié en problématiques scientifiques,

trop peu investi par les mathématiciens,

et qui devient – et deviendra encore – de plus en plus complexe

mediametrie