

STA108 Enquêtes et sondages. Introduction

Gilbert Saporta

CNAM, octobre 2013

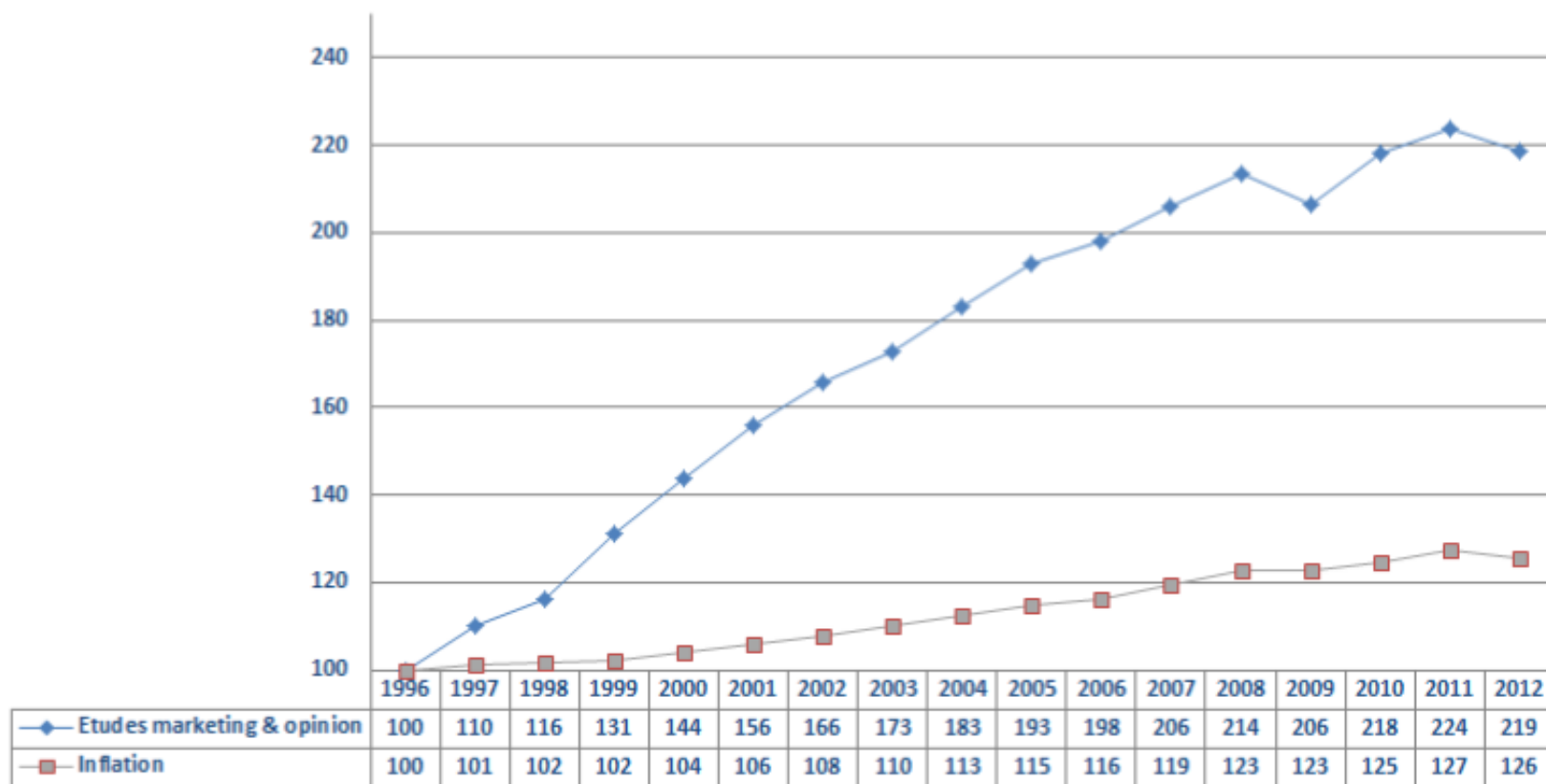
Introduction: aperçu du secteur

- La statistique publique: 8000 employés dont 5800 à l'INSEE
- Une organisation ternaire:
 - Le **Conseil national de l'information statistique** (Cnis) assure en amont la concertation entre ses producteurs et ses utilisateurs.
 - Le **service statistique public** (Insee et services statistiques ministériels) est le moteur dans sa conception, sa production et sa diffusion.
 - L'**Autorité de la statistique publique** veille au respect des principes d'indépendance professionnelle, d'impartialité, d'objectivité, de pertinence et de qualité dans son élaboration et sa diffusion.

un secteur privé qui ne connaît pas la crise

- Près de **400 instituts d'étude de marché et d'opinion** identifiés en France
- Marché estimé de **2.1 milliards d'euros en 2011**
- **Environ 12 000 personnes**, hors enquêteurs

TAUX DE CROISSANCE DU MARCHE DES ETUDES MARKETING & OPINION, 1996-2012 (BASE 100 EN 1996)



- L'opinion: une faible part de l'activité des instituts

Effectifs et chiffres d'affaires des principaux instituts de sondages

Instituts de sondages	Nombre de salariés (au 31/12/2009)	Chiffre d'affaires en 2009 (en millions d'€)	Part des sondages politiques dans le chiffre d'affaires* (en %)
BVA	230	53	1 %
CSA	106	32	16 %
IFOP	159 (au 30/04/10)	35,2	20 à 25 %
IPSOS	574	98,7	1 %
LH2	95	19	3 à 5 %
Opinion Way	45	9,1	6 %
TNS-Sofres	559** (en 2006)	126** (au 31/12/2008)	NC
Viavoice	NC	0,8**	25 %

** données communiquées à vos rapporteurs par les instituts*

*** source Infogreffe*

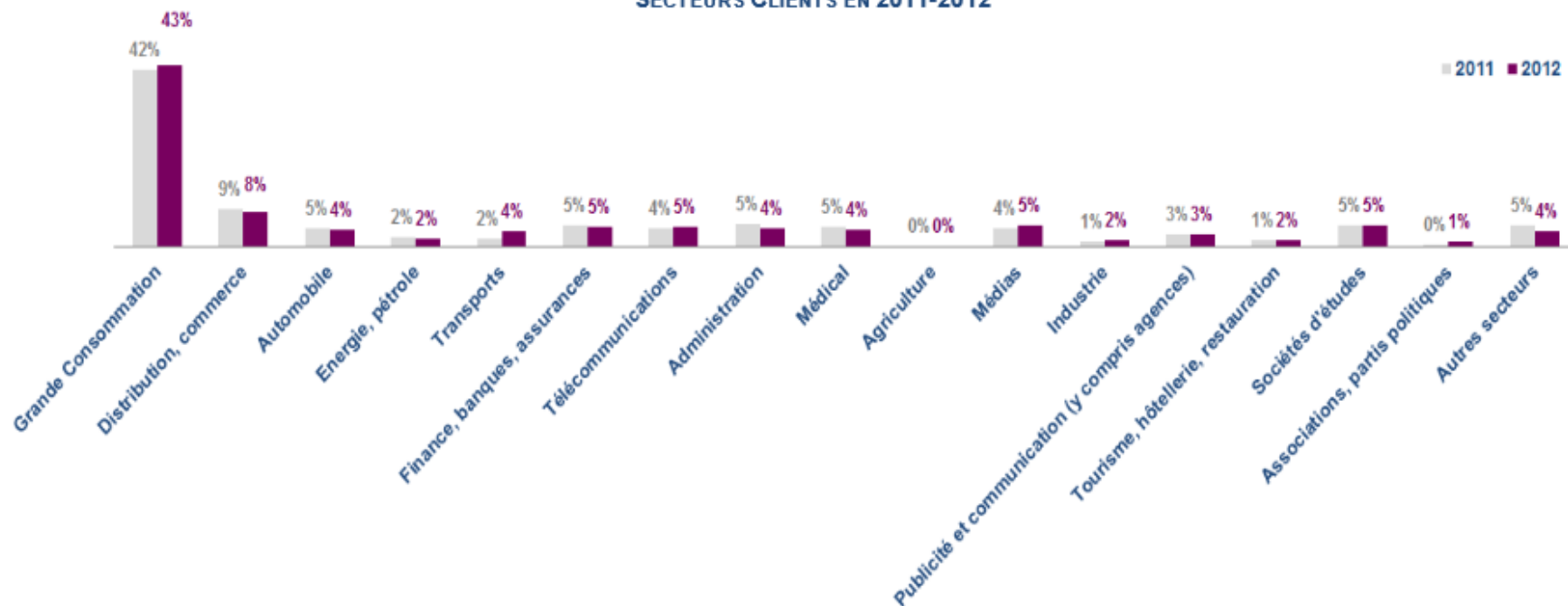
NC = non communiqué à vos rapporteurs

Source: rapport Portelli-Sueur, Sénat, 2011:
<http://www.senat.fr/rap/r10-054/r10-0541.pdf>

Secteurs Clients : la grande consommation reste le principal secteur Client

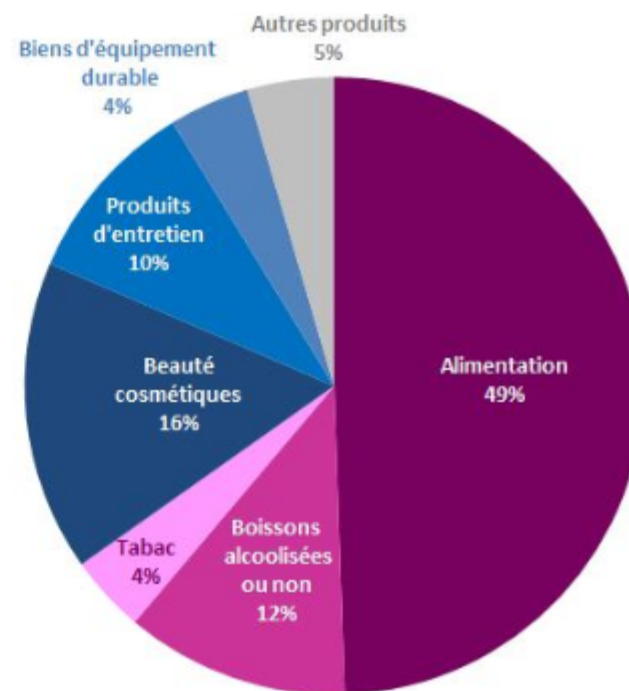
La grande consommation affiche 43% de la demande totale. Les autres secteurs d'activités présentent une répartition stable entre 2011 et 2012, la variation la plus significative (-1,6%) revenant aux Transports (personne et fret).

SECTEURS CLIENTS EN 2011-2012



L'alimentation, le tabac, les **boissons alcoolisées et non alcoolisées**, constituent 65% du marché de la **grande consommation**.

MARCHES DE LA GRANDE CONSOMMATION EN 2012



Une histoire récente

- **1895** – Kiaer, dénombrements représentatifs
- **1925** – Jensen,
- **1934** – Neyman: la théorie
- **1936** – Election de Roosevelt
- **1938** – Fondation de l'IFOP par J.Stoetzel
- **1952** – Horvitz et Thompson, Sondages à probabilités inégales
- **1965** – Ballottage De Gaulle

Galerie



Anders Nikolai Kiaer
1838- 1919



Jerzy Neyman
1894-1981



Jean Stoetzel
1910-1987

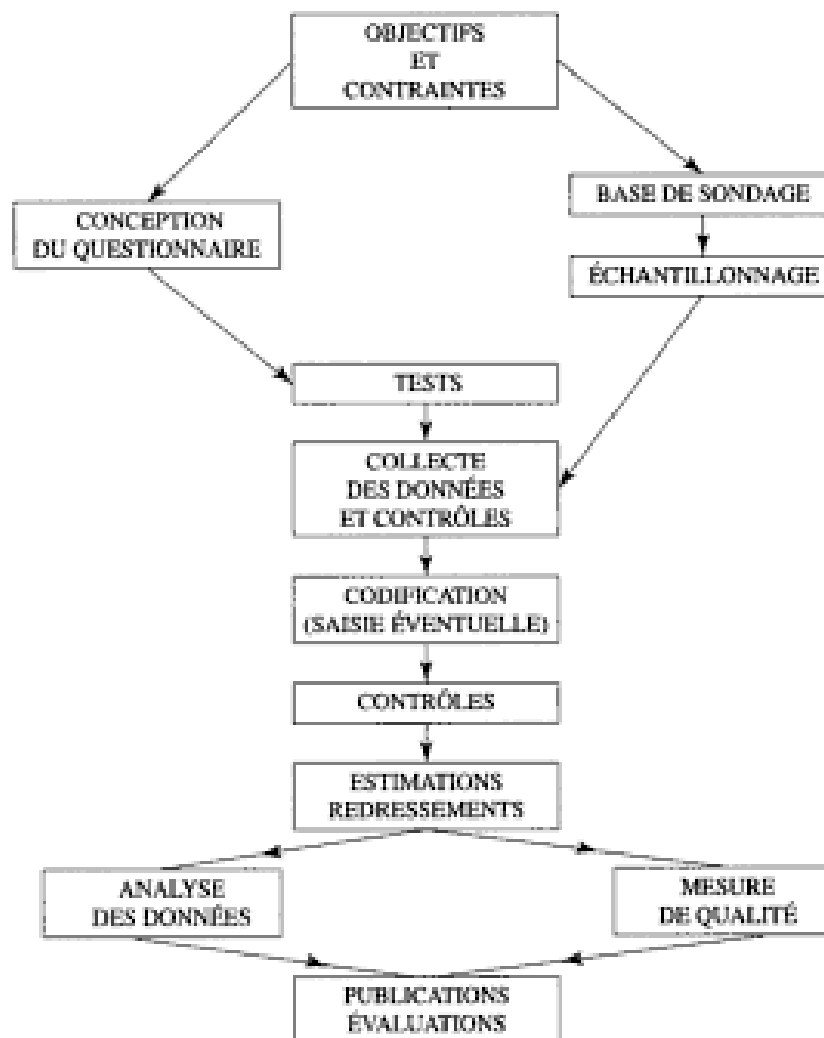
Un contre-exemple célèbre



LES TECHNIQUES DE SONDAGE

- Problèmes essentiels:
 - Sélection de l'échantillon;
 - Agrégation des réponses
 - ✓ estimateur;
 - ✓ précision;

Les principales étapes



source: P.Ardilly

LES TECHNIQUES DE SONDAGE

- Méthodes aléatoires:

Plans de sondage

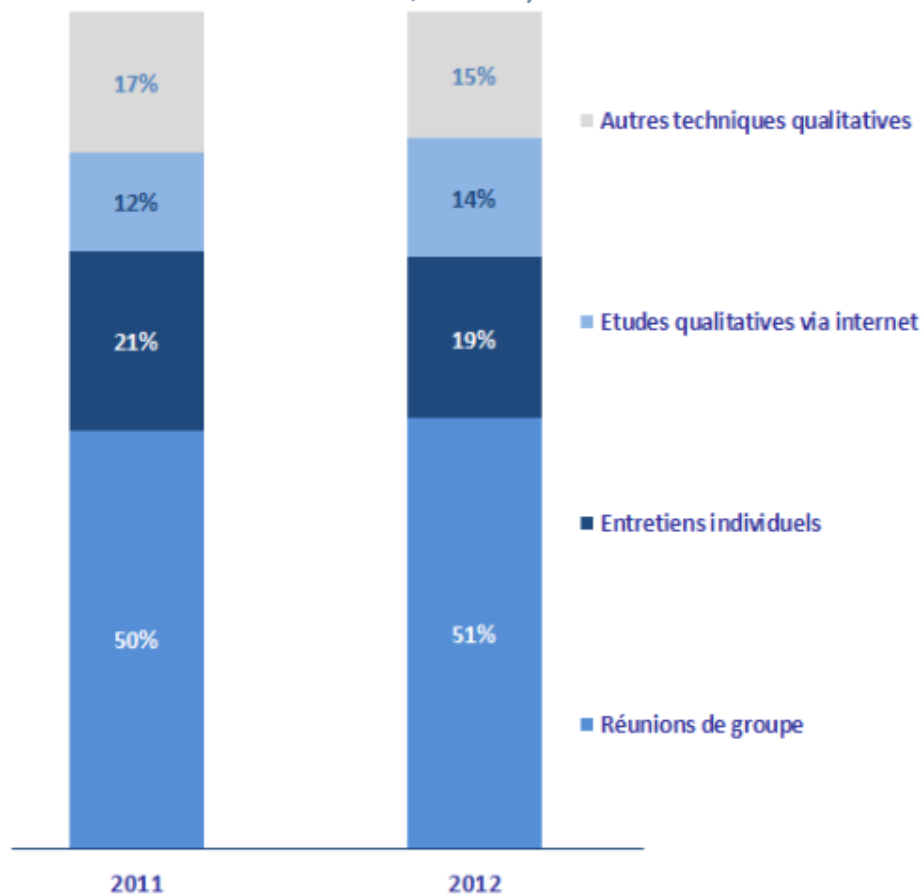
- **Simples**: - à probabilités égales
- à probabilités inégales

- **Complexes**: - stratifié
- en grappe
- plusieurs degrés

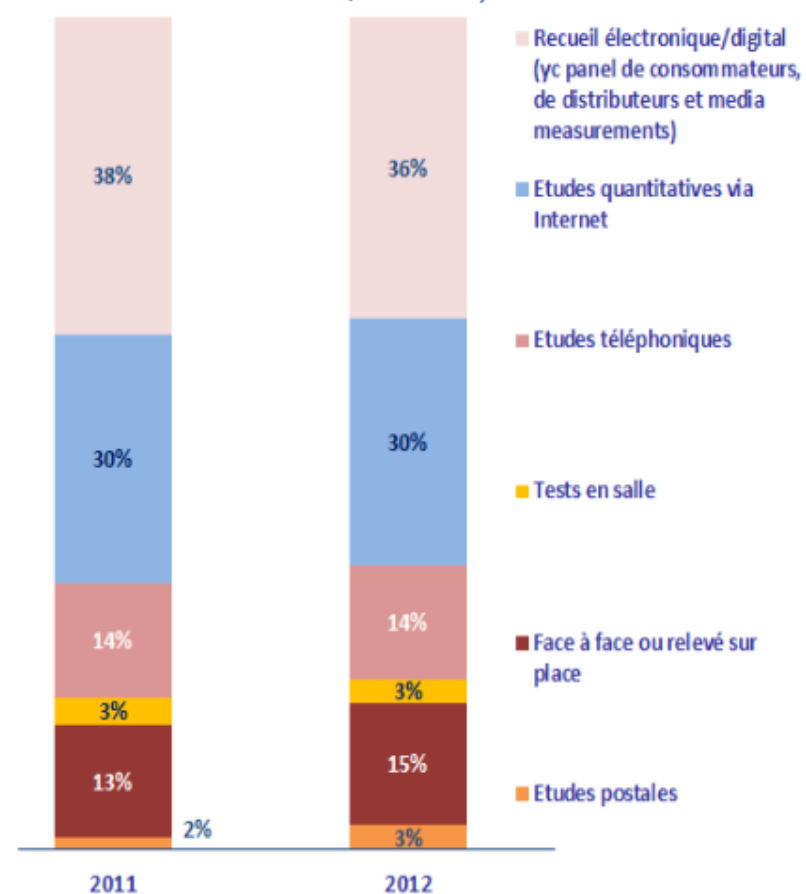
LES TECHNIQUES DE SONDAGE

- Méthodes par choix raisonné ou judicieux:
 - Quotas;
 - Itinéraires;
 - Unités – types;
 - Volontariat;
 - Échantillonnage sur place;

REPARTITION DES METHODES DE RECUEIL ET D'ETUDES
DANS LE CA QUALITATIF, 2011-2012



REPARTITION DES METHODES DE RECUEIL ET D'ETUDES
DANS LE CA QUANTITATIF, 2011-2012



Représentativité

Yves Tillé, *Théorie des sondages*, Dunod, 2001

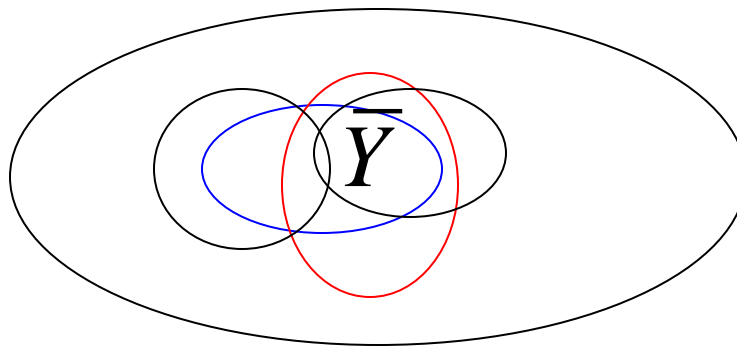
Voir invoquée la "représentativité" dans un rapport d'enquête pour justifier de la qualité d'un sondage peut presque à coup sûr laisser soupçonner que l'étude a été réalisée dans une méconnaissance totale de la théorie de l'échantillonnage. Le concept de représentativité est aujourd'hui à ce point galvaudé qu'il est désormais porteur de nombreuses ambivalences. Cette notion, d'ordre essentiellement intuitif, est non seulement sommaire mais encore fausse et, à bien des égards, invalidée par la théorie. Raison pour laquelle ce terme sera volontairement évité dans cet ouvrage.

Représentativité

- Notion peu scientifique
- Souvent confondue avec le respect de certaines proportions (modèle réduit)
- Un sondage à probabilités inégales , un sondage stratifié ou à plusieurs degrés peuvent être représentatifs en un autre sens:
 - **Sondage extrapolable** : probabilités d'inclusion connues et non nulles

Fluctuations et biais

- Fluctuations d'échantillonnage : avec les mêmes probabilités d'inclusion, répéter q fois un sondage donnera q résultats différents



$$\bar{y}_1, \bar{y}_2, \dots, \bar{y}_q$$

- Sans biais: si la moyenne des moyennes de tous les échantillons possibles est égale à la moyenne de la population (pas d'écart systématique)

SONDAGE ALEATOIRE SIMPLE

■ Notations:

- Population ou base de sondage: **N**
- Identifiant: **i**
- Variable d'intérêt: **Y** (Y_1, Y_2, \dots, Y_N)

$$\bar{Y} = \frac{1}{N} \sum_{i=1}^N Y_i; \quad T = \sum_{i=1}^N Y_i$$

$$\sigma^2 = \frac{1}{N} \sum_{i=1}^N (Y_i - \bar{Y})^2; \quad S^2 = \frac{1}{N-1} \sum_{i=1}^N (Y_i - \bar{Y})^2 = \frac{N}{N-1} \sigma^2$$

SONDAGE ALÉATOIRE SIMPLE

- Définition: tirage équiprobable sans remise de n unités;
- Taux de sondage: $\frac{n}{N} = \tau$
- C_N^n échantillons possibles;
- π_i probabilité d'inclusion (plan de taille fixe): $\sum_{i=1}^N \pi_i = n$
- Équiprobabilité: $\pi_i = \frac{n}{N} = \tau$
- Remarque: $\pi_i = \sum_{s \ (i \in s)} p(s)$

SONDAGE ALÉATOIRE SIMPLE

- Estimation du total et de la moyenne:

\bar{y} - estimateur de \bar{Y}

$N\bar{y}$ - estimateur de T

$$E(\bar{y}) = \bar{Y} \quad ; \quad E(N\bar{y}) = T$$

- Démonstration avec les variables de Cornfield

$$\delta_i = \begin{cases} 1 & \text{si } i \in s \\ 0 & \text{si } i \notin s \end{cases} \quad \begin{aligned} E(\delta_i) &= \pi_i \\ V(\delta_i) &= \pi_i(1 - \pi_i) \quad \text{cov}(\delta_i; \delta_j) = \pi_{ij} - \pi_i\pi_j \end{aligned}$$

$$\frac{N}{n} \sum_{i \in s} y_i = \hat{T} = \sum_{i \in s} \frac{y_i}{\pi_i} = \sum_{i=1}^N \frac{Y_i}{\pi_i} \delta_i$$

y_i = variable aléatoire;

Y_i = variable non aléatoire

$$E(\hat{T}) = \sum_{i=1}^N \frac{Y_i}{\pi_i} E(\delta_i) = \sum_{i=1}^N Y_i = T$$

- **Jerome Cornfield (1912-1979)**
plus connu comme biostatisticien
(lien entre tabac et cancer du poumon)



On samples from finite populations
JASA, 39,236-239, 1944

SONDAGE ALEATOIRE SIMPLE

■ Covariance entre variables de Cornfield

$$\text{cov}(\delta_i; \delta_j) = \pi_{ij} - \pi_i \pi_j = \pi_{ij} - \tau^2$$

$$\pi_{ij} = \sum_{s \{i, j \in s\}} p(s) = \frac{C_{N-2}^{n-2}}{C_N^n} = \frac{n(n-1)}{N(N-1)} = \tau \frac{n-1}{N-1}$$

$$\text{cov}(\delta_i; \delta_j) = -\frac{\tau(1-\tau)}{N-1}$$

■ Variance de la moyenne

$$\begin{aligned} V(\bar{y}) &= V\left(\frac{1}{n} \sum_{i=1}^N Y_i \delta_i\right) = \frac{1}{n^2} \left[\sum_{i=1}^N Y_i^2 V(\delta_i) + \sum_{i \neq j} \sum Y_i Y_j \text{cov}(\delta_i; \delta_j) \right] \\ &= \frac{\tau(1-\tau)}{n^2} \left[\sum_{i=1}^N Y_i^2 - \sum_{i \neq j} \sum \frac{Y_i Y_j}{N-1} \right] = \frac{\tau(1-\tau)}{n^2} N S^2 = (1-\tau) \frac{S^2}{n} \end{aligned}$$

SONDAGE ALÉATOIRE SIMPLE

- Variances:

$$V(\bar{y}) = (1 - \tau) \frac{S^2}{n}$$

$$V(\hat{T}) = N^2 (1 - \tau) \frac{S^2}{n}$$

Estimation de S^2 :

$$s^2 = \frac{1}{n-1} \sum_{i \in s} (y_i - \bar{y})^2$$

$$E(s^2) = S^2$$

$$\Rightarrow \begin{cases} \widehat{V(\bar{y})} = (1 - \tau) \frac{s^2}{n} \\ \widehat{V(\hat{T})} = N^2 (1 - \tau) \frac{s^2}{n} \end{cases}$$

SONDAGE ALÉATOIRE SIMPLE

- Intervalles de confiance pour un paramètre d'intérêt (« fourchette »)
 - Intervalle ayant une probabilité $1-\alpha$ (niveau de confiance) de contenir la vraie valeur du paramètre. α risque d'erreur, généralement partagé de façon symétrique $\alpha/2$ et $\alpha/2$
 - Nécessite de connaître au moins approximativement la distribution de probabilité de l'estimateur
 - La longueur de l'intervalle diminue avec n et augmente avec le niveau de confiance et avec la variance de l'estimateur (elle-même fonction de la variance de la population)

Le théorème « central limite »

- La moyenne d'un échantillon de n observations indépendantes issues d'une population de moyenne μ et d'écart-type σ converge si n augmente vers une loi normale:

$$N(\mu; \frac{\sigma}{\sqrt{n}})$$

- Illustration animée:
 - http://www.vias.org/simulations/simusoft_cenliit.html
 - $n > 30$ est souvent suffisant



Cenlimit.exe

Intervalle de confiance théorique pour une moyenne

- Tirages indépendants (avec remise) et $n > 30$

$$\bar{y} - u_{\alpha/2} \frac{\sigma}{\sqrt{n}} < \bar{Y} < \bar{y} + u_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

$$\text{pour } \alpha = 5\% \quad u_{\alpha/2} \simeq 2$$

- Tirages sans remise

- On pourra admettre que:

$$\bar{y} - u_{\alpha/2} \frac{S}{\sqrt{n}} \sqrt{1 - \tau} < \bar{Y} < \bar{y} + u_{\alpha/2} \frac{S}{\sqrt{n}} \sqrt{1 - \tau}$$

- Si le taux de sondage est faible la précision ne dépend pas de N

Intervalles de confiance estimés à 95%

- Pour une moyenne:

$$\bar{y} - 2s\sqrt{\frac{1-\tau}{n}} < \bar{Y} < \bar{y} + 2s\sqrt{\frac{1-\tau}{n}}$$

- Pour un pourcentage:

$\bar{y} = \hat{p}$ fréquence observée

$$\mathbf{Y}_i = \begin{cases} 1 \\ 0 \end{cases} \quad \bar{Y} = p$$

$$V(\hat{p}) = (1-\tau) \frac{p(1-p)}{n} \frac{N}{N-1}$$

$$\hat{V}(\hat{p}) = (1-\tau) \frac{\hat{p}(1-\hat{p})}{n-1} \simeq \frac{\hat{p}(1-\hat{p})}{n} \text{ si } \tau \text{ faible}$$

$$\boxed{\hat{p} - 2\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} < p < \hat{p} + 2\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}}$$

Calculs de taille d'échantillon

- Pour une précision fixée

$$\Delta = 2S\sqrt{\frac{1-\tau}{n}} \quad \text{d'où} \quad n = N \frac{1}{1 + \frac{N\Delta^2}{4S^2}}$$

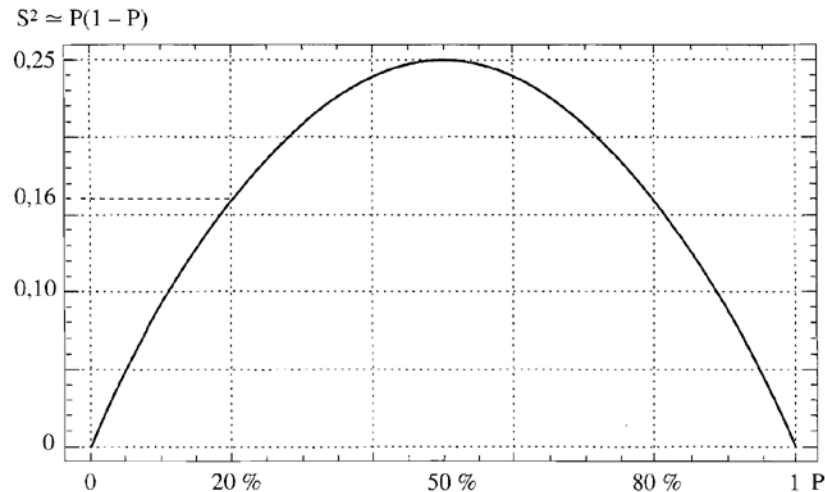
- Nécessite de connaître S !

Pour une proportion

- Si n grand et τ faible

$$\Delta = 2\sqrt{\frac{p(1-p)}{n}} \text{ d'où } n = \frac{4p(1-p)}{\Delta^2}$$

- Utile si on connaît approximativement p a priori



Ardilly, 2006

$\Delta \backslash p$	0,05	0,10	0,20	0,30	0,40	0,50
$\pm 0,005$	7 600	14 400	25 600	33 600	38 400	40 000
$\pm 0,01$	1 900	3 600	6 400	8 400	9 600	10 000
$\pm 0,02$	475	900	1 600	2 100	2 400	2 500
$\pm 0,03$	211	400	711	933	1 066	1 111
$\pm 0,04$	119	225	400	525	600	625
$\pm 0,05$	76	144	256	336	384	400

■ Solution prudente (ou pessimiste)

Se placer dans le cas $p=0.50$

avec $\alpha=0.05$

$$n \simeq \frac{1}{\Delta^2}$$

- Pour τ fort , dans le cas $p=0.50$ avec un niveau de confiance de 95%:

$$n \simeq \frac{N}{1 + N\Delta^2}$$

- Précision absolue ou précision relative?
 - Pour une population rare, on aboutit à une taille d'échantillon souvent excessive
 - Viser un Δ/p change tout
- Compromis à faire quand il y a plusieurs variables d'intérêt
- Attention aux non-réponses: la précision dépend du nombre de répondants