ENQUETES et SONDAGES
STA 108 2010-2011

Intervenants :
G.Saporta (CNAM), O.Marchese (IPSOS), S.Rousseau (INSEE)

<u>Plan du cours:</u>

1	8 octobre	Introduction GS et OM
ED1	11 octobre	Rappels - Sondage aléatoire simple 1 SR **(18h15-20h45)**
2	15 octobre	Sondage aléatoire simple GS
ED2	18 octobre	Sondage aléatoire simple 2 SR **(18h15-20h45)**
3	22 octobre	Sondages à probabilités inégales GS
ED3	25 octobre	Plans à probabilités inégales SR **(18h15-20h45)**
4	29 octobre	Algorithmes de tirage GS **Salle 39.2.64**
5	5 novembre	Stratification GS
ED4	8 novembre	TP simulations de tirage SR **(18h15-20h45)**
6	12 novembre	Sondages à deux degrés et grappes SR
ED6	15 novembre	Plans stratifiés 1 SR
7	19 novembre	Redressement (quotient, régression, post-strates) SR
ED7	22 novembre	Plans stratifiés 2 SR
8	26 novembre	Données manquantes et non-réponses GS et SR
ED8	29 novembre	Plans par grappes SR

9	3 décembre	Effets et pratique des redressements OM
ED9	6 décembre	Plans à plusieurs degrés SR
10	10 décembre	La méthode des quotas OM
ED10	13 décembre	TP correction de la non-réponse SR
11	17 décembre	Les panels GS et OM
12	3 janvier	Le recensement SR
13	7 janvier	Sources d'erreur et biais OM
14	10 janvier	Questionnaires, enquêteurs et enquêtés OM
15	14 janvier	Modes de recueil (avec et sans enquêteur) OM
ED11	17 janvier	Redressement 1 SR
ED12	21 janvier	Redressement 2 SR
ED13	24 janvier	TP redressement SR
ED14	28 janvier	Compléments et révisions SR

Examen 1ère session : 4 février
Examen 2ème session : 15 avril

Un projet pratique est également exigé, la note finale de STA108 sera la moyenne arithmétique équipondérée de la note d'examen (1ère ou 2ème session) et de la note de projet.

Références:

Ouvrages recommandés

- J.ANTOINE Histoire des sondages (Odile Jacob,2005)
- P.ARDILLY Les techniques de sondage, 2ème édition (Technip, 2006)
- P.ARDILLY, Y.TILLE Exercices corrigés de méthodes de sondage (Ellipses, 2003)
- A.M. DUSSAIX, J.M. GROSBRAS Exercices de sondages (Economica, 1992)
- SYNTEC Etudes Marketing et Opinion - Fiabilité des méthodes et bonnes pratiques (Dunod, 2007)
- Y.TILLÉ Théorie des sondages (Dunod, 2001)

Sites internet

- Cours de statistique : http://www.agro-montpellier.fr/cnam-lr/statnet/
- Autorité de la statistique publique http://www.autorite-statistique-publique.fr
- CNIS http://www.cnis.fr/
- INSEE : http://www.insee.fr
- IPSOS: http://www.ipsos.fr/
- Assoc. Intern. Statisticiens d'enquête: http://isi.cbs.nl/iass/allFR.htm
- SYNTEC Etudes http://www.syntec-etudes.com/

INTRODUCTION

- Aperçu du secteur
  - Statistique publique

La statistique publique est gouvernée par 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 est le moteur dans sa conception, sa production et sa diffusion.
Il est composé de l'Insee et des services statistiques ministériels.
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.

INSEE	5 800 employés
Ministères	2 200
Total	8 000

- Secteur privé:

  près de 400 grands groupes et sociétés identifiés **en France** consolident un marché estimé de **1.93 milliards d'euros en 2009**, avec un **effectif total d’environ 12000 personnes**, hors enquêteurs

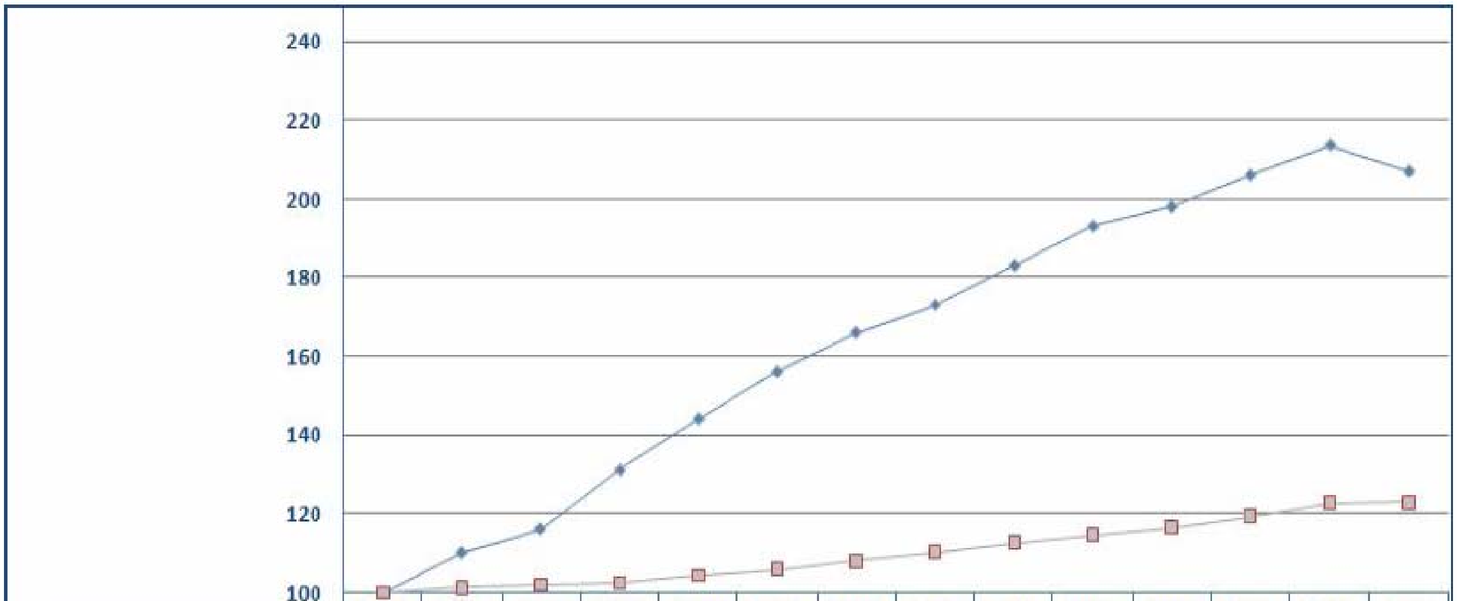
SYNTEC Etudes Marketing & Opinion

LE MARCHE FRANÇAIS DES ETUDES EN 2009

2009 enregistre le 1er recul de son activité depuis 13 ans

Progression du CA des 61 membres

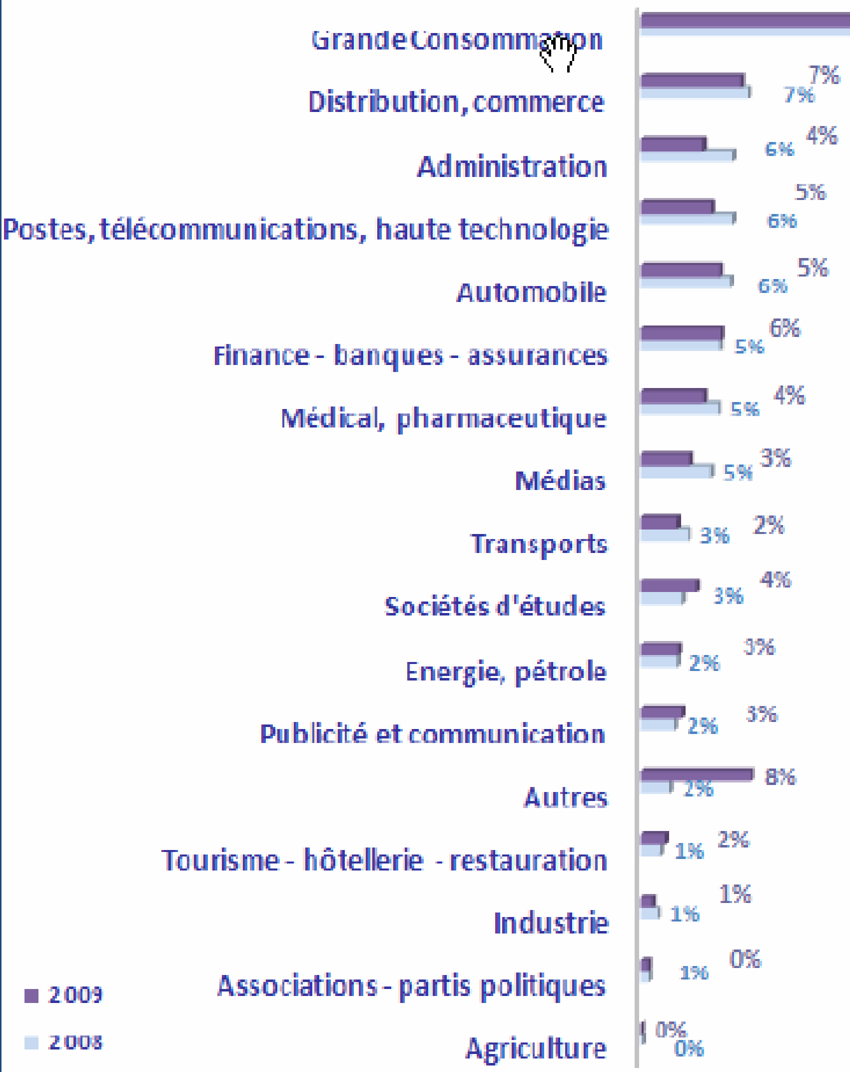
TAUX DE CROISSANCE DU MARCHE DES ETUDES MARKETING & OPINION, 1996-2009



	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009
Etudes marketing & opinion	100	110	116	131	144	156	166	173	183	193	198	206	214	207
Inflation	100	101,2	101,9	102,4	104,2	105,9	108	110,2	112,6	114,6	116,4	119,4	122,7	122,8

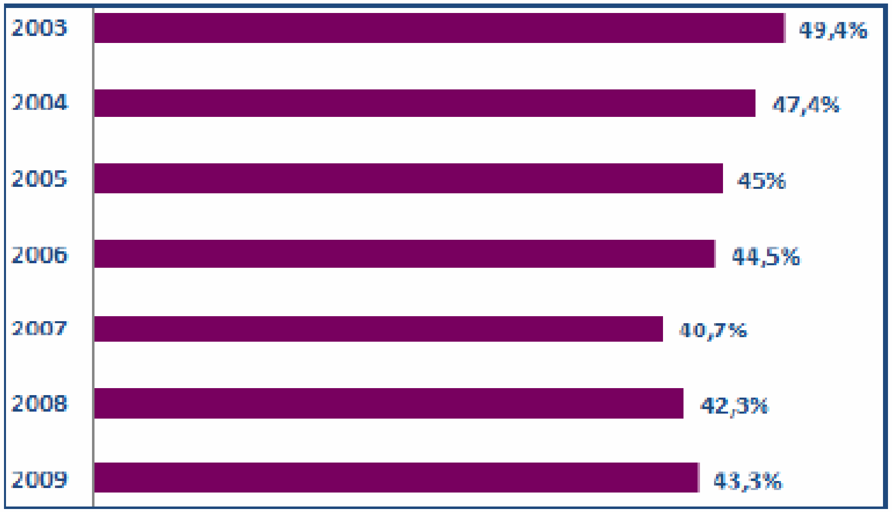
Secteur clients : la grande consommation confirme sa légère reprise engagée en 2008

LA REPARTITION SECTORIELLE, 2008-2009



L'activité générée par le principal secteur client, la **grande consommation**, inverse la tendance de ces dernières années et gagne 2,6 points depuis 2007.

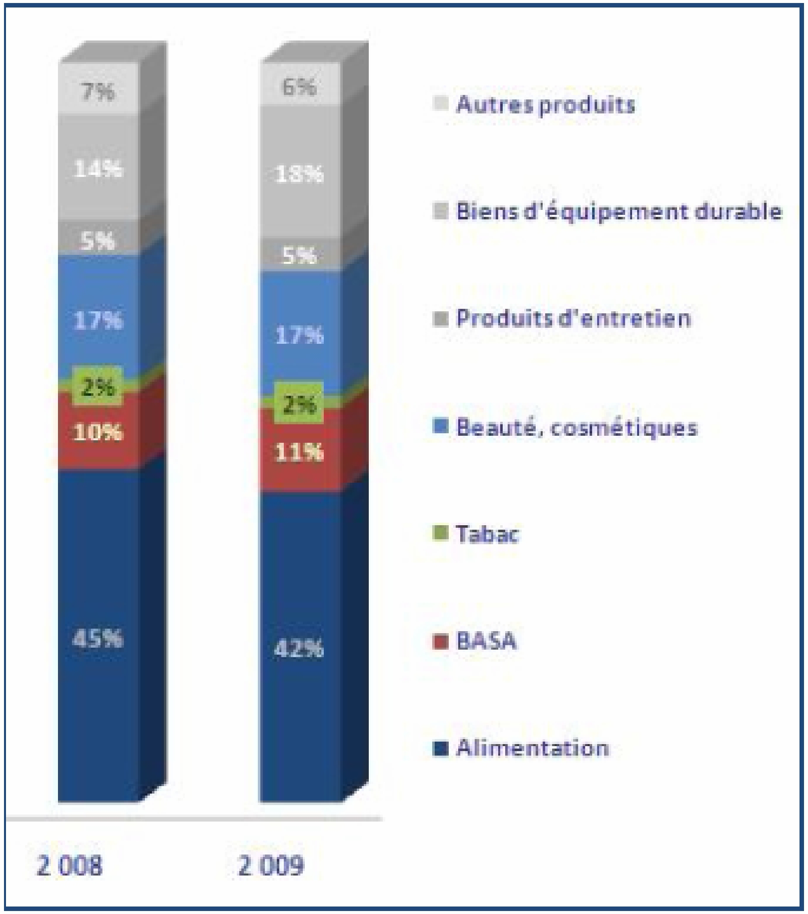
PART DE LA GRANDE CONSOMMATION DANS L'ACTIVITE DES INSTITUTS, 2003-2009



Les **services financiers**, les **sociétés d'études**, le **secteur des énergies**, le **tourisme**, l'univers de la publicité et de la **communication** voient leur part relative augmenter dans l'activité des instituts au cours de 2009.

SYNTEC Etudes Marketing & Opinion

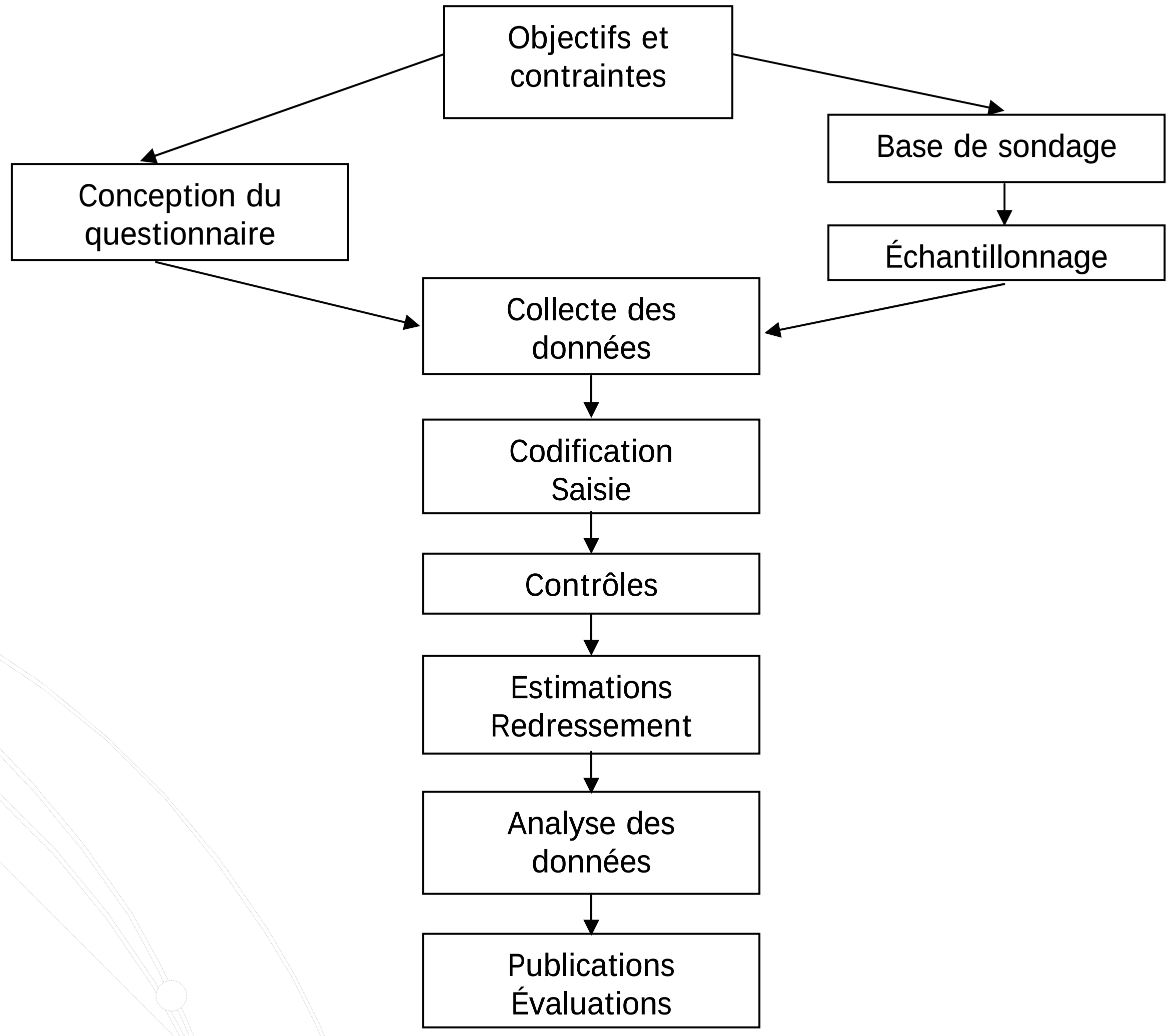
MARCHE DE LA GRANDE CONSOMMATION, 2008-2009



INTRODUCTION

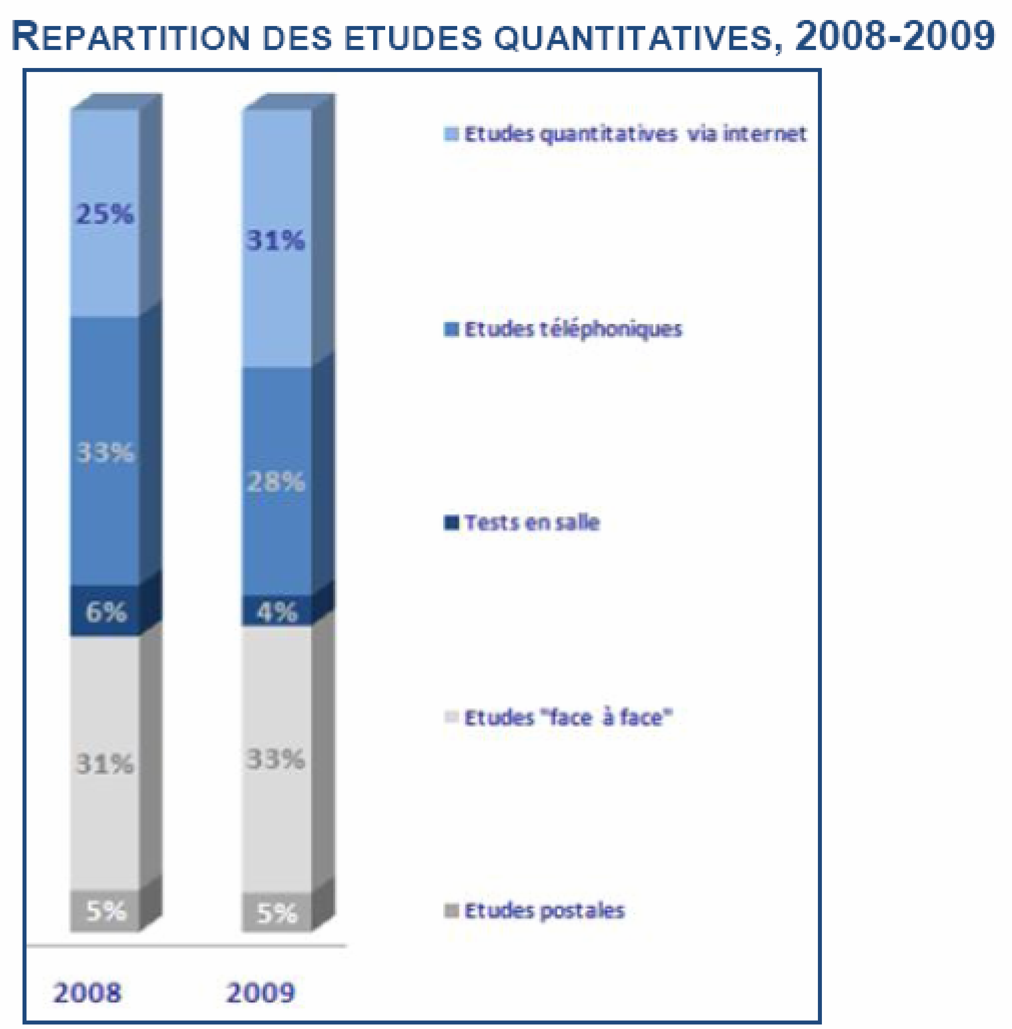
- Histoire récente
  - **1895** – Kiaer, dénombrements représentatifs
  - **1925** – Jensen
  - **1934** – Neyman, Sondages à 2 degrés
  - **1936** – Election de Roosevelt
  - **1938** – Fondation de l'IFOP
  - **1952** – Horvitz et Thompson, Sondages à probabilités inégales
  - **1965** – Ballottage De Gaulle

INTRODUCTION



Modes de recueil:

SYNTEC Etudes Marketing & Opinion



LES TECHNIQUES DE SONDAGE

- Méthodes aléatoires:

  Plans de sondage

  - **<u>Simples</u>:** - à probabilités égales
    - à probabilités inégales

  - **<u>Complexes</u>:** - stratifié
    - en grappe
    - plusieurs degrés

LES TECHNIQUES DE SONDAGE

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

LES TECHNIQUES DE SONDAGE

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

SONDAGE ALEATOIRE SIMPLE

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

$$\overline{Y} = \frac{1}{N}\sum_{i=1}^{N} Y_i \; ; \qquad T = \sum_{i=1}^{N} Y_i$$

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

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

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

- Démonstration avec les variables de Cornfield

$$\delta_i = \begin{cases} 1 \ si \ i \in s \\ 0 \ si \ i \notin s \end{cases}$$

$$E(\delta_i) = \pi_i$$

$$V(\delta_i) = \pi_i(1-\pi_i) \quad \mathrm{cov}(\delta_i;\delta_j) = \pi_{ij} - \pi_i\pi_j$$

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

y_i=variable aléatoire;

Y_i= variable non aléatoire

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

SONDAGE ALEATOIRE SIMPLE

- Covariance entre variables de Cornfield

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

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

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

- Variance de la moyenne

$$V(\bar{y}) = V\left(\frac{1}{n}\sum_{i=1}^{N} Y_i\delta_i\right) = \frac{1}{n^2}\left[\sum_{i=1}^{N} Y_i^2 V(\delta_i) + \sum_{i\neq j}\sum Y_iY_j\,\mathrm{cov}(\delta_i;\delta_j)\right]$$

$$= \frac{\tau(1-\tau)}{n^2}\left[\sum_{i=1}^{N} Y_i^2 - \sum_{i\neq j}\sum\frac{Y_iY_j}{N-1}\right] = \frac{\tau(1-\tau)}{n^2}NS^2 = (1-\tau)\frac{S^2}{n}$$

SONDAGE ALÉATOIRE SIMPLE

- Variances:

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

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

Estimation de S^2:

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

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

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

SONDAGE ALÉATOIRE SIMPLE

- Intervalles de confiance pour un paramètre d'intérêt (« fourchette »)
  - Intervalle ayant une probabilité 1-α (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 connaitre 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_cenlimit.html
- n>30 est souvent suffisant

Intervalle de confiance théorique pour une moyenne

- Tirages indépendants (avec remise) et n>30

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

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

- 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

$$Y_i=\begin{cases}1\\0\end{cases} \qquad \bar{Y}=p$$

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

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

Calculs de taille d'échantillon

- Pour une précision fixée

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

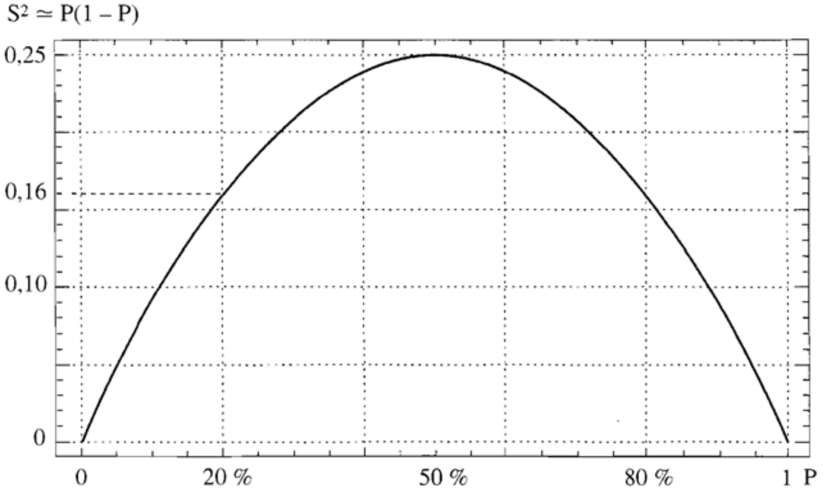
  - Nécessite de connaitre S !

Pour une proportion

- Si n grand et τ faible

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

  - Utile si on connait approximativement *p* a priori



Ardilly, 2006

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

- Solution prudente (ou pessimiste)

  Se placer dans le cas p=0.50

  avec α=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