le cnam

# ENQUETES et SONDAGES
# STA 108 2011-2012

**Intervenants**
**G.Saporta (CNAM), P.Périé (IPSOS), S.Rousseau (INSEE)**

**Plan du cours:**

| | | |
|---|---|---|
| 1 | 7 octobre | Introduction GS et PP |
| ED1 | 10 octobre | Rappels - Sondage aléatoire simple 1 SR |
| 2 | 14 octobre | Sondage aléatoire simple GS |
| ED2 | 17 octobre | Sondage aléatoire simple 2 SR |
| 3 | 21 octobre | Sondages à probabilités inégales GS |
| ED3 | 24 octobre | Plans à probabilités inégales SR |
| 4 | 28 octobre | Algorithmes de tirage PP |
| ED4 | 31 octobre | |
| 5 | 4 novembre | Stratification PP |
| ED5 | 7 novembre | TP simulations de tirage SR |
| ED6 | 14 novembre | Plans stratifiés 1 SR |
| 6 | 18 novembre | Sondages à deux degrés et grappes PP |
| ED7 | 21 novembre | Plans stratifiés 2 SR |
| 7 | 25 novembre | Redressement (quotient, régression, post-strates) GS |
| ED8 | 28 novembre | Plans par grappes SR |
| 8 | 2 décembre | Données manquantes et non-réponses GS et SR |
| ED9 | 5 décembre | Plans à plusieurs degrés SR |
| 9 | 9 décembre | Sources d'erreur et biais PP |
| ED10 | 12 décembre | TP correction de la non-réponse SR |
| 10 | 16 décembre | La méthode des quotas PP |

| | | |
|---|---|---|
| 11 | 6 janvier | Les panels GS et PP |
| ED11 | 9 janvier | Redressement 1 SR |
| 12 | 13 janvier | Effets et pratique des redressements PP |
| ED12 | 16 janvier | Redressement 2 SR |
| 13 | 20 janvier | Le recensement SR |
| ED13 | 23 janvier | TP redressement SR |
| 14 | 27 janvier | Questionnaires, enquêteurs et enquêtés PP |
| ED14 | 30 janvier | Compléments et révisions SR |
| 15 | 3 février | Modes de recueil (avec et sans enquêteur) PP |

**Examen 1ère session : 10 février 2012**
**Examen 2ème session : 20 avril 2012**

**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.**

# 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*

- 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.

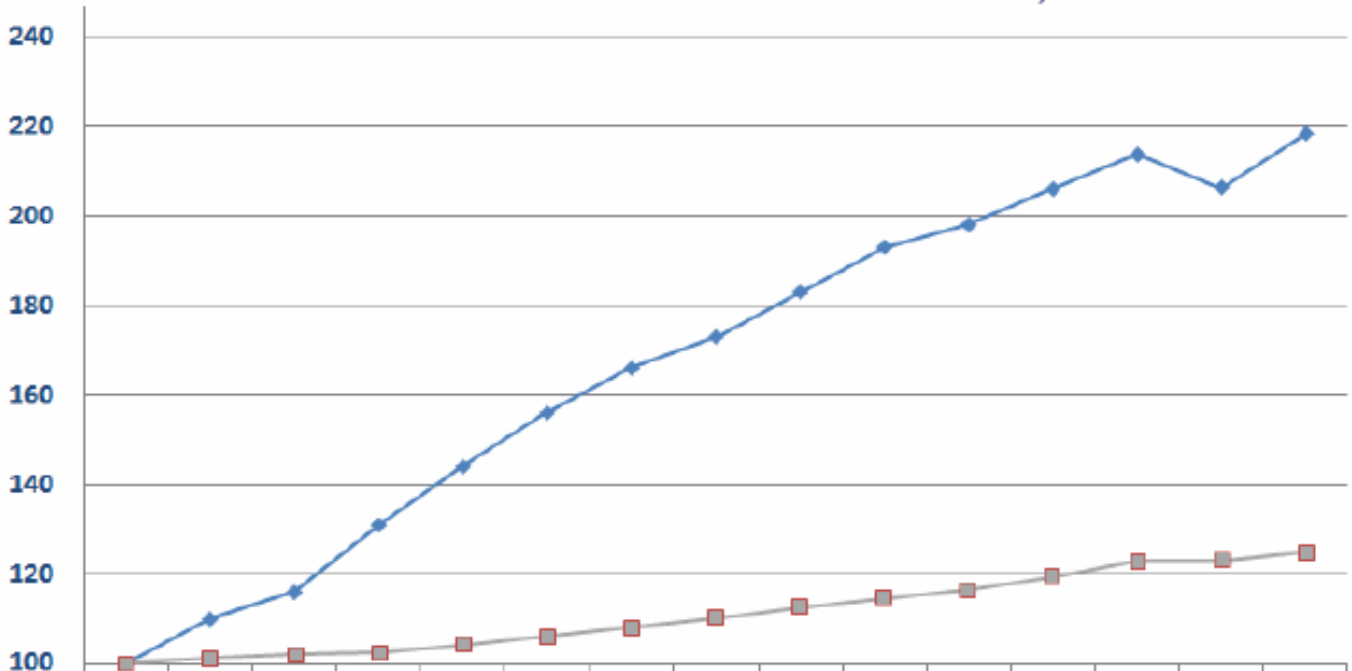
  - Le **service statistique public** (Insee et services statistiques ministériels ) est le moteur dans sa conception, sa production et sa diffusion.

  - 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 connait pas la crise*

- Près de **400 instituts d'étude de marché et d'opinion** identifiés en France
- Marché estimé de **2 milliards d'euros en 2010**
- **Environ 12 000 personnes**, hors enquêteurs

# TAUX DE CROISSANCE DU MARCHE DES ETUDES MARKETING & OPINION, 1996-2010




| | 1996 | 1997 | 1998 | 1999 | 2000 | 2001 | 2002 | 2003 | 2004 | 2005 | 2006 | 2007 | 2008 | 2009 | 2010 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Etudes marketing & opinion | 100 | 110 | 116 | 131 | 144 | 156 | 166 | 173 | 183 | 193 | 198 | 206 | 214 | 206 | 218 |
| Inflation | 100 | 101 | 102 | 102 | 104 | 106 | 108 | 110 | 113 | 115 | 116 | 119 | 123 | 123 | 125 |

SYNTEC Etudes Marketing & Opinion

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

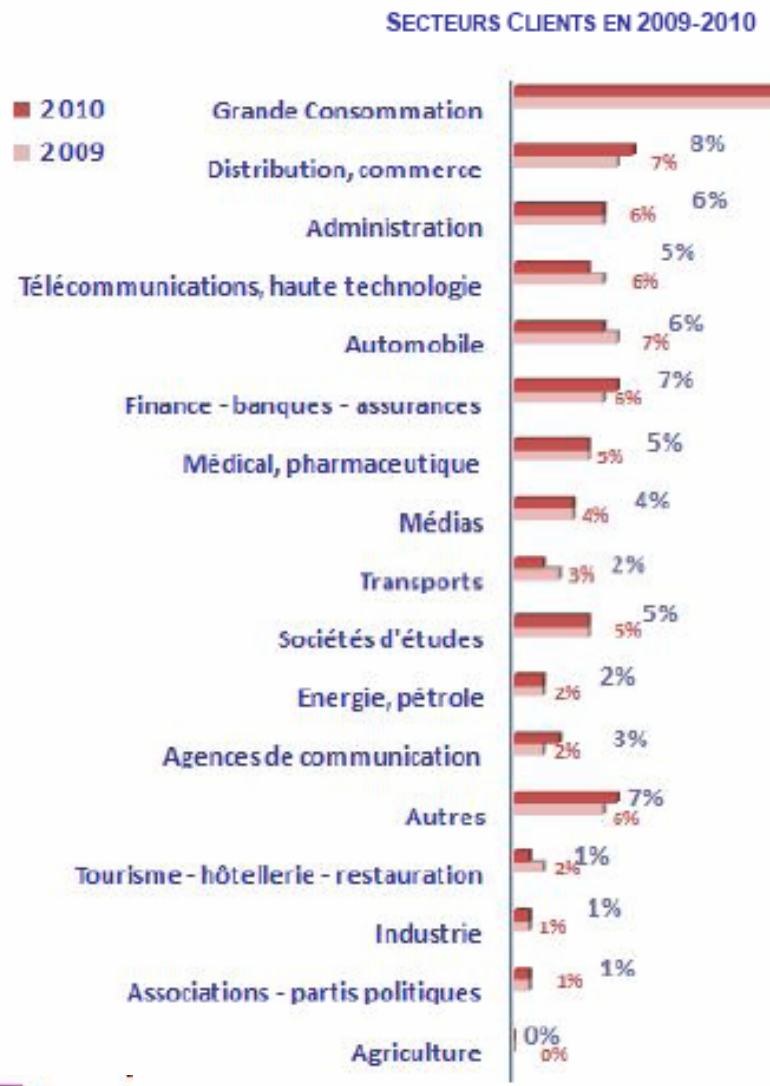
**Effectifs et chiffres d'affaires des principaux instituts de sondages**

| 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*

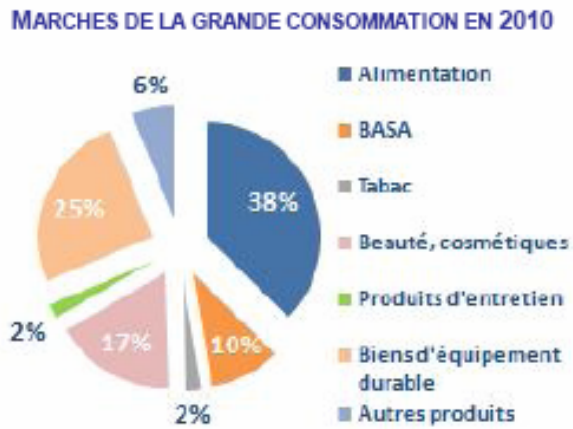
Source: rapport Portelli-Sueur, Sénat

# Secteurs Clients : diminution de la part relative de la grande consommation




Le principal secteur Client, la **grande consommation**, voit la part relative de son chiffre d'affaires diminuer régulièrement depuis 2003.

Les autres secteurs présentent une répartition stable entre 2009 et 2010. Une évolution légèrement positive est notée dans le même temps pour la **distribution**, les **services financiers** et les **agences de communication**.




SYNTEC Etudes Marketing & Opinion

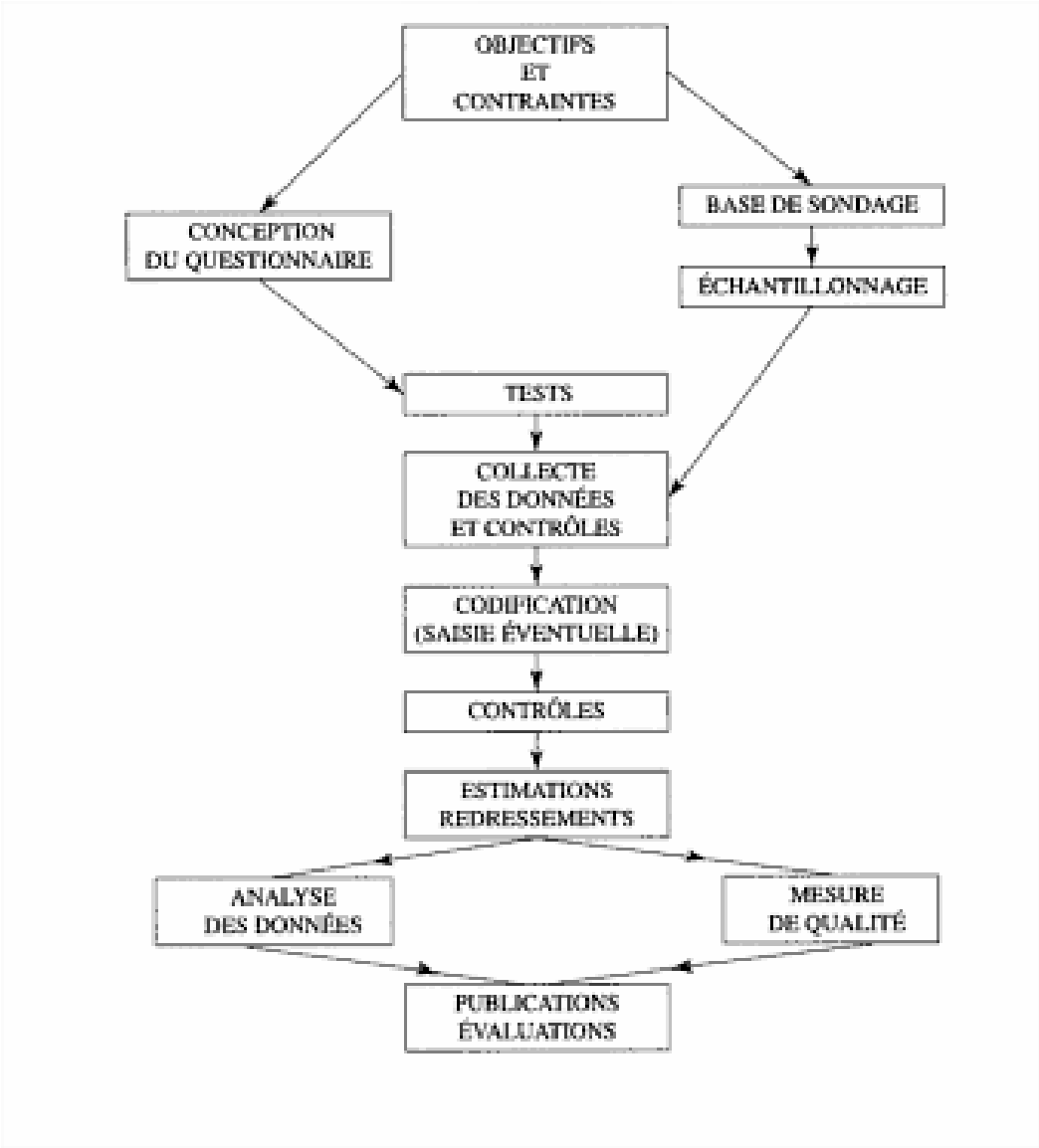
# *Une histoire récente*

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

# *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

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

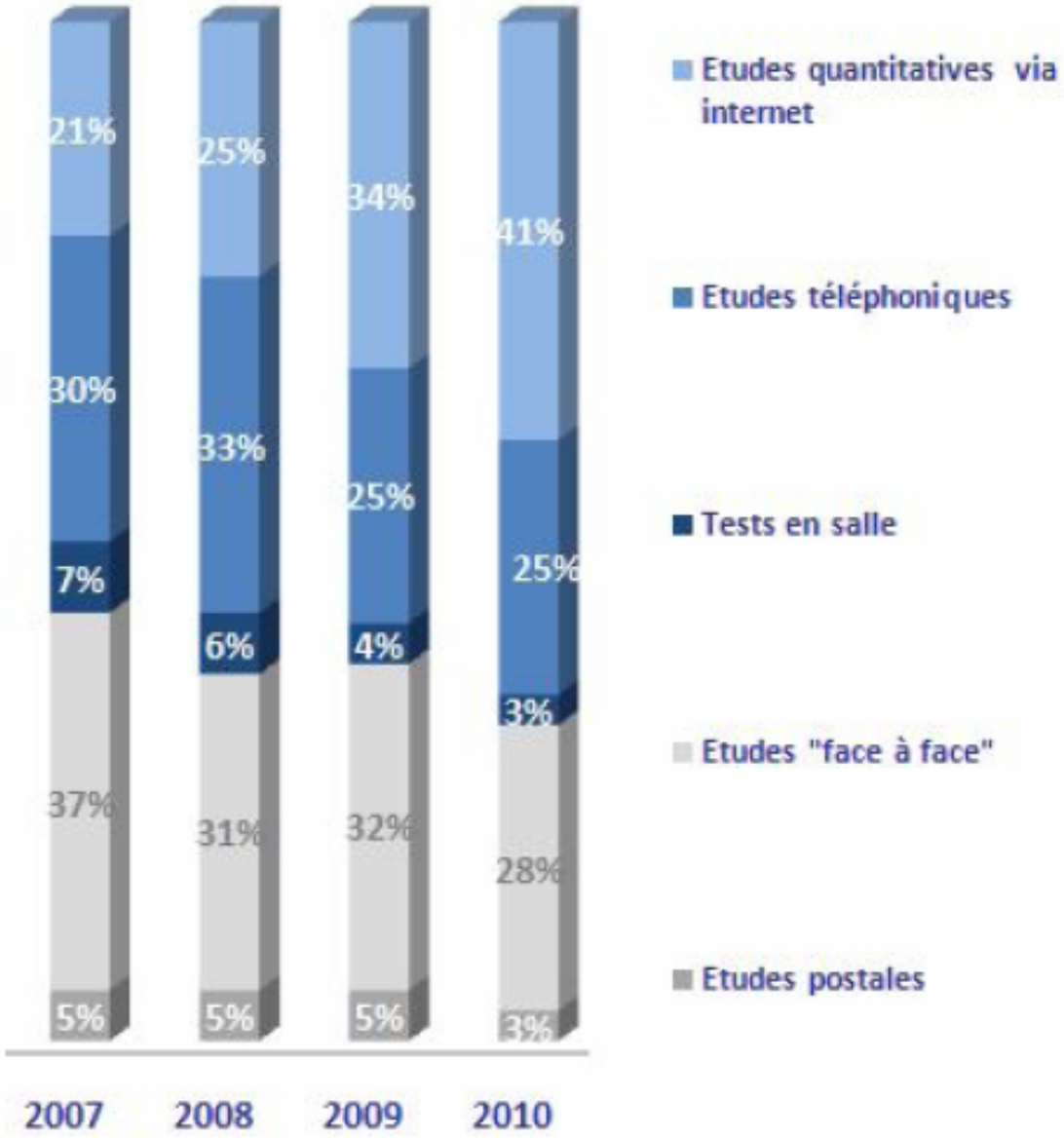
  - **<u>Complexes</u>:** - stratifié
    - en grappe
    - plusieurs degrés

# *LES TECHNIQUES DE SONDAGE*

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

SYNTEC Etudes Marketing & Opinion

# REPARTITION DES METHODES DE RECUEIL ET D'ETUDES DANS LE CA QUANTITATIF, 2007-2010



# Représentativité

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

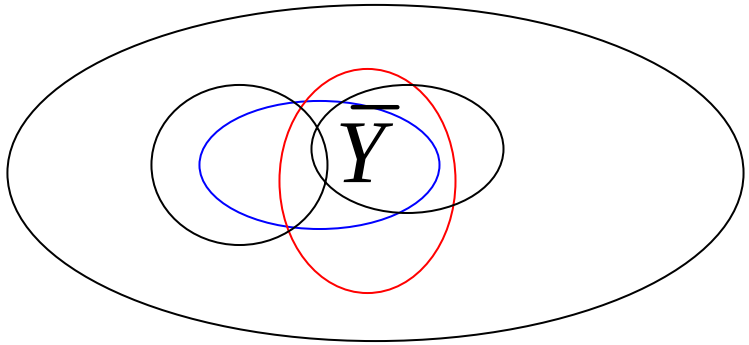
*Voir invoquée la "représentativité" dans un rapport d'enquête pour justifier de la qualité d'un sondage peut presque à coup sûr laisser soupçonner que l'étude a été réalisée dans une méconnaissance totale de la théorie de l'échantillonnage. Le concept de représentativité est aujourd'hui à ce point galvaudé 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




$\overline{y}_1, \overline{y}_2, \ldots, \overline{y}_q$

- 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** (Y1, Y2.....YN)

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

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

# SONDAGE ALÉATOIRE SIMPLE

- Définition: tirage équiprobable sans remise de n unités;
- Taux de sondage: $\frac{n}{N} = \tau$
- $C_N^n$ échantillons possibles;
- $\pi_i$ 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

$$\delta_i = \begin{cases} 1 \ si \ i \in s \\ 0 \ si \ i \notin s \end{cases} \qquad E(\delta_i) = \pi_i$$

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

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

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

$y_i$=variable aléatoire;

$Y_i$= variable non aléatoire

# 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:

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

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

Estimation de $S^2$:

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

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

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

# *SONDAGE ALÉATOIRE SIMPLE*

- Intervalles de confiance pour un paramètre d'intérêt (« fourchette »)
  - Intervalle ayant une probabilité 1-$\alpha$ (niveau de confiance) de contenir la vraie valeur du paramètre. $\alpha$ 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 $\mu$ et d'écart-type $\sigma$ 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

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

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

- 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

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

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

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

# *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 !

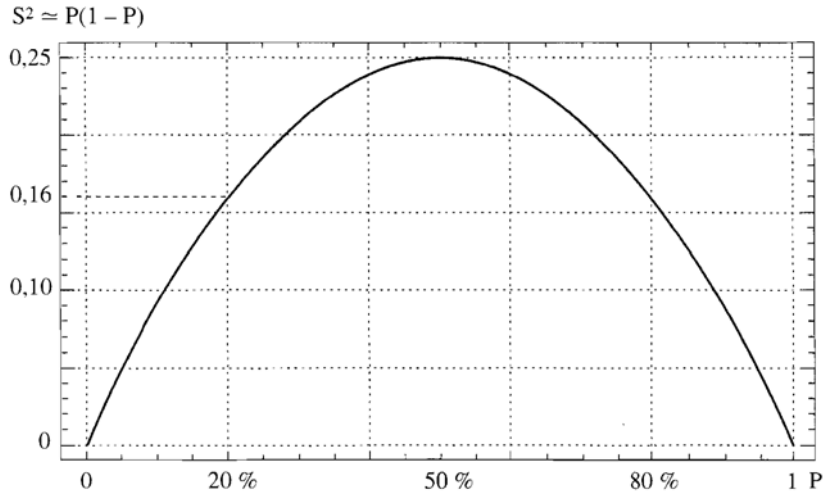
## *Pour une proportion*

- Si n grand et $\tau$ 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 $\alpha$=0.05

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

- Pour $\tau$ 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 $\Delta/p$ change tout
- Compromis à faire quand il y a plusieurs variables d'intérêt
- Attention aux non-réponses: la précision dépend du nombre de répondants