ENQUETES et SONDAGES
STA 108 2009-2010

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

Plan du cours:

1 9 octobre	Introduction GS et OM
ED1 12 octobre	Rappels de probabilité et de statistique SR
2 16 octobre	Sondage aléatoire simple GS
ED2 19 octobre	Sondage aléatoire simple 1 SR
3 23 octobre	Sondages à probabilités inégales GS
ED3 2 novembre	Sondage aléatoire simple 2 SR
4 6 novembre	Algorithmes de tirage GS
ED 4 9 novembre	Plans à probabilités inégales 1 SR
5 13 novembre	Stratification GS
6 16 novembre	Sondages à deux degrés et grappes GS
ED5 20 novembre	Plans à probabilités inégales 2 SR
ED6 23 novembre	TP simulations de tirage SR
7 27 novembre	Redressement (quotient, régression, post-strates) GS
ED7 30 novembre	Plans stratifiés 1 SR

8	4 décembre	Données manquantes et fusion de fichiers GS
ED8	7 décembre	Plans stratifiés 2 SR
9	11 décembre	Sources d'erreur et biais OM
ED9	14 décembre	Plans par grappes SR
10	18 décembre	Effets et pratique des redressements OM
ED10	4 janvier	Plans à plusieurs degrés SR
11	8 janvier	La méthode des quotas OM
ED11	11 janvier	Redressement 1 SR
12	15 janvier	Les panels GS et OM
ED12	18 janvier	Redressement 2 SR
13	22 janvier	Le recensement SR
ED13	25 janvier	TP redressement SR
14	29 janvier	Questionnaires, enquêteurs et enquêtés OM
15	1 février	Modes de recueil (avec et sans enquêteur) OM
ED14	5 février	Compléments et révisions SR

Examen 1ère session : 12 février
Examen 2ème session : 23 avril

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/
- INSEE : http://www.insee.fr
- IPSOS: http://www.ipsos.fr/
- Assoc. Intern. Statisticiens d'enquête: http://www.cbs.nl/isi/iass/
- SYNTEC Etudes http://www.syntec-etudes.com/ http://www.syntec-

INTRODUCTION

- Aperçu du secteur
  - statistique publique

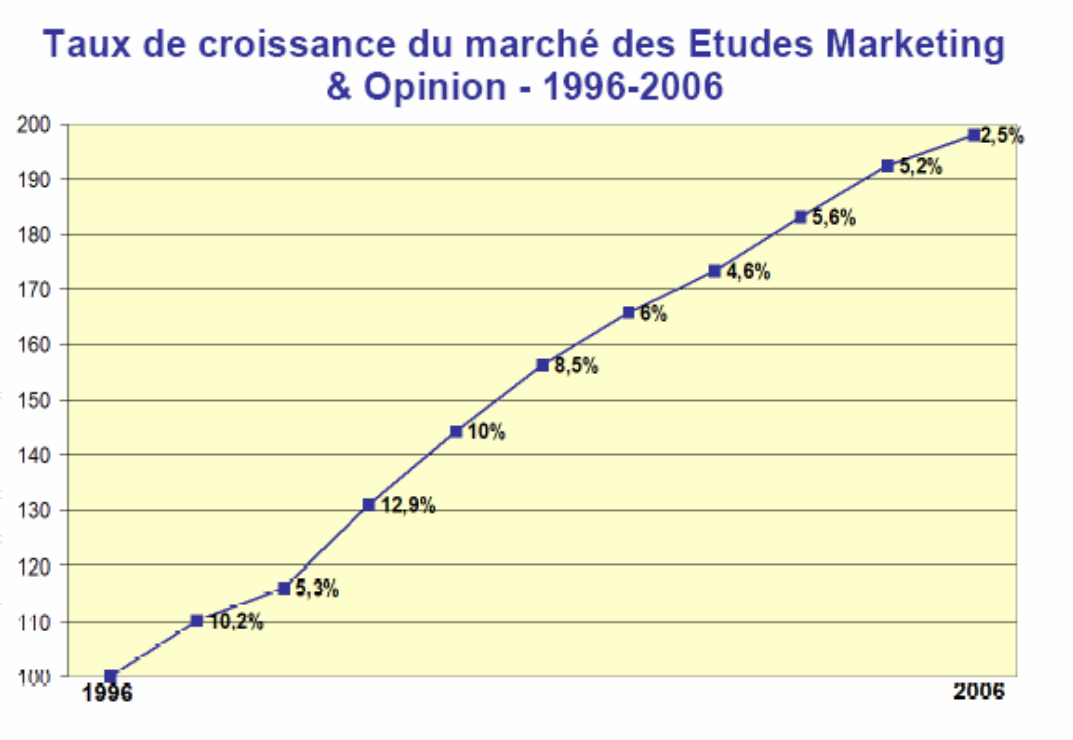
    CNIS

    INSEE – 7 000 employés
  - Secteur privé:

    près de 400 grands groupes et sociétés identifiés **en France** consolident un marché de **1.85 milliards d'euros en 2006**, avec un **effectif total d'environ 12000 personnes**, hors enquêteurs

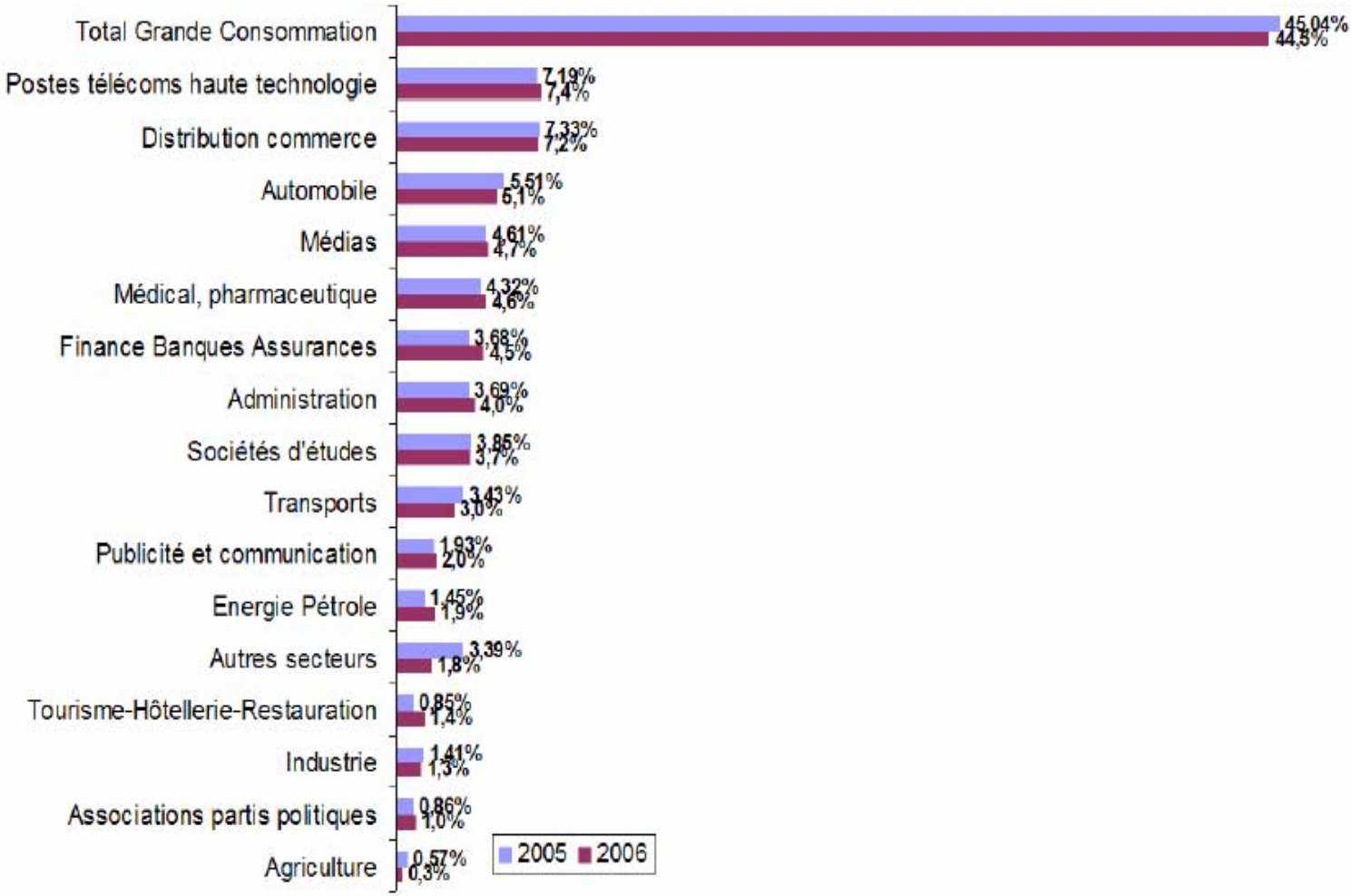
INTRODUCTION

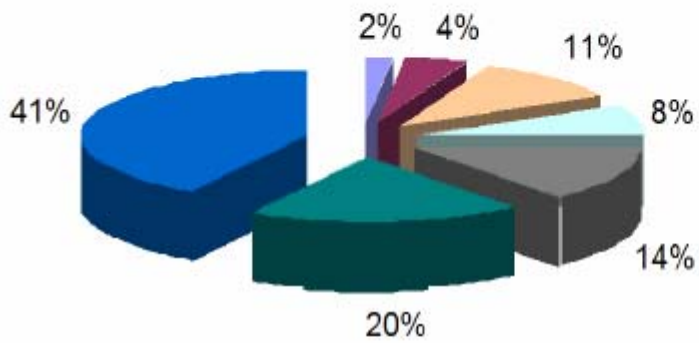
Progression du CA des 63 membres de Syntec Etudes Marketing et Opinion



INTRODUCTION

Secteurs d'activité Clients : une répartition stable d'année en année





Tabac | Autres | Boissons | Produits d'entretien | Biens d'équipement durable | Beauté Cosmétique | Alimentation

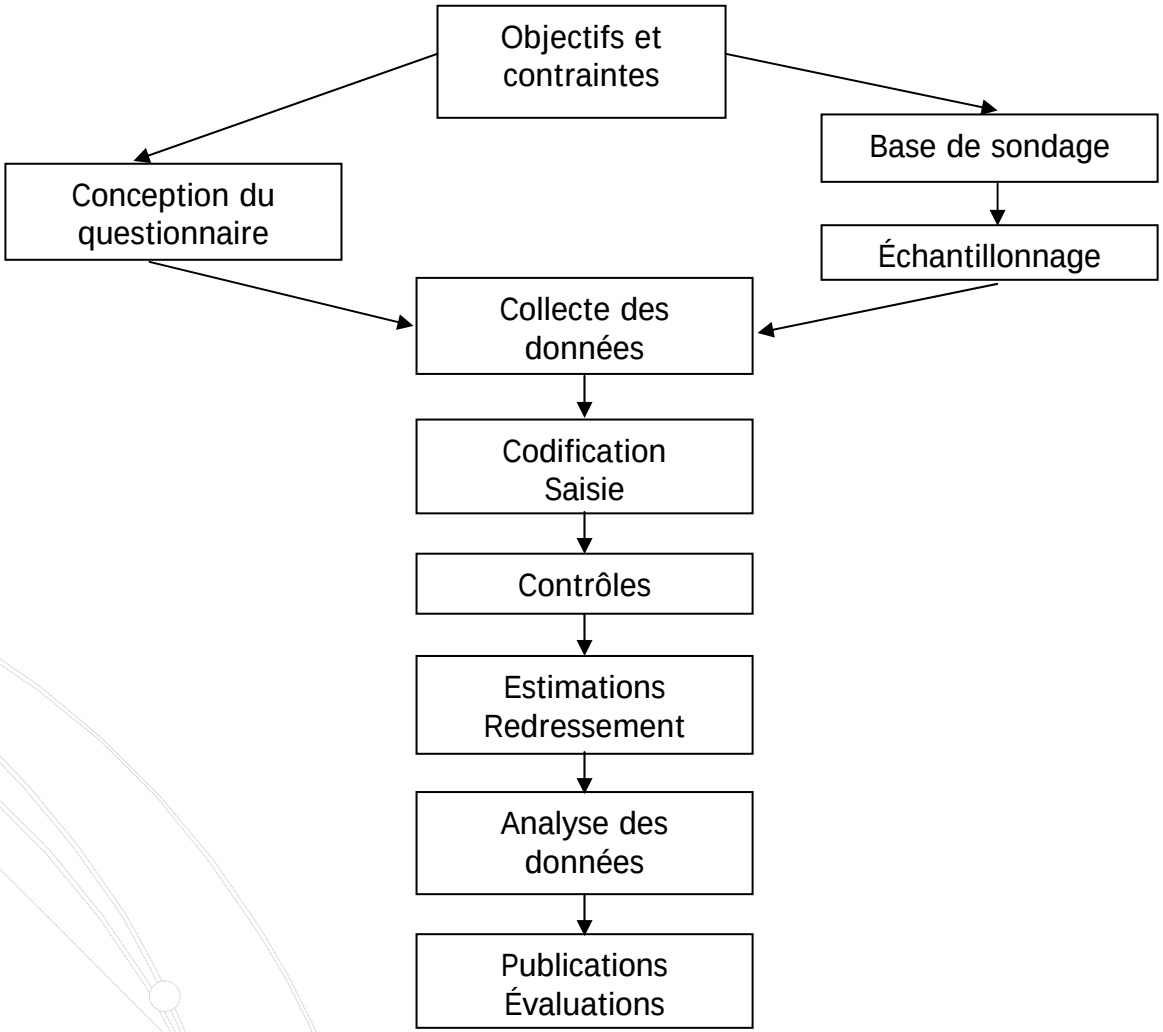
Comme en 2005, **l'alimentation** (41%), **les cosmétiques** (20%) et les **biens d'équipement durable** (14%) constituent les premiers marchés de la **grande consommation** à solliciter les instituts d'études.

2

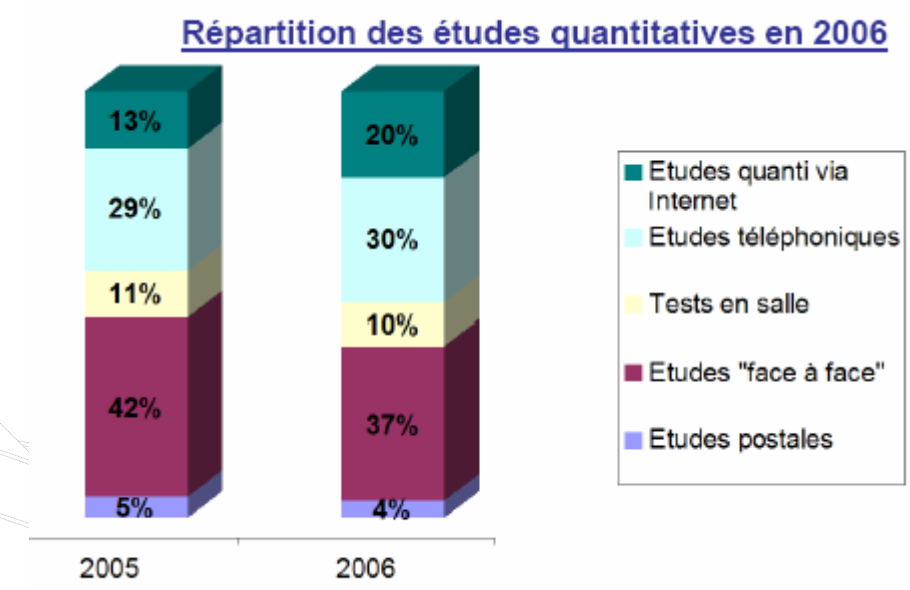
INTRODUCTION

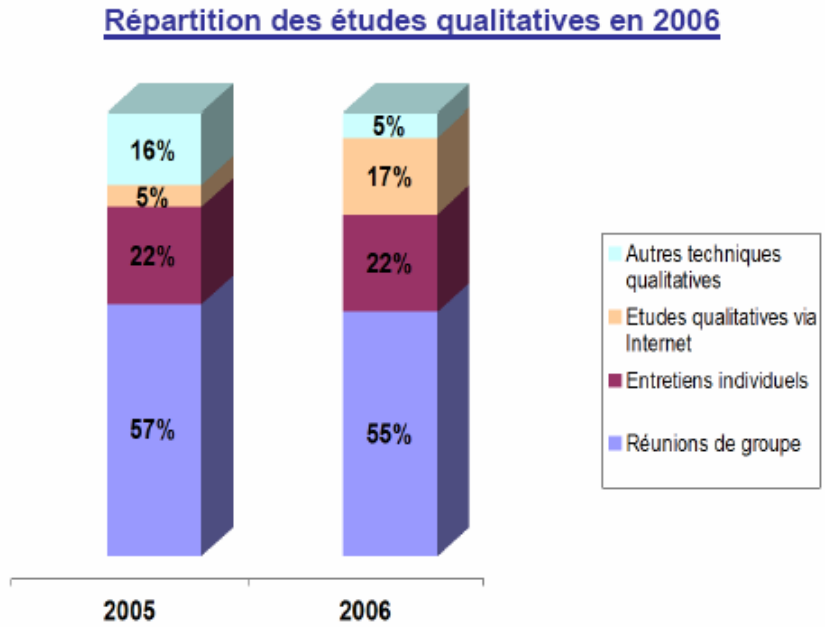
- Histoire récente
  - **1895** – Kiaer, dénombrements représentatifs
  - **1925** – Jensen
  - **1934** – Neyman, Sondages à 2 degrés
  - **1936** – Election de Roosevelt
  - **1938** – Fondation de l'IFOP
  - **1952** – Horvitz et Thompson, Sondages à probabilités inégales
  - **1965** – Ballottage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

INTRODUCTION



Modes de recueil





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;
  - Sondage « à chaud ».

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

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

- 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

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

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

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-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

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

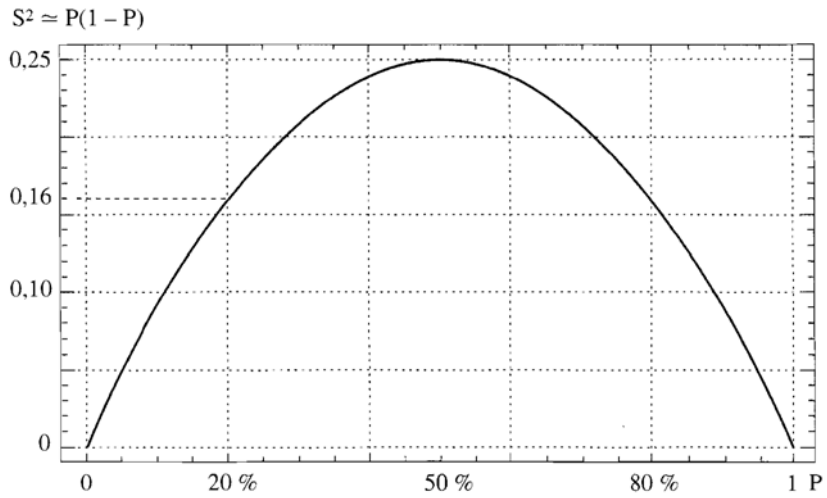
  - Nécessite de connaitre S !

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

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

- 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