INTRODUCTION AUX SONDAGES

1. Les différentes étapes d'une enquête par sondage

2. Différents types d'échantillons et estimateurs associés

3. La construction du questionnaire

1. LES DIFFÉRENTES ÉTAPES D'UNE ENQUÊTE PAR SONDAGE

A. Définition de l'objet d'étude

B. Préparation

C. Collecte

D. Dépouillement

E. Traitement statistique, interprétation et communication

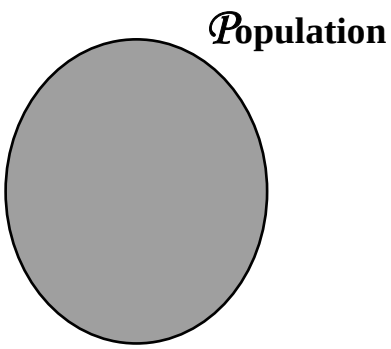
F. Qualité d'une enquête par sondage

A. DÉFINITION DE L'OBJET D'ÉTUDE

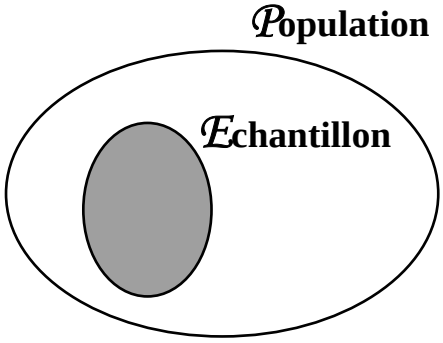
1. POPULATION CONCERNÉE

Problème de définition dans l'espace et dans le temps
Problème de nomenclature

2. RECENSEMENT ou SONDAGE



Toute le population est enquêtée.



Une partie de la population, *l'échantillon*, est enquêtée.

- **Comment choisir l'échantillon ?**
- **Comment extrapoler les résultats obtenus sur l'échantillon à toute la population ?**

Note :
En statistique, sondage s'oppose à recensement.
En anglais, on a deux mots : "poll" pour le sondage d'opinion, "sampling" pour l'enquête par sondage. En espagnol aussi : "sondeo" pour le sondage d'opinion, "muestreo" pour l'enquête par sondage.
Les statisticiens devraient parler "d'enquêtes par échantillonnage" pour les distinguer des "sondages d'opinion" réalisés par les sondeurs et les politologues.

3. AVEC QUESTIONNAIRE ou SANS QUESTIONNAIRE

1. Enquête d'opinion auprès des enseignants

2. Enquête auprès des ménages sur leurs déplacements

3. Résultats concours en fonction du dossier scolaire

4. Enregistrement des ventes de logements à la Chambre des Notaires

4. INFORMATION RECHERCHÉE : LES PARAMÈTRES D'INTÉRÊT

1/ Deux types de variables

- **Variables catégorielles**

ex 1 :

sexe	**2 modalités**	**H, F**
établissement	**4 modalités**	**école primaire, collège, lycée professionnel, autre lycée**

ex 4 :

type de logement	**5 modalités**	**Studio, T2, T3, T4, T5 et +**
zone d'habitation	**6 modalités**	**Z_1, Z_2, …, Z_6**

- **Variables réelles**

ex 3 : note à la première épreuve d'écrit

ex 4 : surface du logement (en m^2)
prix de la transaction (en euros)

2/ Les paramètres d'intérêt selon le type des variables

- **Variables catégorielles**

  ***a/ la proportion* d'une catégorie *A* d'une variable catégorielle**

  $$\boldsymbol{P = \frac{N_A}{N} = \frac{\text{Effectif de la population dans la catégorie A}}{\text{Effectif total de la population}}}$$

  b/ éventuellement l'effectif N_A appelé aussi la taille

  c/ une différence de deux proportions

- **Variables réelles**

  a/ la moyenne μ d'une variable réelle Y

  $$\mu = \frac{1}{N}\sum_{i=1}^{N} Y_i = \frac{\tau}{N} \quad \text{où } \tau = \sum_{i=1}^{N} Y_i \text{ est le total}$$

  b/ éventuellement le total τ

  c/ une différence de deux moyennes

  d/ la variance σ^2 d'une variable réelle Y

  $$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}\left(Y_i - \mu\right)^2 = \frac{1}{N}\sum_{i=1}^{N} Y_i^2 - \mu^2$$

  e/ une corrélation entre deux variables réelles

f/ un ratio

- **$R = \frac{\mu_Y}{\mu_Z}$ rapport de deux moyennes inconnues**

  exemple 4 : Prix au m^2 Prix/Surface = $\frac{\mu_{Prix}}{\mu_{Surface}}$

  (ce rapport est différent de la moyenne des prix au m^2 des logements)

- **Moyenne par rapport à une sous-population de taille inconnue**
  exemple 4 : le prix moyen des logements avec parking

  $\mu_{Pk} = \frac{1}{N_{Pk}} \sum_{i \text{ avec } Pk} Y_i$ **(avec N_{Pk} inconnu)**

- **Proportion par rapport à une sous-population de taille inconnue**
  exemple 4 : parmi les logements avec parking, proportion de ceux qui sont de construction récente (*CR*)

  $P_{CR/Pk} = \frac{N_{CR \text{ et } Pk}}{N_{Pk}}$

- **Un rapport de deux variances inconnues**

B. PRÉPARATION

Recherche de l'information disponible (information auxiliaire)

Etablissement de la liste des contraintes et des moyens disponibles : coût, temps, faisabilité

Choix de la période d'enquête et choix des moyens

Choix des variables statistiques : les instruments de mesure (éventuellement, construction du questionnaire)

Dans le cas d'un sondage aléatoire :

- réactualisation de la base de sondage

- choix du type de sondage (taille, répartition, ...) et des estimateurs avec évaluation de leur précision

Pré-enquête

Choix et formation des enquêteurs

C. COLLECTE

Tirage de l'échantillon aléatoire (utilisation de table de chiffres au hasard (cf. annexe 1) ou de générateur de nombres pseudo-aléatoires sur calculatrice ou ordinateur)

Enquête sur le terrain

Contrôle du déroulement de l'enquête, du travail des enquêteurs, respect de l'échantillon tiré.

D. DÉPOUILLEMENT

Codage

Saisie sur ordinateur

Vérification

Contrôle de validité

Calcul de pondérations (poids d'échantillonnage)

Calcul des coefficients de redressement (par exemple pour les non-réponses)

E. TRAITEMENT STATISTIQUE, INTERPRÉTATION, COMMUNICATION

Le traitement statistique dépend de l'échantillon

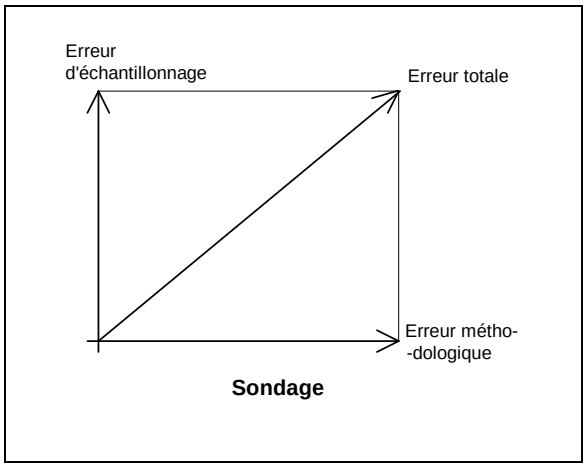
Dans le cas d'un sondage aléatoire,
estimation par intervalle de confiance des paramètres d'intérêt selon le plan d'échantillonnage utilisé

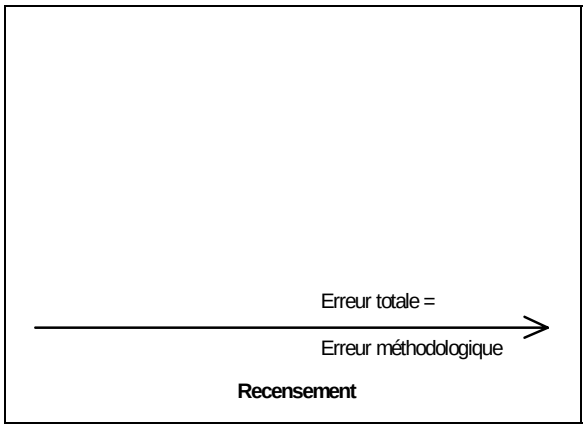
F. LA QUALITÉ D'UNE ENQUÊTE PAR SONDAGE

Dans le cas d'une enquête par sondage, les paramètres que l'on cherche à connaître sur la population seront *estimés* à partir de l'échantillon. La différence entre la valeur du paramètre et la valeur estimée est appelée *erreur d'échantillonnage*. La théorie des sondages permet de contrôler cette erreur en mesurant *la précision des estimateurs*.

Dans une enquête, il existe bien d'autres erreurs, que l'on peut regrouper sous le nom d'*erreur méthodologique* (« non sampling errors » en anglais), qui peuvent se produire lors des différentes étapes de l'enquête :
- définition imprécise des objectifs de l'enquête et de la population concernée,
- mauvaise qualité de la base de sondage ou de la méthode de tirage,
- rédaction ou présentation du questionnaire insatisfaisante,
- incompétence de l'enquêteur,
- refus de répondre de l'enquêté.

On admet que les deux erreurs sont orthogonales et que l'erreur totale est la somme des deux.





Pour certaines enquêtes, plutôt que de faire un recensement sur 100 000 individus qui peut induire une erreur méthodologique importante, on peut avoir une meilleure précision en faisant un sondage sur 1 000 individus.

2. DIFFÉRENTS TYPES D'ÉCHANTILLONS ET ESTIMATEURS ASSOCIÉS

A. Échantillon aléatoire simple

B. Échantillon aléatoire stratifié

C. Échantillon aléatoire par grappes

D. Poids d'échantillonnage

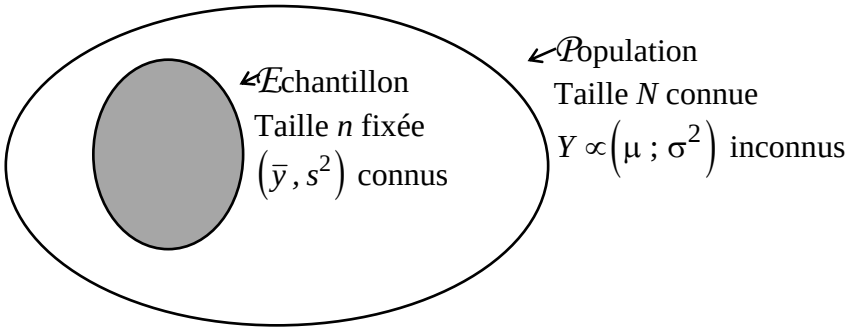
E. Redressement d'échantillons

F. Méthode empirique des quotas

A. ÉCHANTILLON ALÉATOIRE SIMPLE

La liste des individus de la population, appelée ***base de sondage***, est disponible. Un échantillon aléatoire simple à probabilités égales de taille *n* est extrait de la population. On envisagera le cas d'un tirage avec remise ou sans remise.

- ***Échantillon aléatoire simple : estimation d'une moyenne μ***



$\mu = \frac{1}{N}\sum_{i=1}^{N} y_i$ et $\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(y_i - \mu)^2$ sont la moyenne et la variance de la variable Y sur la population ; ce sont des paramètres inconnus qu'il s'agit d'estimer,

$\bar{y} = \frac{1}{n}\sum_{i\in E} y_i$ et $s^2 = \frac{1}{n-1}\sum_{i\in E}(y_i - \bar{y})^2$ sont la moyenne et la variance corrigée observées sur l'échantillon de taille *n*.

$\bar{y}$ et s^2 sont des observations de variables aléatoires réelles $\bar{Y}$ et S^2.
Dans le cas d'un sondage ***avec remise,*** on a :

$$E(\bar{Y}) = \mu, \quad V(\bar{Y}) = \frac{\sigma^2}{n} \quad \text{et} \quad E(S^2) = \sigma^2$$

$\bar{Y}$ et S^2 sont des ***estimateurs sans biais*** de μ et σ^2 respectivement.

On pose $\hat{V}(\bar{Y}) = \frac{S^2}{n}$, alors $\hat{V}(\bar{Y})$ est un estimateur sans biais de $V(\bar{Y})$.

Dans le cas d'un sondage ***sans remise***, on a :

$$E(\bar{Y}) = \mu, \quad V(\bar{Y}) = \frac{\sigma^2}{n}\frac{N-n}{N-1} \quad \text{et} \quad E(S^2) = \sigma^2\frac{N}{N-1}$$

On pose $\hat{V}(\bar{Y}) = \frac{S^2}{n}(1-\frac{n}{N})$, alors $\hat{V}(\bar{Y})$ est un estimateur sans biais de $V(\bar{Y})$.

On retrouve les résultats du sondage avec remise dans le cas où le taux de sondage n/N est négligeable. La variance est nulle lorsque n = N, cas d'un recensement.

- ***Théorème limite central***

Pour *n* suffisamment grand :

$$\sqrt{n}\left(\frac{\overline{Y}-\mu}{\sigma}\right) \propto N(0;1) \quad \text{et} \quad \sqrt{n}\left(\frac{\overline{Y}-\mu}{S}\right) \propto N(0;1)$$

- ***Estimation de la moyenne μ par intervalle de confiance à 95%***

(cas où le taux de sondage est négligeable)

$$\left[\bar{y} - 1.96\frac{s}{\sqrt{n}} \; ; \bar{y} + 1.96\frac{s}{\sqrt{n}}\right]$$

que l'on notera aussi $\bar{y} \pm 1.96\frac{s}{\sqrt{n}}$, intervalle centré sur $\bar{y}$.

La ***précision absolue*** à 95% de confiance est égale à $1.96\frac{s}{\sqrt{n}}$.

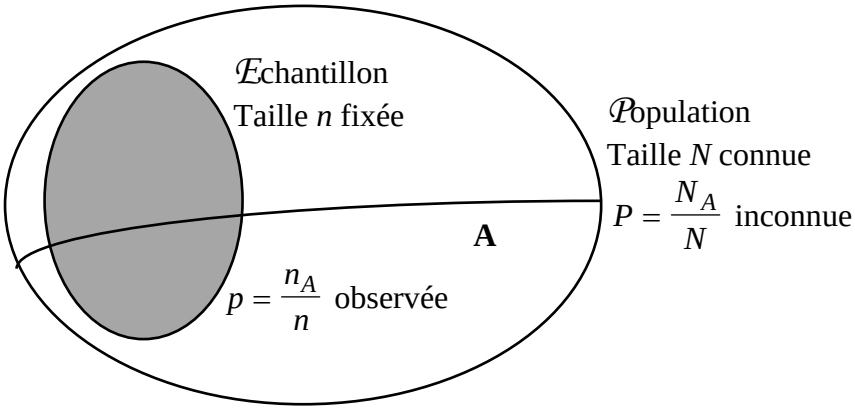
La ***précision relative*** à 95% de confiance est égale à $1.96\frac{s}{\bar{y}\sqrt{n}}$.

Ces définitions ne sont pas très heureuses car la précision de l'estimateur est d'autant plus grande que les mesures de la précision (telles que définies ici) sont petites.

Dans le cas de population de grande taille *N*, le taux de sondage est généralement négligeable et **on remarquera que la précision ne dépend que de la taille de l'échantillon** (et de la variabilité des données).

Dans le cas où le taux de sondage n'est plus négligeable les précisions sont multipliées par le facteur de correction $\sqrt{1-\frac{n}{N}}$, coefficient inférieur à 1.

- ***Échantillon aléatoire simple : estimation d'une proportion*** $\boldsymbol{P = \frac{N_A}{N}}$



p, proportion observée sur l'échantillon, est l'observation d'une variable aléatoire de moyenne *P*, la proportion inconnue sur la population, et de variance $\frac{P(1-P)}{n}\frac{N-n}{N-1}$ (et donc $\frac{P(1-P)}{n}$ dans le cas où le taux de sondage est négligeable ($\frac{n}{N} < 0.1$)).

Il s'agit d'un cas particulier d'estimation d'une moyenne ; la variable réelle *Y* est l'indicatrice de *A*, on a alors :
$\mu = P$, $\sigma^2 = P(1-P)$, $\bar{y} = p$ et $s^2 = p(1-p)$ (en confondant *n* et $n-1$).

Estimation de P par intervalle de confiance à 95% (cas où $\frac{n}{N} < 0.1$)

$$p \pm 1.96\sqrt{\frac{p(1-p)}{n}}$$

On a en particulier, pour $p = 0.5$, valeur de p (compris entre 0 et 1) rendant $p(1-p)$ maximal :

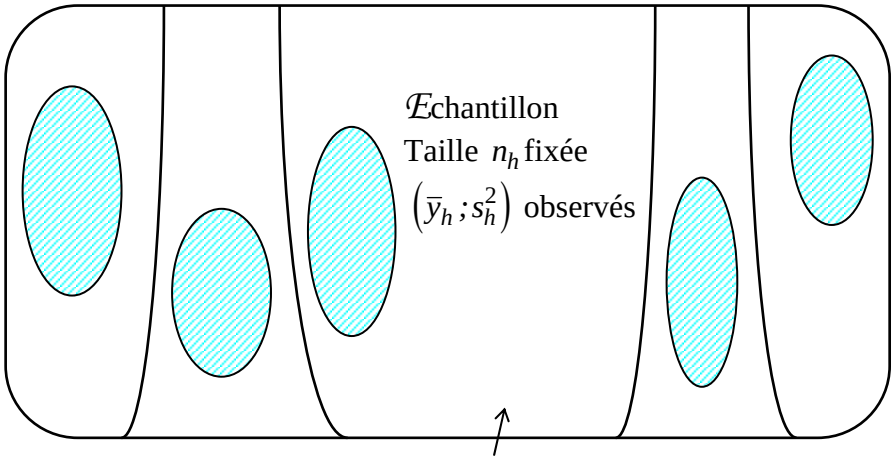
$$p \pm \frac{1}{\sqrt{n}}$$

Cf. Annexe 2, Table de précision absolue à 95% de l'estimation d'une proportion.

B. ÉCHANTILLON ALÉATOIRE STRATIFIÉ

- ***Estimation d'une moyenne μ***

H échantillons indépendants de taille fixée : n_h , $h = 1,\ldots,H$; $\sum_{h=1}^{H} n_h = n$



Décomposition de la moyenne et de la variance

$$\mu = \sum_{h=1}^{H} \frac{N_h}{N} \mu_h$$

$$\sigma^2 = \sum_{h=1}^{H} \frac{N_h}{N} (\mu_h - \mu)^2 + \sum_{h=1}^{H} \frac{N_h}{N} \sigma_h^2$$

$$= \quad \sigma^2_{\text{inter}} \quad + \quad \sigma^2_{\text{intra}}$$

= variance des moyennes + moyenne des variances

Estimation ponctuelle de μ :

$$\bar{y}_s = \sum_{h=1}^{H} \frac{N_h}{N} \bar{y}_h$$

Variance de $\bar{y}_s$ (cas de taux de sondage négligeables) :

$$V(\bar{y}_s) = \sum_{h=1}^{H} \left(\frac{N_h}{N} \right)^2 \frac{\sigma_h^2}{n_h}$$

Estimation de la variance de $\bar{y}_s$:

$$\hat{V}(\bar{y}_s) = \sum_{h=1}^{H} \left(\frac{N_h}{N} \right)^2 \frac{s_h^2}{n_h}$$

- ***Cas de la stratification proportionnelle*** : $n_h = \frac{N_h}{N}\, n \quad h = 1,\ldots,H$

L'estimation par intervalle de confiance à 95% de μ est alors (en supposant $\frac{n}{N}$ négligeable et en posant $s^2_{\text{intra}} = \sum_{h=1}^{H} \frac{N_h}{N}\, s_h^2$, moyenne des variances) :

$$\bar{y}_s \pm 1.96\, \frac{s_{\text{intra}}}{\sqrt{n}}$$

L'échantillon obtenu est souvent appelé "échantillon représentatif" car il constitue un modèle réduit de la population. Nous préférons l'appeler échantillon stratifié proportionnel car, en théorie des sondages, tous les échantillons aléatoires sont représentatifs dans le sens où on sait estimer les paramètres de la population et évaluer la précision obtenue.
On remarque que l'on a : $s_{\text{intra}} \leq s$ et égalité lorsque les moyennes des strates sont égales.

- ***Cas de la stratification optimale*** : $n_h = \frac{N_h\, \sigma_h}{\sum N_h\, \sigma_h}\, n$

On montre que, à taille globale égale à *n*, l'allocation de l'échantillon qui minimise la variance de l'estimateur de stratification est proportionnelle au produit de la taille de la strate par son écart-type.

L'estimation par intervalle de confiance à 95% de μ est alors (en supposant $\frac{n}{N}$ négligeable et en posant $\bar{s} = \sum_{h=1}^{H} \frac{N_h}{N}\, s_h$, moyenne des écart-types) :

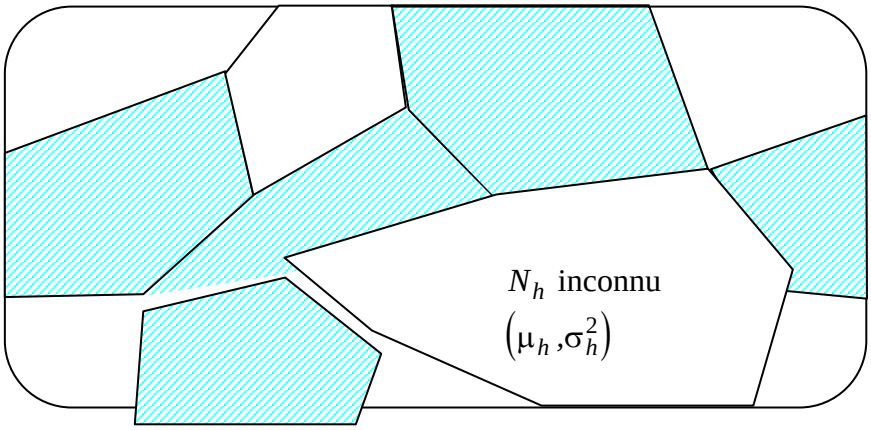
$$\bar{y}_s \pm 1.96\, \frac{\bar{s}}{\sqrt{n}}$$

On remarque que l'on a : $\bar{s} \leq s_{\text{intra}}$ et égalité lorsque les écart-types des strates sont égaux.

En conclusion

Lorsque la variable catégorielle de stratification est fortement liée à la variable réelle *Y* d'intérêt, c'est-à-dire, lorsque le rapport de corrélation défini par $\eta = \sigma_{\text{inter}} / \sigma$ est proche de 1, la stratification proportionnelle améliore considérablement la précision des estimateurs ; si de plus, les dispersions des valeurs de *Y* sont très différentes d'une strate à l'autre, alors la stratification optimale améliore encore davantage la précision des estimateurs.

C. ÉCHANTILLON ALÉATOIRE PAR GRAPPES



M grappes (*M* connu)
N individus (*N* inconnu)
$Y \propto (\mu, \sigma^2)$ inconnus

On tire un échantillon aléatoire simple de *m* grappes (*m* fixé) parmi les *M* grappes.
On observe (N_h, μ_h, σ_h^2) pour la grappe *h* de l'échantillon.

1. *Unités statistiques : les grappes*

Pour le traitement statistique concernant les grappes, on est dans le cadre de l'échantillon aléatoire simple. (Cas particulier : enquête de dénombrement.)

2. *Unités statistiques : les individus*

Dans le cas où l'on veut estimer la moyenne μ par individu d'une variables réelle Y, on est ramené au cas de l'estimation d'un ratio. On montre que, par rapport à un échantillon aléatoire simple de même taille, on perd en précision.

Coefficient de grappe

C'est le rapport entre les précisions absolues du sondage aléatoire par grappe sur le sondage aléatoire simple de même taille. Dans le cas où N_h est connu et égal à N_0 pour toutes les grappes ce coefficient est :

$$\sqrt{N_0}\,\frac{s_{\text{inter}Y}}{s_Y} \leq \sqrt{N_0}$$

où $s^2_{\text{inter}Y}$ et s^2_Y sont les estimations de la variance inter grappes et de la variance totale de la variable *Y*.

D. POIDS D'ÉCHANTILLONNAGE

Le poids d'échantillonnage est le poids p_i qu'il faut affecter à chaque individu *i* de l'échantillon pour que les totaux de l'échantillon ainsi pondéré fournissent les estimations ponctuelles des totaux de la population.

Ces poids vérifient $\sum_{i \in E} p_i = N$.

Exemples :
Dans le cas d'un sondage aléatoire simple à probabilités égales de taille *n*, on a, pour tout *i* de *E* : $p_i = \frac{N}{n}$.
Dans le cas d'un sondage stratifié à probabilités égales de taille $(n_1,...,n_H)$, on a, pour tout *i* de la strate *h* : $p_i = \frac{N_h}{n_h}$.

E. REDRESSEMENT DE L'ÉCHANTILLON

Malgré son nom un peu sulfureux, le redressement d'échantillon est une technique d'estimation utilisant a posteriori une information auxiliaire afin d'améliorer l'estimation. La technique de redressement la plus utilisée est la post-stratification.

- ***Post-stratification***

Exemple : on sait (information auxiliaire) que la population d'intérêt est répartie selon trois tranches d'âge (moins de 25 ans, de 25 à 65 ans, plus de 65 ans) de la façon suivante :

Tranches d'âges	< 25	25 à 65	> 65	Ensemble
Répartition	25%	42%	33%	100%

Or, on observe sur un échantillon aléatoire simple de taille 1200 la répartition suivante :

Tranches d'âges	< 25	25 à 65	> 65	Ensemble
Effectifs	330	370	500	1200
Répartition	27.5%	30.8%	41.7%	100%

Si la répartition théorique avait été respectée on aurait dû observer :

Tranches d'âges	< 25	25 à 65	> 65	Ensemble
Effectifs	300	500	400	1200

On décide alors d'affecter à chacun des individus *i* de la 1ère tranche d'âges le poids $\frac{300}{330}$, à chacun des individus *i* de la 2ème tranche le poids $\frac{500}{370}$ et à chaque individu de la 3ème tranche le poids $\frac{400}{500}$.

Dans le cas du redressement d'échantillon, la somme des poids est égale à la taille de l'échantillon. On a en effet :

$$\sum_{i\in E} p_i = \sum_{h=1}^{H} \sum_{i\in E_h} p_i = 330\left(\frac{300}{330}\right)+370\left(\frac{500}{370}\right)+500\left(\frac{400}{500}\right)=1200\,.$$

Redressement de l'échantillon : consiste à redresser la structure par rapport à des répartitions connues.

L'estimation d'une moyenne μ est alors :

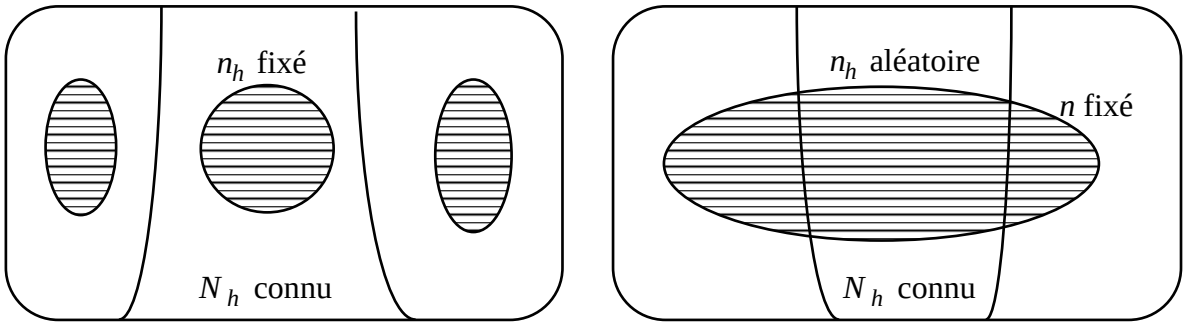
$$\hat{\mu}=\frac{1}{1200}\sum_{i\in E} p_i\, y_i = \frac{1}{1200}\left(\frac{300}{330}\sum_{i\in E_1} y_i+\frac{500}{370}\sum_{i\in E_2} y_i+\frac{400}{500}\sum_{i\in E_3} y_i\right)$$

$$\hat{\mu} = \frac{300}{1200}\,\bar{y}_1+\frac{500}{1200}\,\bar{y}_2+\frac{400}{1200}\,\bar{y}_3\,.$$

Or $\frac{300}{1200} = 25\% = \frac{N_1}{N}$,… , et donc $\hat{\mu} = \sum_{h=1}^{3} \frac{N_h}{N}\,\bar{y}_h\,.$

C'est l'écriture de l'estimateur de stratification mais ici les tailles d'échantillons n_h sont aléatoires.

Stratification ***Post-stratification***



H échantillons aléatoires simples indépendants 1 échantillon aléatoire simple

F. LA METHODE EMPIRIQUE DES QUOTAS

1. INTRODUCTION

Avantage

L'immense avantage de la méthode des quotas est qu'elle ne nécessite pas de disposer d'une base de sondage d'où, comparativement à un sondage aléatoire de même type, un très faible coût et une très grande rapidité.

Inconvénient

L'inconvénient est qu'il n'est pas possible de calculer la précision des estimateurs obtenus. Cependant, il est courant de les voir accompagnés des calculs de la précision qui aurait été obtenue avec un échantillon aléatoire. Une personne non avertie ne voit aucune différence entre les deux méthodes si ce n'est le coût !

A condition d'être utilisée avec de grandes précautions (Cf. § 3), cette méthode donne des résultats satisfaisants.

2. Un ersatz de sondage aléatoire stratifié proportionnel

La méthode des quotas consiste à constituer de façon empirique un échantillon ***représentatif*** de la population suivant certaines variables. En effet, la seule contrainte imposée à l'enquêteur est de respecter certains quotas établis suivant les valeurs de quelques variables, dites ***de contrôle***, faciles à identifier sur la population. Le tableau suivant fournit un exemple de sondage au 1/50ème, représentatif de la population pour le sexe, l'âge et la C.S.P..

		POPULATION	ÉCHANTILLON
SEXE	Hommes	12 500	250
	Femmes	12 500	250
	Ensemble	25 000	500
ÂGE	< 25	10 000	200
	25 à 55	12 500	250
	>55	2 500	50
	Ensemble	25 000	500
CSP	Cad. Sup.	5 000	100
	Cad. Moy.	5 000	100
	Ouvriers	7 500	150
	Employés	7 500	150
	Ensemble	25 000	500

Si on a 10 enquêteurs et que chacun réalise 10 enquêtes par jour, chacun aura pour la semaine à enquêter 50 personnes réparties de la façon suivante : 25 hommes, 25 femmes, 20 de moins de 25 ans , 25 de 25 à 55 ans, 5 de plus de 55 ans, 10 cadres supérieurs, 10 cadres moyens, 15 ouvriers, 15 employés.
Des quotas croisés seraient préférables mais ce n'est pas toujours possible et c'est plus difficile à gérer. L'enquêteur doit savoir raisonner à l'intérieur de ses quotas pour éviter d'avoir à « boucler » son échantillon avec 5 femmes de moins de 25 ans cadres supérieurs.

Finalement, la méthode des quotas propose de constituer un ***échantillon stratifié proportionnel*** mais de façon empirique.

3. CONDITIONS D'UTILISATION

Par analogie avec le sondage aléatoire stratifié représentatif, les résultats seront d'autant plus satisfaisants que les variables de contrôle sont fortement liées à l'objectif de l'enquête.

Cependant, ***comme toute méthode empirique, des biais importants peuvent être dus à des variables qui influent fortement sur les réponses et auxquelles on ne pense pas a priori.***

Par exemple, une enquête par quotas sur l'utilisation des transports en commun, entièrement réalisée en plein centre ville, a donné des résultats inutilisables. Il

est évident qu'une variable essentielle à prendre en compte est la répartition selon les quartiers de la ville.

Un autre exemple est le suivant. Lors d'une enquête par quotas auprès de personnes âgées, on s'aperçoit qu'aucune ne juge pénible de monter les escaliers. Après vérification, elles habitaient toutes au rez-de-chaussée. Si les enquêteurs peuvent boucler leurs quotas sans monter des escaliers, pourquoi le feraient-ils ?

Une enquête sur la santé est réalisée par quotas auprès de la population de plus de 15 ans d'une ville. On réalise qu'il y a dans l'échantillon une proportion anormalement élevée de personnes malades : l'enquête avait eu lieu à domicile pendant les heures de bureau !

Ces exemples peuvent paraître évidents et l'on peut se sentir à l'abri de telles erreurs. Prenons alors encore un exemple. En 1968, deux enquêtes sur l'intention d'achat des ménages en biens d'équipements ont été réalisées. L'une par la SOFRES par la méthode des quotas suivant l'âge, le sexe et la CSP du chef de ménage, l'autre par l'INSEE à partir d'un échantillon aléatoire stratifié de ménages.
Les résultats sont assez différents. Ayant tout à fait confiance à la méthode aléatoire, on cherche quelle est la variable de contrôle qui aurait due être prise en compte dans la méthode des quotas. Il s'avère qu'il s'agit du nombre de personnes dans le ménage.
On obtient sur cette variable les résultats suivants :

Nbre de personnes	1	2	3	4	5	+	Ensemble
INSEE	19.1	27.3	19.1	15.9	8.9	9.5	100
SOFRES	8.5	23.8	19.8	19.7	12.3	15.9	100

Après redressement de l'échantillon de la SOFRES sur cette variable les résultats des deux enquêtes sont assez proches.

Biens	INSEE	SOFRES
Automobile	52.7	53.1
Télévision	58.0	53.1
Réfrigérateur	68.5	68.7
Machines à laver	47.5	44.9
Aspirateur	49.8	50.3

La méthode des quotas est désormais validée par la méthode aléatoire ; elle peut être réutilisée, avec évidemment comme variable de contrôle supplémentaire "le nombre de personnes dans le ménage".

La méthode des quotas est très souvent utilisée au dernier degré d'un sondage complexe.
Par exemple, pour un sondage auprès de la population française, on peut construire un échantillon aléatoire stratifié selon les régions, le type d'agglomération, la commune, le secteur d'habitation et utiliser, dans les secteurs d'habitation retenus de façon aléatoire, un échantillon par quotas des individus à enquêter.

Les consignes données aux enquêteurs sont alors très nombreuses. En plus des quotas à respecter, il aura à répartir les enquêtes selon les heures de la journée, à respecter des procédures définies a priori pour le choix du logement, selon le nombre d'étages, le nombre de logements à l'étage retenu, etc. Enfin, lorsqu'un logement est retenu, le choix de l'individu dans le logement peut également être déterminé selon une procédure définie a priori : par exemple, l'individu dont l'anniversaire est le plus proche. Toutes ces contraintes sont imaginées de façon à ce que l'échantillon empirique ressemble le plus possible à un échantillon aléatoire, et que l'enquêteur n'ait à aucun moment le loisir de choisir telle personne plutôt que telle autre !

De même, lors d'une enquête à la sortie d'un musée par exemple, il faudra définir une règle pour le choix de l'échantillon : par exemple, dès qu'un enquêté a fini de répondre, l'enquêteur doit s'adresser à la 4ème personne qui se présente !

3. LA CONSTRUCTION DU QUESTIONNAIRE

A. Introduction

B. Les modes de passation du questionnaire

C. Les étapes de la construction du questionnaire

D. Les types de questions

E. La formulation des questions

F. L'architecture du questionnaire

A. LE QUESTIONNAIRE : INTRODUCTION

Le questionnaire est l'instrument de mesure et, comme tout instrument de mesure, il peut avoir un effet sur ce qu'il mesure. Comme il a été dit précédemment, de nombreuses études ont été réalisées pour mesurer ces effets. Ces effets sont autant de sources d'erreurs possibles qu'il convient d'éviter lors de la construction du questionnaire.

B. MODE DE PASSATION DU QUESTIONNAIRE

Le questionnaire peut être administré par un enquêteur (***en face à face*** ou ***par téléphone***) ou bien autoadministré (***par la poste***, par exemple).

Dans le cas de questionnaires administrés par des enquêteurs, il est essentiel de préparer avec beaucoup d'attention ces enquêteurs. Ils doivent bien comprendre l'objectif de l'enquête et du questionnaire, bien comprendre chaque question et les poser telles qu'elles sont rédigées, respecter scrupuleusement les consignes données. Il faut évidemment qu'ils travaillent sérieusement et qu'ils connaissent les éléments mis en place pour les contrôler.

Il existe à présent des aides informatiques à la passation du questionnaire par enquêteur ; il s'agit des systèmes CAPI (Computer Assisted Personal Interview) pour les enquêtes en face à face et CATI (Computer Assisted Telephone Interview) pour les enquêtes par téléphone. Ces outils permettent de réaliser simultanément le recueil des données, leur codage et leur saisie. Ils proposent de plus des présentations aléatoires des modalités possibles afin d'éviter le biais dû à l'ordre de ces modalités. Lors d'une réponse incompatible avec les réponses précédentes ou d'une erreur de saisie, des tests de cohérence et de vérification permettent d'en avertir immédiatement l'enquêteur.

L'enquête postale est peu coûteuse mais elle ne peut pas être administrée à toutes les populations. Le questionnaire doit être particulièrement bien construit puisqu'une incompréhension ne peut pas être éclaircie par un enquêteur. Pour obtenir un bon taux de réponse, il faut accompagner le questionnaire d'une lettre explicative et motivante et fournir une enveloppe pour la réponse dispensée d'affranchissement. Il faut également pratiquer des relances.

C. LES ETAPES DE LA CONSTRUCTION DU QUESTIONNAIRE

Les différentes étapes de la construction du questionnaire sont les suivantes.

1 - définir précisément l'objectif du questionnaire, à qui il s'adresse, dans quelles conditions il est soumis à l'enquêté,

2 - réaliser une étude qualitative préalable ; il s'agit d'organiser des entretiens de quelques personnes issues de la population concernée afin de préciser la manière d'aborder les questions, le vocabulaire utilisé,

3 - rédiger une première version du questionnaire

- structurer le questionnaire en fonction de l'objectif (questions d'identification, thème 1, thème 2,...), renvoyer les questions sensibles en fin de questionnaire, organiser les questions filtres et les renvois à des blocs de questions.
- penser à la codification, la saisie et le traitement statistique (tableaux, sous-populations d'intérêt).
- respecter les règles qui suivent, concernant la formulation des questions et l'articulation du questionnaire
- penser aux documents annexes à fournir (présentation, explications), soit à l'enquêté, soit à l'enquêteur,

4 - tester le questionnaire auprès de quelques personnes issues de la population concernée, afin de vérifier si les questions sont bien comprises par les enquêtés, si le questionnaire « passe » bien,

5 - rédiger une version définitive du questionnaire.

D. LES TYPES DE QUESTIONS

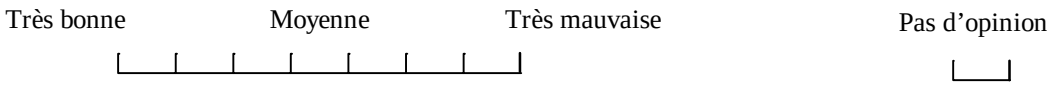
On distingue, quant au fond, deux types de questions :
- les questions concernant les faits ("***questions objectives***"),
- les questions concernant des opinions, des attitudes, des motivations, des préférences ("***questions subjectives***").

On distingue, quant à la forme, deux types de questions :
- les ***questions fermées*** c'est-à-dire pour lesquelles l'ensemble des réponses possibles est proposé (***étape confirmatoire*** du sujet de l'étude),
- les ***questions ouvertes*** pour lesquelles il n'est pas possible a priori d'envisager la diversité des réponses (***étape exploratoire*** du sujet), L'enquêté répond alors librement à la question.

Pour les ***questions fermées***, la réponse peut être unique, à modalités ordonnées ou non. On a une réponse unique lorsque la réponse est un nombre. Par exemple l'âge ou le nombre d'enfants.

Exemple de réponse unique à modalités ordonnées :

Opinion sur tel homme politique.



Faut-il proposer ou non une modalité « pas d'opinion » ?
Faut-il proposer ou non une modalité centrale, donc un nombre impair de modalités ?
Sur ces deux questions, les avis sont partagés. Le fait de proposer une modalité « pas d'opinion » et une modalité centrale permet de distinguer :
- ceux qui se prononcent (favorablement ou défavorablement)
- ceux qui ont une opinion moyenne
- ceux qui n'ont pas d'opinion
- ceux qui refusent de répondre à la question.

D'après les dernières expériences, il semble que la présence ou non d'une modalité « pas d'opinion » ou d'une modalité centrale ne modifie pas les pourcentages de ceux qui ont choisi une autre modalité (opinion favorable ou

défavorable) par rapport à ceux qui se sont prononcés. (Cf. Grémy in Grangé et Lebart, 1992).
La réponse peut être multiple, avec ou sans classement, avec ou sans indication du nombre de réponses à donner.

Exemple sans classement et sans indication sur le nombre de réponses à donner :
« Concernant le rôle d'un professeur, quels sont, parmi les énoncés suivants, ceux dont vous vous sentez le plus proche ? » (Cochez la case correspondant à chacun de vos choix).

énoncé 1 ○

énoncé 2 ○

....

énoncé 8 ○

Le traitement statistique se fera alors à l'aide de 8 variables indicatrices, une pour chaque énoncé. L'interprétation des réponses à ce type de question n'est pas toujours aisé.

Exemple avec classement et avec indication du nombre de réponses à donner :
« Concernant le rôle d'un professeur, indiquez par ordre de préférence 1, 2, 3, les trois énoncés parmi les énoncés suivants dont vous vous sentez le plus proche. »

énoncé 1 ○

énoncé 2 ○

.....

énoncé 8 ○

Ces questions, appelées aussi ***questions de préférence***, sont particulièrement difficile à interpréter. En effet, à partir des préférences individuelles il n'est pas possible d'établir une préférence collective : c'est le célèbre paradoxe de Condorcet sur l'agrégation des préférences.

Supposons qu'il y ait trois énoncés A, B, C et trois individus qui classent les énoncés de la façon suivante :

	Énoncé A	Énoncé B	Énoncé C
Individu 1	1	2	3
Individu 2	3	1	2
Individu 3	2	3	1

Si l'on veut comparer les trois énoncés sur l'ensemble des trois individus on a :

A placé devant B par 2 voix contre 1
B placé devant C par 2 voix contre 1
C placé devant A par 2 voix contre 1 !

Le traitement des ***questions ouvertes*** peut se faire de deux manières :

- soit en les fermant a posteriori (post codage) ; on repère les types de réponse revenant le plus souvent et on crée, pour chaque type de réponse, une variable indicatrice (présence, absence) de cette réponse,
- soit en utilisant, sur les réponses telles qu'elles sont transmises, des logiciels spécialisés dans le traitement statistique de données textuelles (SPAD-T ou Alceste, par exemple).

Il existe aussi des ***questions semi-ouvertes***, où seule la dernière modalité de réponse est de la forme :

○ Autre modalité. Précisez.

Cf. en annexe 4 le codage du questionnaire en fonction des types de question.

E. LA FORMULATION DES QUESTIONS

1. Vérifier la structure logique, éviter les formulations interro-négatives.
Ex : "Êtes-vous d'accord pour désapprouver ceux qui s'opposent au projet de loi sur le code de la nationalité ?"

2. Vérifier que la question soit comprise par l'enquêté (en utilisant éventuellement une question filtre), qu'elle soit comprise de la même façon par tous les enquêtés, éviter les termes techniques, TVA, CEE, SMIC, TUC, SME, ou les termes vagues.
Ex : "Avez-vous lu un quotidien hier ? "
L'interprétation du verbe lire peut être très différente d'un individu à l'autre.

3. Vérifier que l'enquêté puisse répondre sans difficulté pratique (question de mémoire, de typologie, de finalité, ...).
Ex : "En quelle année avez-vous acheté votre téléviseur ?"
"Quelle a été votre consommation totale de surgelés ce mois-ci ?"
Les enquêteurs utilisent parfois, pour les enquêtes sur le budget "dépenses" ou le budget "temps", des carnets de décomptes que les enquêtés doivent tenir pendant plusieurs semaines. Mais la tenue de tels carnets n'entraîne-t-elle pas un changement de comportement des enquêtés ? Les personnes acceptant une telle contrainte sont-elles « représentatives » de l'ensemble de la population ?
"Êtes-vous actif au sens de l'INSEE ? Je vous rappelle la définition : ... " (suivent 10 lignes !)
"Avez-vous cessé de boire ?" (question éminemment difficile)
"Pour réduire le trafic de drogue, pensez-vous que les douaniers devraient être plus sévères dans les aéroports ?". Que ne ferait-on pas pour réduire le trafic de drogue ? Il y a ici deux idées dans la même question.

4. Vérifier que la question soit acceptable par l'enquêté, qu'il veuille y répondre sincèrement. Quelle confiance accorder à la promesse d'anonymat statistique pour des questions sensibles, voire réprimées par la loi ?
Ex : "Avez-vous avorté durant les 12 derniers mois ?"
Pour ce type de question on pourra faire appel à la méthode des ***réponses aléatoires*** (ou anonymisées) :
En l'absence de l'enquêteuse, l'enquêtée jette un dé : si le dé marque 1 ou 2 elle donne la réponse correcte à la question, si le dé marque 3, 4, 5 ou 6 elle donne la réponse fausse à la question. Ainsi l'enquêteuse ne peut pas savoir si l'enquêtée a ou non avorté au cours des douze derniers mois, mais ce dispositif permet d'estimer le pourcentage des femmes l'ayant fait. Une telle enquête a donné un pourcentage de 3.2% alors que des enquêtes parallèles en face à face et par questionnaire anonyme ont donné respectivement 0.3% et 0.8%.

Éviter les termes chargés de jugement de valeur : l'enquêté peut vouloir impressionner favorablement l'enquêteur. Il est courant d'observer une sur-consommation de livres, de concerts, de théâtres, savon, dentifrice, et une sous-consommation d'alcool, de tabac, de drogue, de revues pornographiques,...
Ex : "Avez-vous l'intention de prendre une assurance sur la vie ?"
"Connaissez-vous le célèbre François Duval ?" (illustre inconnu qui amène 16% de réponses positives).
"Avez-vous voté aux élections partielles de décembre dernier ?" (33% de réponses positives alors qu'il n'y a pas eu d'élections).
"Avez-vous entendu parlé du projet de loi sur...?" (53% de oui sur un projet imaginé pour les besoins de l'enquête).
"Pourquoi n'avez-vous pas voté ?" Les réponses arrivent dans l'ordre : 1) force majeure, 2) raisons diverses, 3) manque d'intérêt.
"A votre avis, pourquoi les abstentionnistes ne votent-ils pas ?" Les réponses arrivent dans l'ordre 3),2),1).

5. Réfléchir au choix des mots.

Ex :"Doit-on permettre les discours contre la démocratie ?" (oui 21%, non 62%, sans opinion 17%)

"Doit-on interdire les discours contre la démocratie ?" (non 39%, oui 46%, sans opinion 15%)

"Approuvez-vous la création de dispensaires où les femmes pourraient être renseignées sur les moyens à employer :

pour éviter la grossesse ?" (oui 71%)

pour avoir le nombre d'enfants qu'elles désirent ?" (oui 83%)

6. Éviter qu'il y ait une même opinion pour des raisons différentes.

Par exemple : « Êtes-vous pour l'emprisonnement à perpétuité ? »

Peuvent répondre « non » à cette question des personnes qui pensent qu'il faut laisser un espoir même au pire criminel et des personnes qui sont contre la prison à vie car ils sont pour la peine de mort !

7. Pour les questions fermées, veiller à ce que toutes les modalités de réponse soient proposées et qu'elles soient exclusives.

F. L'ARCHITECTURE DU QUESTIONNAIRE

1. Éliminer les questions inutiles, redondantes, celles qui amènent nécessairement 100% de oui ou de non.
Ex : "Voulez-vous davantage de crèches ?"

2. Vérifier l'ordre des questions :

Faire attention au biais d'acquiescement
Il faut savoir qu'une grande partie de la population a tendance à répondre « oui » pour ne pas contrarier l'interlocuteur, ou « d'accord » pour ne pas avoir à discuter. Cette tendance est d'autant plus forte que les personnes ont un plus faible niveau d'éducation.
Si, les réponses oui ou non semblent s'imposer, pour des questions d'opinions par exemple, une même catégorie d'individus ne doit pas avoir à répondre toujours oui.

Faire attention au biais de réponse sur liste, l'ordre de présentation des réponses a un effet sur les réponses. Dans l'exemple suivant (Grémy in Grangé et Lebart, 1992), un questionnaire a été présenté sous deux formes, chacune proposant dans un ordre inverse l'un de l'autre les réponses proposées à la question suivante :
« Selon vous, quel est aujourd'hui le problème le plus grave ? »

Réponses	Ordre direct 1186 individus	Ordre inverse 994 individus
le chômage	34.2%	19.2%
le terrorisme	8.5%	7.0%
la faim dans le monde	21.3%	17.7%
la guerre	15.3%	18.3%
la surpopulation	1.5%	1.7%
le racisme	5.0%	8.5%
le non respect des droits de l'homme	8.0%	10.4%
l'insuffisante formation professionnelle des jeunes	1.3%	6.2%
la délinquance	1.2%	8.2%

Faire attention à ***l'effet de halo*** c'est-à-dire l'influence d'une question sur la question suivante,

Ex : "Pensez-vous que la grande criminalité soit en progression ?"

"Êtes-vous favorable au rétablissement de la peine de mort ?"

De façon générale l'enquêté cherche a avoir un discours cohérent. Pour un sondage pré-électoral par exemple, lorsqu'il change d'intention de vote, l'enquêté a tendance à falsifier la réponse concernant son dernier vote pour ne pas montrer qu'il a changé d'avis.

3. Veiller au rythme du questionnaire, ne pas passer du coq à l'âne, éviter les questions d'arrêt sur lesquelles l'enquêté peut hésiter à répondre, reporter en fin de questionnaire les questions sensibles. Lorsque l'enquêté réalise qu'il répond à un questionnaire, il peut avoir envie d'arrêter.

4. Vérifier que le questionnaire ne soit pas trop long, ennuyeux, difficile, indiscret, partial.

5. Évaluer l'adéquation du questionnaire à son objectif.

On renvoie le lecteur à de Singly (1992) pour une présentation plus détaillée de la construction du questionnaire.

TABLE DE PRÉCISION ABSOLUE À 95%
DE L'ESTIMATION D'UNE PROPORTION

Précision absolue à 95% : $2\sqrt{\frac{p(1-p)}{n}}$ exprimée en %.

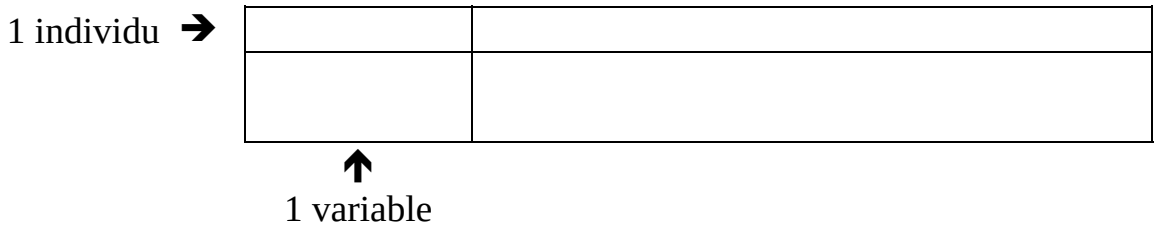
Prop. observée *p* Taille échant. *n*	5% ou 95%	10% ou 90%	15% ou 85%	20% ou 80%	25% ou 75%	30% ou 70%	35% ou 65%	40% ou 60%	45% ou 55%	**50%**
100	4.36	6.00	7.14	8.00	8.66	9.17	9.54	9.80	9.95	**10.00**
150	3.56	4.90	5.83	6.53	7.07	7.48	7.79	8.00	8.12	**8.16**
200	3.08	4.24	5.05	5.66	6.12	6.48	6.75	6.93	7.04	**7.07**
250	2.76	3.79	4.52	5.06	5.48	5.80	6.03	6.20	6.29	**6.32**
300	2.52	3.46	4.12	4.62	5.00	5.29	5.51	5.66	5.74	**5.77**
350	2.33	3.21	3.82	4.28	4.63	4.90	5.10	5.24	5.32	**5.35**
400	2.18	3.00	3.57	4.00	4.33	4.58	4.77	4.90	4.97	**5.00**
500	1.95	2.68	3.19	3.58	3.87	4.10	4.27	4.38	4.45	**4.47**
600	1.78	2.45	2.92	3.27	3.54	3.74	3.89	4.00	4.06	**4.08**
700	1.65	2.27	2.70	3.02	3.27	3.46	3.61	3.70	3.76	**3.78**
800	1.54	2.12	2.52	2.83	3.06	3.24	3.37	3.46	3.52	**3.54**
900	1.45	2.00	2.38	2.67	2.89	3.06	3.18	3.27	3.32	**3.33**
1000	1.38	1.90	2.26	2.53	2.74	2.90	3.02	3.10	3.15	**3.16**
1500	1.13	1.55	1.84	2.07	2.24	2.37	2.46	2.53	2.57	**2.58**
2000	0.97	1.34	1.60	1.79	1.94	2.05	2.13	2.19	2.22	**2.24**
3000	0.80	1.10	1.30	1.46	1.58	1.67	1.74	1.79	1.82	**1.83**
5000	0.62	0.85	1.01	1.13	1.22	1.30	1.35	1.39	1.41	**1.41**
10000	0.44	0.60	0.71	0.80	0.87	0.92	0.95	0.98	0.99	**1.00**

LE CODAGE

① L'OBJECTIF

Questionnaire avec les réponses ➔ Réponses codées pour la saisie informatique

Il s'agit d'obtenir un tableau rectangulaire (matrice des données).



② LE PROBLÈME

A une question peut correspondre une variable ou plus.

Le nombre de variables dépend du type de question.

③ LES DIFFÉRENTES POSSIBILITÉS

- A une question on répond par un nombre.

Ex : Année de naissance, mesure, prix, ...

On code la réponse par le nombre indiqué.
Ensuite il est possible de recoder en classes pour obtenir une variable catégorielle.

1 question ➔ 1 variable réelle

- Question fermée avec réponse unique

Toutes les réponses possibles ont été prévues et on ne peut donner qu'une seule réponse (ne pas oublier dans la question, si nécessaire, l'option "Autre" ou "Ne sait pas").

Question : "Où habitez-vous ?"

			Codage
Centre ville	☑	1	**1**
Banlieue	☐	2	
A l'extérieur de la ville	☐	3	

1 question ➔ 1 variable catégorielle

- Question fermée avec réponses multiples

On peut donner plusieurs réponses.

On crée une variable pour chaque réponse possible, appelée variable indicatrice de présence absence (1 ou 0).

Question : "Quels sont les sports que vous aimez ?"

			Codage
Foot Ball	☐	Sp1	**0**
Natation	☐	Sp2	**0**
Tennis	☑	Sp3	**1**
Rugby	☑	Sp4	**1**
Equitation	☐	Sp5	**0**
Autre	☑	Sp6	**1**

1 question ➔ 6 variables indicatrices

Autant de variables que de réponses possibles

- Question avec réponses multiples ordonnées

On peut donner plusieurs réponses en les classant (en général par ordre de préférence).

Question : " Quels sont vos trois sports préférés parmi les sports suivants (noter 1, 2, 3 par ordre de préférence) ?"

			Codage		
Foot Ball	☐	1	Sp1	Sp2	Sp3
Natation	☐	2	**3**	**5**	**6**
Tennis	**1**	3			
Rugby	☐	4			
Equitation	**2**	5			
Autre	**3**	6			

1 question ➔ 3 variables catégorielles à 6 modalités chacune

Il est difficile d'exploiter ce genre de question.

- Question ouverte

Réponse sous forme de commentaire libre.

On peut procéder alors à un post-codage.

1/ Lecture de plus ou moins 50 réponses et identification des thèmes qui se répètent

2/ Création, pour chaque thème, d'une variable de présence absence (1, 0) du thème considéré.

1 question ➔ autant de variables indicatrices que de thèmes identifiés

Remarque : il existe des logiciels spécialisés (SPAD-T) qui ne nécessitent pas de post-codage pour les questions ouvertes. Il faut alors saisir la réponse dans sa globalité.

BIBLIOGRAPHIE SUCCINCTE

ARDILLY, P. (1994) *Les techniques de sondage*, Technip, Paris.

BROSSIER, G. et DUSSAIX, A.M. (1999). *Enquêtes et sondages*, Dunod, Paris.

DE SINGLY, F. (1992). *L'enquête et ses méthodes : le questionnaire,* Nathan, Paris.

DUSSAIX, A.M. et GROSBRAS, J.M. (1993) *Les sondages : principes et méthodes*, P.U.F., Coll. Que sais-je ? n° 701, Paris.

GRANGÉ, D. et LEBART, L. (eds) (1993) *Traitement statistique des enquêtes*, Dunod, Paris.