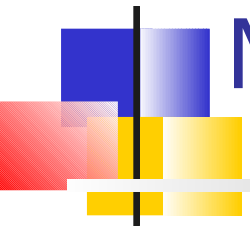
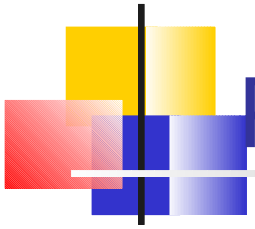


Technique d'Enquête et Méthode de Sondage

Matthieu NEVEU

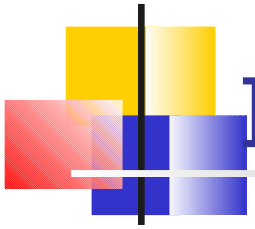
Licence d'Économétrie

Année 2004-2005



Références

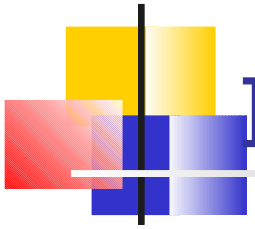
- Ouvrages généraux
  - Brossier G et Dussaix AM (1999). Enquêtes et Sondages. Dunod
  - Ardilly Pascal (1994). Les techniques de sondage.
  - Arkin Herbert (1963). Handbook of sampling for auditing and accounting.
  - Gouriéroux C. (1981). La théorie des sondages, Economica.
  - Rea Louis (1992). Designing and conducting survey research : a comprehensive guide.
  - Tillé Y (2001). Théorie des sondages. Dunod
- Echantillonnage
  - Beaud J. P. (1997). L'échantillonnage, in : Gauthier, op.cit.
  - Satin A., Shastry W (1983). L'échantillonnage : un guide non mathématique. Statistique Canada, Ottawa. 69 p.
- Questionnaires
  - De Singly (2001). L'enquête et ses méthodes : le questionnaire. Nathan Université. 127 p.
  - Gauthier (1997). Recherche sociale : de la problématique à la collecte des données. Presses de l'Université du Québec. 529 p.
  - Sudman, Seymour, Bradburn, Norman (1987). Asking questions : a practical guide for questionnaire design. Jossey -Bass, San Francisco. 397 p.
- Manuels de statistiques
  - Bouget D., Viénot A. (1995). Traitement de l'information : statistiques et probabilités.
  - Melton J.S., Arnold C.J. Introduction to probability and statistics, Mc Graw-Hill, International Edition.
  - Wonnacott R. et Wonnacott T. (1999). Statistiques. Economica.



Introduction

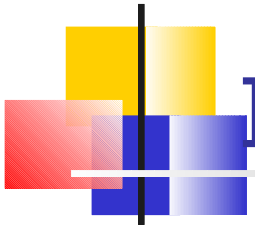
- Champ d'exercice des sondages :
  - Enquêtes sur les intentions de vote, baromètres et côtes de popularité, mesures d'audience…
  - Information économique et sociale (démographie, conditions de vie, emploi, consommation, santé, éducation, transports, loisirs, logement, prix…).
  - Contrôle des comptes : Conseil Supérieur de l'ordre des experts comptables et comptables agréés.
  - Contrôle de qualité : fabrication et réception de produits industriels.

- Acteurs
  - CNIS (Conseil National de l'Information Statistique)
  - INSEE (secrétariat du CNIS)
  - Organismes publics et para-publics mènent un grand nombre d'études et de recherches.
  - Entreprises privées : études de marché, analyses de clientèle, activités de conseil, sondages politiques et études d'opinion : Nielsen, SECODIP (panels), BVA, IFOP, IPSOS, ISL, SOFRES, Médiamétrie (audience des médias), CESP (étude des supports publicitaires)…



Introduction - Définition -

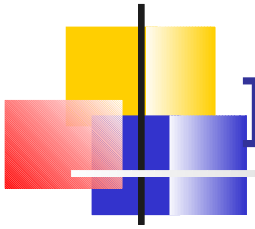
- **L'Enquête**
  - Est une recherche d'information.
- **Le Recensement**
  - Est une collecte d'information sur la totalité des personnes concernées par l'enquête.
- **L'Échantillonnage**
  - Est une opération consistant à sélectionner une fraction d'une population plus vaste.
    - Sondages ou Tests
  - Représentativité de l'échantillon
- **L'analyse Statistique et l'analyse économétrique**



Introduction - Définition -

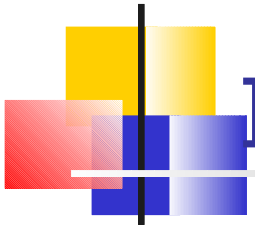
Avant tout travail, pensez à définir :

- **Le champs de l'enquête**
- **Les unités d'observations (ou individus)**
- **L'unité d'échantillonnage**
- **La base d'échantillonnage ou la base de sondage**
- **L'échantillonnage**
- **Variables d'intérêts**



Introduction - Etapes de l'Enquête par sondage -

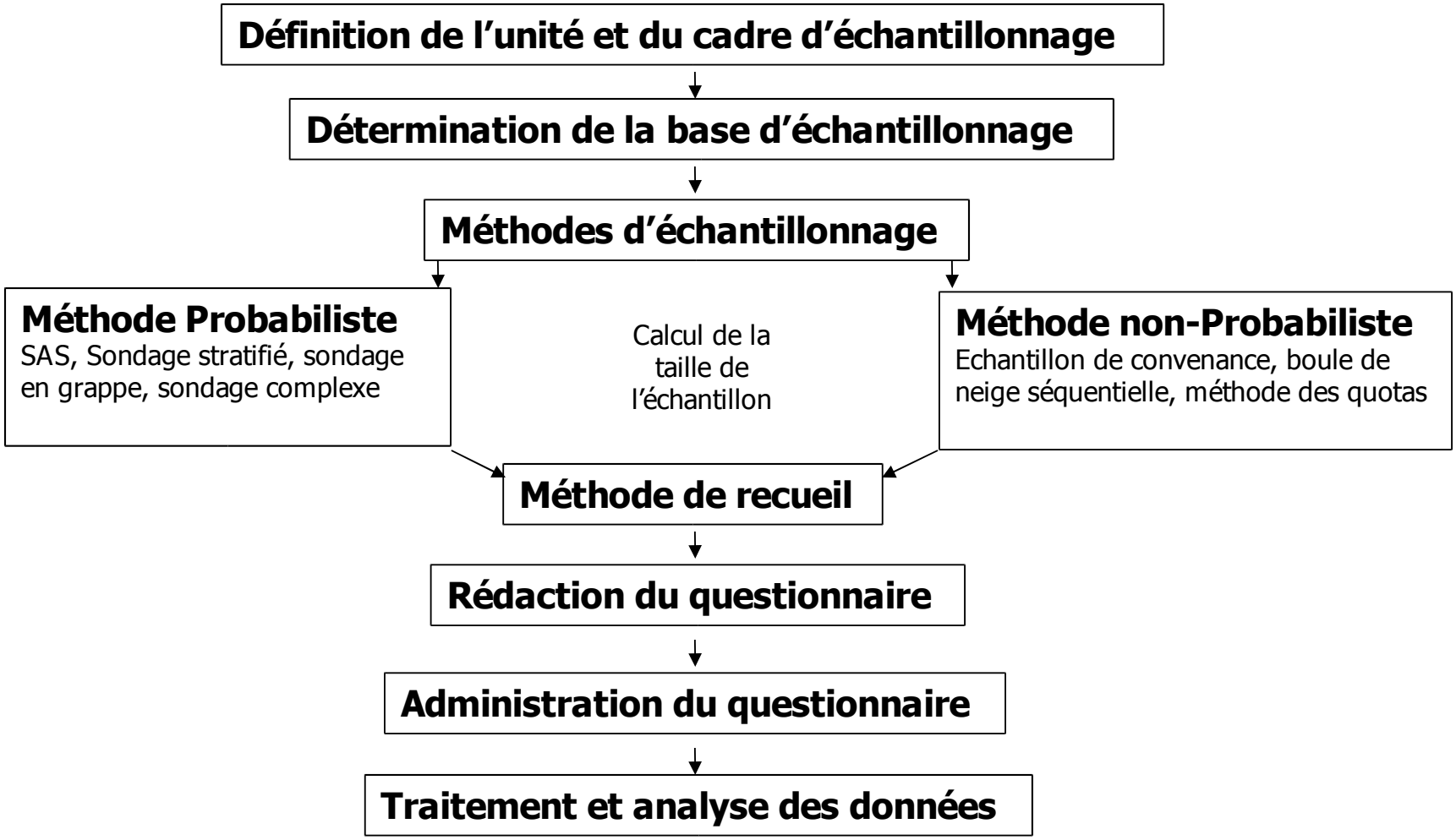
- **Pertinence** : s'assurer que les informations recherchées n'existent pas déjà.
- **Réflexion générale et théorique sur le sujet** : élaboration d'hypothèses qui seront confirmées ou infirmées par les observations d'enquête.
- **Faisabilité :** matériellement réalisable à un coût raisonnable. S'assurer que l'enquête produira des informations statistiques de bonne qualité.
- **Conception générale de l'enquête :**
  - Définition des objectifs de l'enquête
  - Détermination de la taille de l'échantillon selon budget et précision souhaitée.
  - Définition d'une technique de recueil de l'information
  - Définition simultanée de la méthode d'échantillonnage.

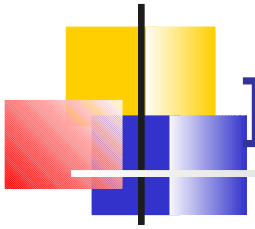


Introduction - Etapes de l'Enquête par sondage -

- **Rédaction du questionnaire :**
  - Première rédaction en utilisant, si possible, les résultats d'études exploratoires ou qualitatives préalables.
  - Pré-test du questionnaire
  - Rédaction définitive incluant le pré-codage
- **Administration du questionnaire :**
  - Réunion d'information des enquêteurs
  - Administration des questionnaires sur le terrain.
  - Contrôle de la qualité du travail des enquêteurs.
- **Traitement et analyse des données :**
  - Vérification de l'exhaustivité et de la vraisemblance des informations.
  - Repérage de valeurs aberrantes (apurement).
  - Codage de certaines variables (age, profession, éducation…).
  - Saisie informatique des questionnaires.
  - Traitement par logiciels spécialisés de traitement de données.

Introduction L'Échantillonnage : une succession d'étapes





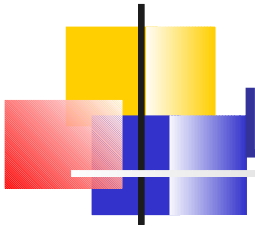
Introduction - Méthode d'Echantillonnage -

Méthode aléatoire ou probabiliste

- Le **sondage aléatoire simple**
- La **stratification**
- Le **sondage par grappes**
- Le **sondage à plusieurs degrés**

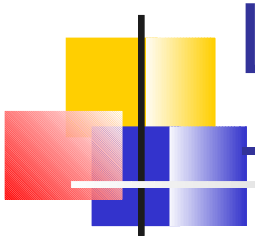
Méthodes empiriques ou « à choix raisonné »

- **Absence de base de sondage.**
- **Méthode des unités types**
- **Méthode des quotas**



L'Approche Probabiliste

- Partie de la théorie des sondages qui s'appuie sur la **théorie des probabilités**
- **Équiprobabilité des alternatives**
- **Le problème du tirage au hasard**
  - **Logique d'affinité** : risque que les unités d'échantillonnage suivent une logique d'affinité.
  - **Biais de sélection** : ce biais est lié à la population de référence.
- Lorsque les conditions sont réunies, il est possible d'engager un **processus d'échantillonnage probabiliste,** dont l'avantage essentiel est de pouvoir être évalué rigoureusement.
- Le modèle fondamental le plus simple est celui du **sondage aléatoire simple**

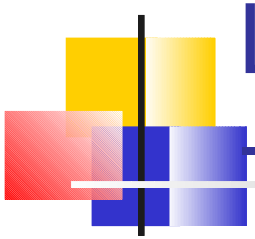


L'Approche Probabiliste
- le Sondage Aléatoire Simple (SAS) -

- **Modèle de référence, modèle le plus simple**

- **Procédure de tirage aléatoire d'une fraction de la population. Tous les échantillons sont possibles avec la même probabilité.**
  - Echantillon avec remise (peu répandu en pratique).
  - Echantillon sans remise.

- **Tirages avec remise** : Risque d'interroger plusieurs fois la même personne au lieu d'unités différentes.
  - Equiprobabilité : tous les individus ont la même probabilité (1/N) d'être choisis à chaque tirage.
  - Probabilité que l'individu α ne soit pas choisi au cours d'un tirage est 1 - 1/N.
  - Probabilité qu'il ne figure pas dans l'échantillon est : Pr ({α non choisi au 1er tirage} $\cap$ {α non choisi au 2ème tirage} $\cap$ …{α non choisi au n^{ième} tirage}).
  - Tout individu a la même probabilité de figurer dans l'échantillon. Lorsque N est grand, cette probabilité est peu différente du taux de sondage n/N.

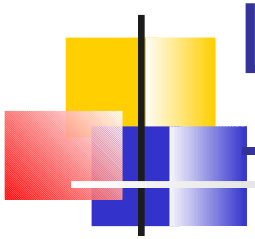


L'Approche Probabiliste
- le Sondage Aléatoire Simple (SAS) -

- **Tirages sans remise** : variables aléatoires non indépendantes.
  - Probabilité que l'individu α (α = 1 à N) appartienne à l'échantillon avant le ième tirage (i = 1 à n) :
    - i - 1 individus ont été tirés
    - N – (i-1) individus peuvent encore être tirés au cours du ième choix.
    - Probabilité que l'individu α n'ait pas été désigné au cours des i-1 premiers tirages, donc qu'il figure parmi les N-1 restants : (N-i+1)/N
    - Probabilité que l'individu α soit tiré au ième choix, sachant qu'il ne l'a pas été avant : 1/(N-i+1)
  - Probabilité que l'individu α soit choisi au ième tirage est égale au produit de la probabilité qu'il ne l'ait pas été avant par la probabilité qu'il le soit à ce ième tirage : 1/N
  - A chaque tirage, un individu a donc la même probabilité 1/N d'être choisi. La probabilité qu'il figure dans l'échantillon est :
    - Pr ({α choisi au 1er tirage} U {α choisi au 2ème tirage} U … {α choisi au nième tirage})
  - Les évènements {α choisi au ième tirage} étant incompatibles, cette probabilité est égale à la somme des probabilités {α choisi au ième tirage}, soit n/N.

- **Le sondage sans remise est donc représentatif** puisque chaque individu de la population peut figurer dans l'échantillon avec la même probabilité connue n/N.

- Si le taux de sondage f=n/N est inférieur à 0.05 (0.10 selon la précision souhaitée), **l'échantillon sans remise peut être assimilé à un échantillon avec remise.**



L'Approche Probabiliste
- le Sondage Aléatoire Simple (SAS) -

- **Le sondage, c'est l'observation de l'un de ces échantillons.**
  - Cette situation ne reflète que rarement la réalité. Dans la pratique, on essaie de restreindre le nombre de combinaisons en évitant celles qui seraient a priori non souhaitables.

- **Le SAS fournit un cadre de référence** indispensable pour deux raisons :
  - Jugement des autres modèles d'échantillonnage par rapport à ses propriétés. Il sert, en quelque sorte, d'étalon.
  - Il constitue la « brique » élémentaire. Ex. : les sondages stratifiés et les sondages à deux degrés sont des assemblages de sondages simples

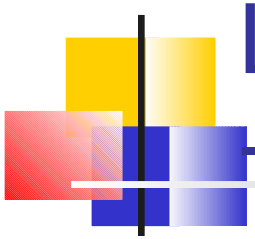
- Il est donc important d'en connaître parfaitement toutes les propriétés.

L'Approche Probabiliste
- le Sondage Aléatoire Simple (SAS) -

- **Exercices : Estimation d'une moyenne**
  - Clients d'une société bancaire.
  - Lancement d'un nouveau produit financier.
  - Variables d'intérêt : caractéristiques de la clientèle, ses motivations et ses réactions éventuelles.
  - Fichier de N titulaires de comptes.
  - Sondage sur un échantillon de n comptes parmi les N.

- **Hypothèses :**
  - N = 5 titulaires de comptes.
  - Echantillon de n = 2.
  - Dépôts sur ces comptes sont : 13, 15, 17, 25 et 30 milliers d'euros. La somme vaut 100 000 €.
  - L'organisme chargé de l'enquête ignore ces montants et se fixe pour objectif d'évaluer leur moyenne à partir des deux valeurs qu'il constatera sur l'échantillon.
  - Soient :
    - y_1 et y_2 les valeurs observées et
      $\bar{y}=\frac{y_1+y_2}{2}$ leur moyenne empirique qui est une variable aléatoire qui dépend de l'échantillonnage

- **Questions :**
  - Recenser les situations possibles dans le cas où l'échantillon est constitué d'unités distinctes (« sans remise »).
  - Calculer la moyenne $\bar{y}$ des 10 valeurs possibles (échantillons) et la moyenne $\bar{Y}$ des 5 valeurs des comptes (base de sondage).
  - Evaluer la dispersion des individus au sein de la population.



L'Approche Probabiliste
- le Sondage Aléatoire Simple (SAS) -

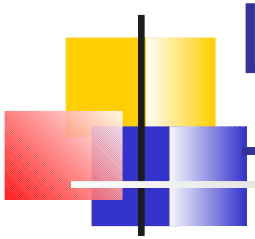
La précision dépend de 3 éléments :

- **la taille *n* de l'échantillon :** plus l'échantillon est grand, plus l'estimation est précise.

- **La variance de la variable d'intérêt :** plus une population est homogène (variance faible), plus le sondage est efficace.

  Si tous les individus sont caractérisés par des valeurs Y_i identiques, un seul suffit à les représenter. A l'inverse, sonder dans une population très hétérogène nécessite des échantillons de taille importante, ou un découpage préalable en sous populations homogènes (principe de stratification).

- **Le taux de sondage *f (=n/N) :*** si le taux de sondage est égal à 1, l'échantillon est la population entière et il n'y a plus d'erreur. Mais, dans la très grande majorité des sondages, les taux de sondage sont très faibles.



L'Approche Probabiliste
- le Sondage Aléatoire Simple (SAS) -

Moyenne, variance, erreur type :

- **La moyenne :** $\bar{y}=\frac{1}{n}\sum_{i=1}^{n} n_i y_i$
- **La variance :** $\sigma^2=\frac{1}{n}\sum_{i=1}^{n} n_i(y_i-\bar{y})$
- **La variance corrigée :** $S^2=\frac{n}{n-1}\sigma^2$
- **L'erreur type :** $\sqrt{v(\bar{y})}=\sqrt{\frac{1}{N}\sum_{i=1}^{N} N_i(Y_i-\bar{Y})}$

L'Approche Probabiliste
- le Sondage Aléatoire Simple (SAS) -

Intervalle de confiance :

- Pour estimer $\bar{Y}$, on procède comme suit :
- D'après la table de Gauss, 95% des valeurs possibles pour $\bar{y}$ se situent à une distance inférieure à $1{,}96\sqrt{V(\bar{y})}$, soit environ à moins de 2 erreurs-type de $\bar{Y}$
- Ayant obtenu la valeur de $\bar{y}$ par l'échantillon, on en déduit un intervalle contenant le paramètre $\bar{Y}$, avec une probabilité de 95% :

$$\bar{Y} \in \left[\bar{y} - 1{,}96\sqrt{V(\bar{y})}, \bar{y} + 1{,}96\sqrt{V(\bar{y})}\right]$$

- L'intervalle de confiance à 90% s'obtient en remplaçant 1.96 par 1.65, et dans l'intervalle de confiance à 99% par 2.58 (coefficients déterminés par la loi de Gauss).
- L'intervalle de confiance véritable fait intervenir la quantité $V(\bar{y})$ qui dépend de σ^2 et n'est donc pas calculable. Dans la pratique, il doit être évalué à partir de l'échantillon observé. La variance σ^2 peut être estimée par la variance corrigée des données recueillies : $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2$
- D'où, une estimation de $V(\bar{y})$ par $V(\bar{y}) = \frac{s^2}{n}$

L'Approche Probabiliste
- le Sondage Aléatoire Simple (SAS) -

- **Estimation d'une proportion**

- Une proportion est un cas particulier de moyenne.
- Construisons la variable qui, à l'individu i, fait correspondre la valeur Yi suivante :
  - Yi = si le client i a l'intention de souscrire au produit.
  - Yi = 0 sinon.
- La proportion P recherchée n'est autre que la moyenne des Yi . On peut donc mettre en œuvre les techniques exposées précédemment.
- Les calculs prennent une forme particulièrement simple:
  - si on note Q=1-P, alors la variance des Yi est égale à : σ^2=P-P^2= (1-P)P=QP.
- Les opérations d'estimation de P sont les suivantes :
  - l'estimateur ponctuel de P est la proportion p observée sur l'échantillon.
  - l'intervalle de confiance à 95% s'écrit :

$$P \in \left[p - 1.96\sqrt{(1-f)\frac{s^2}{n}}, p + 1.96\sqrt{(1-f)\frac{s^2}{n}} \right]$$

  - Ou de façon approchée : $P \in \left[p - 2\sqrt{\frac{qp}{n}}, p + 2\sqrt{\frac{qp}{n}} \right]$
  - Avec q = 1-p

L'Approche Probabiliste
- le Sondage Aléatoire Simple (SAS) -

Estimation d'une proportion …

- La **précision absolue (***PA) : 2*$\sqrt{qp}$. D'après l'échantillon, l'estimation de P se fait à plus ou moins *2*$\sqrt{qp}$ points.

- La **précision relative (***PR***)** vaut *PA/p*. Cela signifie que la marge d'incertitude est de l'ordre de PR de la quantité évaluée.

- Dans notre exemple,

  - La « fourchette » des résultats possibles pour P représentée par cet intervalle de confiance est plus ou moins large et correspond à une **estimation peu précise**.

  - **C'est la taille de l'échantillon qui est en cause :** l'intervalle de confiance est construit d'après l'écart-type, elle-même fonction de *n* comme on l'a vu.

  - En conséquence, pour diviser par deux la largeur de l'intervalle de confiance, il aurait fallu un échantillon de *n* = 800 clients au lieu de 200. Pour diviser encore par deux la fourchette, il aurait fallu *n* = 3200 interrogés…

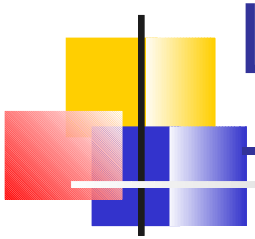
L'Approche Probabiliste
- le Sondage Aléatoire Simple (SAS) -

La taille de l'échantillon

- **« A partir de combien d'éléments un échantillon est-il valable ? »**
  - Si la question est ainsi posée, il n'y a pas de réponse directe. Cela dépend de la contrainte de budget plus ou moins forte.
  - Soit *C* est le budget maximum alloué à l'enquête et *c* est le coût unitaire de sondage, la taille maximale possible est : *C/c*.
  - Mais, cette taille peut être insuffisante pour assurer des résultats suffisamment fiables. La question qui se pose alors est :
- **« Quel budget faudrait-il consacrer pour garantir une précision acceptable ? ».**
  - Même dans ces termes, il n'y a pas de réponses toute faite. Il faut d'abord définir ce qu'on entend par ***précision acceptable.***
  - On peut convenir d'un écartement maximum toléré de l'intervalle de confiance, i.e. fixer une **borne à la précision absolue**

$$2\sqrt{(1-f)\frac{S^2}{n}} \cong 2\sqrt{\frac{\sigma^2}{n}}$$

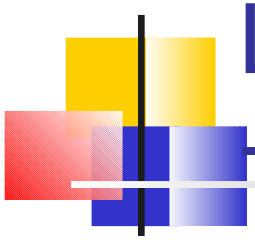
  - Ou bien fixer une **borne à la précision relative** $\frac{2}{\bar{y}}\sqrt{\frac{\sigma^2}{n}}$
- **La difficulté tient dans le fait qu'il faut avoir *a priori* une idée de l'ordre de grandeur des quantités qui doivent intervenir et de leur variance.**



L'Approche Probabiliste
- le Sondage Aléatoire Simple (SAS) -

- Il faut tenter d'obtenir l'information utile.
- On connaît les résultats d'une **enquête similaire réalisée dans un passé pas trop éloigné**, ses résultats peuvent permettre de calibrer l'enquête actuelle.
- Il y a, dans la base de sondage, des informations détaillées relatives à **une variable Z bien corrélée avec la variable Y de l'enquête**.
- On réalise l'enquête en deux phases : on prélève un premier échantillon pour évaluer grossièrement $\bar{Y}$ et σ^2, et on en déduit une taille souhaitable pour l'échantillon véritable.
- Ces situations ne sont pas exhaustives. Elles illustrent le 1er devoir de tout sondeur :

Mobiliser toute l'information disponible a priori
et pertinente au regard de l'enquête qu'il doit effectuer.



L'Approche Probabiliste
- le Sondage Aléatoire Simple (SAS) -

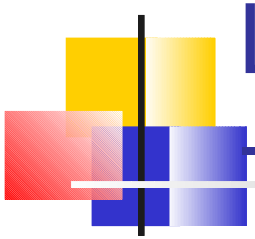
Hypothèse fondamentale du SAS :

- *Toutes* les combinaisons de *n* éléments parmi les N de la population sont réalisables avec la *même* probabilité. Chaque élément a la même chance que les autres d'être sélectionné. Il faut toujours **s'assurer que cette condition est bien vérifiée** sous peine d'utiliser un formulaire inadéquat.

- L'exemple des **sondages « sur place »** (enquête réalisées à la sortie des musées, de spectacles, de centres commerciaux…) montre que ce n'est pas toujours chose aisée :
  - S'il y a des variations d'affluence, et si le rythme des interviews est constant, on ne peut plus parler de sondages avec probabilités égales.
  - Si pendant la période A l'affluence est le double de celle de la période B :
  - Les personnes présentes en A ont deux fois moins de chance d'être interviewées qu'en B, sauf si, par exemple, on double le nombre d'enquêteurs en A.
  - Cela ne veut pas dire que le sondage soit mauvais, mais il faut traiter les observations de façon différente par des **pondérations** adéquates.

- Il est parfaitement légitime (et souvent souhaitable) de réaliser des sondages avec des **probabilités d'inclusion inégales** selon les individus de la population.

- Mais, le traitement des résultats doit en tenir compte et ce n'est pas celui du SAS (Attention donc au maniement sans précaution des logiciels de dépouillement d'enquête !).



L'Approche Probabiliste
- le Sondage Aléatoire Simple (SAS) -

- **Méthodes concrètes de réalisation**, lorsque la base de sondage est constituée par un fichier dont les unités sont identifiées par un numéro de 1 à N.
  - Tables de nombres au hasard.
  - Fonctions disponibles sur les ordinateurs et machines à calculer (« random ») générant des nombres entre 0 et 1.

- **Le tirage systématique** (méthode encore plus simple et très largement utilisée) :
  - numéroter les unités de 1 à N.
  - calculer le « pas » de sondage k = N / n.
  - choisir « au hasard » un départ d entier compris entre 1 et k.
  - l'échantillon est formé des unités identifiées par les numéros les plus proches de d, d+k, d+2.k, …, d+(n-1).k

- La simplicité du tirage systématique fait son succès. Mais, nécessité de **vigilance sur ses propriétés** :
  - **Si le rangement des unités dans le fichier** est indépendant de la variable d'intérêt, la méthode des tirages systématiques est un SAS.
  - **Si les unités sont tirées selon un ordre corrélé avec la variable d'intérêt** : stratification implicite. Le résultat peut être meilleur qu'un SAS au sens strict.
  - **Si périodicité dans le fichier et si le « pas » des tirages est égal à la période** (ou à un multiple) : possibilité de sélectionner des individus très particuliers.
- En pratique :
  - Le SAS n'est jamais utilisé seul.
  - Il ne suppose qu'une seule chose, mais essentielle : **l'existence d'une base de sondage**.
  - La stratification est la première méthode qu'il doit mettre en œuvre.