

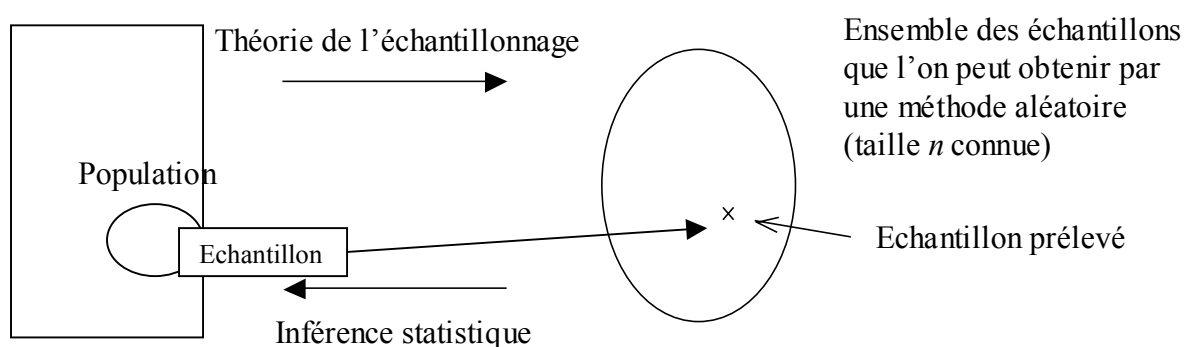
## 5 Estimation

### 5.1 Méthodes de sondage

#### 5.1.1 Jugement sur échantillon et inférence inductive

Ce cours est consacré aux techniques du jugement sur échantillon, par lesquelles nous essayons de « dire des choses » sur la population à partir de la seule observation d'un échantillon. Cette opération s'appelle encore *inférence statistique*. Le terme inférence désignant ici une inférence *inductive* allant du particulier au général.

La déduction est un raisonnement qui opère du général au particulier par le syllogisme (ou l'implication logique) : *si tous les oiseaux volent, alors les poules (qui sont des oiseaux) volent*. Si la prémisse est vraie la conclusion l'est également. Une induction au contraire est un raisonnement qui opère une généralisation d'une observation faite sur quelques cas. Par exemple : « *Je vais en Angleterre. Je vois une anglaise. Elle est rousse. Donc toutes les anglaises sont rousses* ». Ce dernier raisonnement conduit à une erreur évidente dans de nombreux cas comme l'avait déjà remarqué le philosophe anglais David Hume. Soyons moins pressé et répétons l'observation plusieurs fois : « *Je vais en Angleterre. Je vois dix anglaises. Quatre sont rousses. Donc 40% des anglaises sont rousses* », l'erreur de l'inférence inductive existe encore puisque recommençant la même observation de dix anglaises je trouverai vraisemblablement une autre fréquence, mais elle a de grandes chances d'être plus faible. Si l'échantillon observé est le résultat d'un sondage aléatoire (au hasard), alors nous pouvons préciser les propriétés de cette erreur d'échantillonnage grâce au calcul des probabilités. C'est le propos de la théorie de l'échantillonnage.



#### 5.1.2 Exhaustivité, représentativité

Jusqu'à la fin du XIX<sup>ème</sup> siècle, on opposait essentiellement dans les méthodes d'investigation la **monographie**, ou étude intensive d'un très petit groupe d'individus typiques telle qu'elle a été popularisée par Le Play, au **recensement statistique**, ou étude extensive et exhaustive de tous les individus. La seconde méthode seule pouvait donner des informations fiables sur les répartitions, sur le « combien », tandis que la première permettait de répondre plus finement à des questions relatives au « comment » et au « pourquoi ». Toute idée d'une étude partielle qui aurait permis de juger de la population à partir de l'étude d'un échantillon a été longtemps suspecte aux yeux des statisticiens.

On appelle **enquête par sondage** une enquête portant sur une partie d'une population finie appelée **échantillon**. On obtient alors des informations, appelées **informations imparfaites**, sur une ou plusieurs variables. Extrapoler les résultats obtenus dans l'échantillon à la population dont il est issu, c'est faire un raisonnement inférentiel inductif (par opposition à déductif), ce qui conduit à des erreurs, dites **erreurs d'échantillonnage**. Une méthode de sondage sera d'autant plus efficace que les erreurs d'échantillonnage seront plus faibles. On doit donc limiter ces erreurs, d'une part en augmentant la taille de l'échantillon, d'autre part en faisant en sorte que l'échantillon soit « représentatif », c'est-à-dire qu'il ressemble à la population dont il est issu, qu'il en soit une sorte de modèle réduit - *dans les mêmes proportions*. Les erreurs d'observation, à la différence des erreurs d'échantillonnage, sont souvent moins importantes lorsque le nombre d'observations est peu élevé. Pour cette raison, une enquête par sondage sera souvent préférée à une enquête exhaustive (recensement) dont on connaît les autres inconvénients : coût, délai et destruction possible des unités statistiques examinées. Le problème est de choisir la méthode de sondage qui donnera la meilleure information possible tout en respectant des contraintes, généralement de budget et de temps, c'est le propos de la **théorie des sondages**.

Les statisticiens n'ont admis « la méthode représentative » qu'à leur congrès (IIS) de 1925, et sous réserve que la partie étudiée soit représentative du « tout » qui seul est intéressant. Leur résolution distingue deux méthodes pour assurer la représentativité de l'échantillon :

- l'échantillonnage raisonné, où les individus de la population sont choisis suivant des critères judicieux ;
- l'échantillonnage aléatoire, où les individus sont choisis « au hasard ».

### **5.1.3 Méthodes de sondage par choix raisonné**

Ces méthodes sont les plus utilisées par les instituts de sondages (le premier institut de sondage a été créé en 1935 par G. Gallup) car elles sont économiques, rapides et plus faciles à mettre en œuvre en cas d'enquêtes complexes. Mais elles ne permettent pas d'évaluer la précision des informations obtenues.

Méthode des *quota* : la population est partagée en classes suivant les modalités de variables de contrôle : par exemple pour un sondage d'opinion, la population sera partagée selon les catégories socioprofessionnelles, le sexe, l'âge, etc. Les effectifs des classes dans la population étant connus et la taille d'échantillon fixée, on calcule les **quotas**, nombres d'individus à interroger dans chaque classe, de telle façon que chaque classe soit représentée dans l'échantillon proportionnellement à son effectif dans la population. Le choix des individus à interroger dans les classes est laissé à l'initiative des enquêteurs.

Echantillon-type : on choisit un groupe d'individus que l'on considère comme fortement représentatif de la population. Par exemple, les électeurs dans des bureaux de vote dont les résultats ont été très proches des résultats définitifs des élections pendant une longue période. Ou encore les lecteurs des grands quotidiens qui ont servi aux Etats-Unis de méthode de prévision des résultats électoraux avant l'expérience de sondage de Gallup en 1937.

### **5.1.4 Sondages aléatoires**

Les méthodes de sondage aléatoires sont caractérisées par le fait que le choix d'un échantillon est une expérience aléatoire, ou plus précisément une suite d'expériences aléatoires, les tirages d'éléments de la population qui forment un échantillon. Chaque élément de la population a alors une probabilité connue *a priori* d'appartenir à l'échantillon

(probabilités égales = « au hasard », ou inégales). On sait alors calculer la probabilité de chacun des échantillons et on peut évaluer, en termes de probabilité, la confiance que l'on peut avoir dans les conclusions de l'enquête.

Tirages au hasard avec ou sans remise : **méthodes de sondage élémentaires** où chaque élément est choisi « au hasard » dans la population, c'est-à-dire avec une probabilité égale à  $1/N$  ( $N$  étant la taille de la population) ; on distingue les tirages « avec remise » et les tirages « sans remise » (appelé aussi sondage « exhaustif »). Ces méthodes sont coûteuses, difficiles à mettre en œuvre, car elles nécessitent une liste précise des éléments de la population ; mais l'étude des erreurs d'échantillonnage par des méthodes probabilistes est plus simple.

Stratification : la population est partagée en classes, appelées **strates**, suivant les modalités de variables de contrôle supposées en relation avec la variable étudiée ; à l'intérieur de chaque strate, on prélève un échantillon aléatoire. Cette méthode diffère de la méthode des *quotas* par le choix aléatoire des individus dans chaque classe. Le nombre d'individus à interroger dans chaque strate peut être déterminé avec un taux de sondage fixe comme dans la méthode des *quotas* (sondage stratifié homothétique), ou qui varie en fonction de la variabilité dans chaque strate de la quantité étudiée (méthode optimale de Neyman).

Sondage systématique : on prélève « systématiquement » dans une suite des éléments de la population, un élément sur  $r$  où  $r = N/n$  est l'inverse du taux de sondage  $n/N$ , après avoir choisi au hasard le premier élément dans la population. Par exemple, on contrôlera la qualité d'une pièce sur dix (mais un douanier ne contrôlera pas une voiture sur dix à un poste de douane). Cette méthode est particulièrement mauvaise si le critère utilisé pour établir la liste des individus de la population est dépendant de la variable étudiée (exemple d'un fichier classé hiérarchiquement et par service qui pourrait fournir un échantillon exclusivement composé des chefs de service).

Sondage par grappes : au lieu de prélever les éléments isolément, on choisit au hasard des groupes appelés **grappes**, et on examine *tous* les éléments de chaque grappe ; ainsi, un enquêteur interroge tous les habitants d'immeubles choisis au hasard, ou on vérifie toutes les pièces de lots désignés au hasard. Pour un budget fixé, cette méthode de sondage permet souvent de prélever un échantillon de taille plus grande, mais artificiellement homogène.

Sondage à plusieurs degrés : tirages d'échantillons dans des échantillons, lorsque les éléments de la population sont naturellement regroupés en unités. Cette méthode est très utilisée dans les enquêtes sur les ménages ; par exemple, on choisit au hasard des villes, puis dans ces villes, des quartiers, et dans les quartiers des ménages.

Sondage avec probabilités inégales : au lieu de donner à chaque élément de la population la probabilité  $1/N$  d'être choisi, on peut leur affecter des probabilités différentes ; par exemple, les villes seront choisies proportionnellement au nombre de leurs habitants, les entreprises proportionnellement à leur nombre de salariés ou à leur chiffre d'affaires.

Plusieurs de ces procédés peuvent être utilisés simultanément ; par exemple, prélever des grappes après stratification. Il faut alors définir précisément un **plan de sondage**. D'autre part il est nécessaire pour les enquêtes permanentes ou périodiques de conserver le même échantillon : un **panel** est un échantillon permanent de personnes interrogées à intervalles réguliers. Pour des raisons évidentes de « mortalité » des individus, réelle ou par sortie de la catégorie étudiée, il est nécessaire de renouveler partiellement et périodiquement l'échantillon.

## 5.2 Échantillonnage et Estimation

### 5.2.1 Échantillon et Statistique d'échantillonnage

Au premier sens nous appelons échantillon un sous ensemble d'individus  $u_1, u_2, \dots, u_n$  pris dans une population. Mais le plus souvent ce ne sont pas les individus eux mêmes qui nous intéressent mais l'observation ou la mesure d'une variable  $X$  sur ces individus. Nous appelons alors **échantillon fortuit** la suite (ou vecteur)  $\mathbf{x} = (x_1, x_2, \dots, x_i, \dots, x_n)$  de données individuelles qui sont les résultats de cette mesure.

Chacune de ces valeurs cependant est le résultat d'un tirage aléatoire, c'est à dire d'une épreuve aléatoire à laquelle nous avons associé la variable aléatoire  $X$  dont la loi est définie par la fonction de répartition  $F_X(x) = P(X \leq x)$  et la densité  $f_X(x) = dF_X(x)/dx$ . Et le sondage est une épreuve aléatoire composite qui résulte de la répétition de cette épreuve simple consistant à tirer et mesurer un individu. A cette épreuve composite nous associons un  $n$ -uplet de variables aléatoires (ou vecteur aléatoire)  $\mathbf{X} = (X_1, X_2, \dots, X_i, \dots, X_n)$  que nous appelons **échantillon aléatoire**. De même qu'il faut éviter de confondre La variable aléatoire  $X_i$  relative au  $i$ ème tirage et la valeur  $x_i$  qui en est *une* réalisation fortuite, de même ne faut-il pas confondre l'échantillon aléatoire qui est une suite de variables aléatoires et l'échantillon fortuit qui est la suite de leur réalisation numérique.

La loi du vecteur  $\mathbf{X}$  est un peu compliquée sauf si nous nous plaçons dans le cas plus simple, celui d'une **répétition indépendante de  $n$  observations d'une même variable aléatoire  $X$  de loi  $L$** . Les variables  $X_i$  sont de même loi et indépendantes. Nous disons alors que l'échantillon est (une suite de variables) *i.i.d.* (indépendantes et identiquement distribuées). C'est le cas du tirage d'un échantillon d'individus au hasard *avec remise* dans une population finie ou infinie, ou encore, par approximation, d'un tirage sans remise si la taille de l'échantillon  $n$  est faible devant celle de la population  $N$ .

a) Comme pour une variable aléatoire simple, nous pouvons définir la loi de probabilité d'un échantillon aléatoire  $\mathbf{X}$  de taille  $n$ , par la « **vraisemblance** » (Likelihood)  $L(\mathbf{x}) = \text{prob}(\mathbf{X}=\mathbf{x})$  dans le cas discret et par la densité  $L(\mathbf{x})$  dans le cas continu. Pour un échantillon *i.i.d.* on a :

Dans le cas discret, la vraisemblance est le produit des probabilités :

$$L(x_1, x_2, \dots, x_n) = \text{prob}(X_1 = x_1, X_2 = x_2, \dots, X_n = x_n) = \prod_{i=1}^n P(X = x_i)$$

Par exemple, un échantillon *i.i.d.* de taille 3 d'une loi de Poisson de paramètre  $\lambda=2$  aura pour

$$\text{densité : } \text{Prob}(X = x_1, x_2, x_3) = \frac{e^{-\lambda} \lambda^{x_1}}{x_1!} \frac{e^{-\lambda} \lambda^{x_2}}{x_2!} \frac{e^{-\lambda} \lambda^{x_3}}{x_3!} = e^{-6} \frac{2^{x_1+x_2+x_3}}{x_1! x_2! x_3!}$$

Dans le cas continu, la vraisemblance de l'échantillon est le produit des densités :

$$L(x_1, x_2, \dots, x_n) = f_X(x_1, x_2, \dots, x_n) = \prod_{i=1}^n f_X(x_i)$$

b) Nous définissons enfin un **résumé statistique d'échantillon  $\mathbf{T}$**  (quelques fois appelé « une statistique ») comme une fonction  $f$  de l'échantillon aléatoire :

$$T = f(\mathbf{X}) = f(X_1, X_2, \dots, X_n)$$

Donc  $T$  est elle-même une variable aléatoire et

$$t = f(\mathbf{x}) = f(x_1, x_2, \dots, x_n)$$

est une réalisation de cette variable aléatoire pour un échantillon fortuit.

Tout résumé statistique tel que nous l'avons étudié en statistique descriptive est maintenant considéré comme une statistique d'échantillon susceptible de varier avec lui, en particulier les moyennes, les variances et plus généralement les moments (une notion qui généralise les précédentes). Il faut éviter de confondre ces caractéristiques de l'échantillon avec celles de la population car l'objet de cette section est justement d'établir la relation entre les deux notions :

Statistique	Population	Echantillon
<b>a) Moments</b>	<b>Moments théoriques</b>	<b>Moments empiriques</b>
<b>Moyenne</b>	$m_1 = m = E(X)$	$m'_1 = \bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$
<b>Moyenne des carrés</b>	$m_2 = E(X^2)$	$m'_2 = \frac{1}{n} \sum_{i=1}^n X_i^2$
<b>Variance</b>	$\mu_2 = \sigma^2 = E(X-m)^2$	$\mu'_2 = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2 = S^2$
<b>Moment non centré d'ordre k</b>	$m_k = E(X^k)$	$m'_k = \frac{1}{n} \sum_{i=1}^n X_i^k$
<b>Moment centré d'ordre k</b>	$\mu_k = E(X - m)^k$	$\mu'_k = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^k$
<b>b) Fractiles</b>	<b>Fractiles théoriques</b>	<b>Fractiles empiriques</b>
<b>Médiane</b>	$M = \text{Médiane}(X)$	$M' = \text{Médiane échantillon}$
<b>Fractile d'ordre <math>\alpha</math></b>	$F_\alpha / \text{prob}(X < F_\alpha) = \alpha$	$F'_\alpha / \text{Freq cum}(F'_\alpha) = \alpha$

La question est alors de définir les propriétés de ces résumés statistiques, c'est à dire principalement leurs moments et quelques fois leurs lois, en fonction des hypothèses du modèle d'échantillonnage. Puis d'utiliser ces résumés statistiques comme estimateurs des quantités inconnues si elles ont les bonnes propriétés.

### 5.2.2 Qu'est-ce qu'un estimateur ?

Si l'on cherche à connaître les propriétés d'une statistique d'échantillon pour une variable de loi entièrement connue dans la population on est dans un problème d'échantillonnage. Si au contraire on cherche à évaluer une caractéristique  $\theta$  d'une variable  $X$  dans une population - proportion de consommateurs achetant le produit A, moyenne des dépenses d'habillement, moyenne et variance des revenus – à partir de l'observation d'un et un seul échantillon fortuit  $(x_1, x_2, \dots, x_n)$ , on est dans un problème d'estimation.

Dans la **théorie classique** de l'estimation due à Ronald Fisher, la valeur réelle mais inconnue de  $\theta$  est un paramètre de la loi L de la variable X observée n fois dans un échantillon aléatoire  $(X_1, X_2, \dots, X_n)$  *i.i.d.*. Les hypothèses que l'on peut faire a priori sur la loi L sont bien précisées, et l'objectif est d'atteindre la vraie valeur de  $\theta$  avec une erreur minimale. On appelle **estimateur de  $\theta$**  une statistique d'échantillon aléatoire :

$$T = g(X_1, X_2, \dots, X_n)$$

Attention, puisque les  $X_i$  sont des variables aléatoires, T est aussi une variable aléatoire.

On appelle estimation ponctuelle de  $\theta$  la réalisation de cette statistique dans un échantillon fortuit :

$$t = g(x_1, x_2, \dots, x_n)$$

La fonction g décrit le moyen de calculer une estimation de  $\theta$ .

### 5.2.3 Propriétés d'un estimateur

N'importe quelle fonction g ne convient pas pour estimer un paramètre  $\theta$ . L'estimateur doit avoir de bonnes propriétés par rapport à la valeur-cible. En gros on souhaite deux choses : a) que l'estimateur vise en moyenne la vraie valeur b) avec une variabilité autour de cette vraie valeur qui soit la plus faible possible.

a) On appelle biais (erreur moyenne) la quantité  $B = E(T - \theta) = E(T) - \theta$ .

Si le biais est positif, l'estimateur T surestime  $\theta$  en moyenne. Si le biais est négatif, l'estimateur T sous-estime  $\theta$  en moyenne. Une statistique T est un estimateur sans biais pour  $\theta$  si le biais (l'erreur moyenne)  $B = E(T - \theta) = E(T) - \theta$  est nul. On peut se contenter d'un estimateur dont le biais diminue avec la taille de l'échantillon.

T est un estimateur de  $\theta$  **sans biais** si  $E(T) = \theta$ .

T est un estimateur de  $\theta$  **asymptotiquement sans biais** si  $E(T) \xrightarrow{n \rightarrow \infty} \theta$

b) On souhaite par ailleurs que l'estimateur ait une bonne précision. Celle-ci est mesurée par :

L'erreur quadratique = moyenne quadratique des écarts à la valeur cible =  $E((T - \theta)^2)$

Mais celle-ci se décompose en 2 parties : la variance V(T) et le carré du biais  $B^2$  :

$$E((T - \theta)^2) = E((T - E(T))^2) + (\theta - E(T))^2 = V(T) + B^2$$

Donc pour un estimateur qui est déjà sans biais il faut seulement minimiser sa variance. On peut le faire de deux façons :

- exiger que cette variance tende vers 0 quand la taille de l'échantillon augmente indéfiniment. T est un estimateur de  $\theta$  **convergent** si T converge en probabilité vers  $\theta$ , c'est à dire si pour tout  $\varepsilon > 0$  :  $\text{prob}\left(\left\{|T - \theta| > \varepsilon\right\}\right) \xrightarrow{n \rightarrow \infty} 0$ . Il suffit pour cela qu'il soit convergent en moyenne quadratique c'est à dire que  $E((T - \theta)^2) = V(T) + B^2 \xrightarrow{n \rightarrow \infty} 0$ . S'il est déjà sans biais, il suffira que sa variance tende vers zéro.

- exiger que pour une taille donnée d'échantillon sa variance soit la plus petite possible. Les statisticiens Fisher, Cramer et Rao ont montré que cette variance possédait un minimum pour une loi et une taille d'échantillon données. Si la variance de l'échantillon atteint cette limite on dit que l'estimateur est *efficace*.

La plupart du temps nous chercherons donc des estimateurs sans biais et convergents. Cela n'est pas toujours facile. Et parfois contradictoire.

#### 5.2.4 Intervalles de confiance

Plutôt que de fournir une estimation ponctuelle dont on sait qu'elle serait différente avec un autre échantillon, et qu'il faudrait assortir d'une mesure de précision, on préfère souvent fournir un intervalle de confiance (appelé fourchette le soir des élections).

Un *intervalle de confiance* est un intervalle aléatoire  $[A, B]$  qui contient la valeur du paramètre  $\theta$  avec une probabilité donnée :

$$prob(\{\theta \in [A, B]\}) = prob(\{A \leq \theta \leq B\}) = 1 - \alpha \Leftrightarrow prob(\{\theta \notin [A, B]\}) = \alpha$$

On dit que  $[A, B]$  est un intervalle de confiance au *niveau de confiance*  $1 - \alpha$ , habituellement 90 %, 95 % ou 99 %. Généralement il s'agit d'un intervalle de confiance bilatéral de la forme :

$$prob(\{A \leq \theta \leq B\}) = 1 - \alpha \text{ avec } prob(\{\theta < A\}) = prob(\{\theta > B\}) = \alpha/2$$

Les bornes aléatoires  $A$  et  $B$  sont définies à partir d'un échantillon :

$$A = g_1(X_1, X_2, \dots, X_n) \text{ et } B = g_2(X_1, X_2, \dots, X_n)$$

Mais  $A$  et  $B$  prennent les valeurs numériques  $a$  et  $b$  pour un échantillon fortuit :

$$a = g_1(x_1, x_2, \dots, x_n) \text{ et } b = g_2(x_1, x_2, \dots, x_n)$$

$[a, b]$  est une *réalisation de l'intervalle de confiance*  $[A, B]$ .

La définition de l'intervalle de confiance  $[A, B]$  de niveau  $1 - \alpha$  signifie que l'ensemble des échantillons pour lesquels l'intervalle  $[a, b]$  recouvre la vraie valeur du paramètre  $\theta$  a une probabilité égale à  $1 - \alpha$ . Pour une population finie dans laquelle on a fait un sondage au hasard, cela peut s'interpréter ainsi : dans l'ensemble de tous les échantillons possibles, la proportion de ceux pour lesquels la vraie valeur du paramètre  $\theta$  appartient à l'intervalle  $[a, b]$  est de  $1 - \alpha$ . Pour trouver de tels intervalles de confiance on part toujours de l'estimateur ponctuel et de sa loi de probabilité.

#### 5.2.5 Un exemple

On sait par hypothèse qu'une grandeur  $X$  suit une loi continue uniforme entre 0 et  $a$ , mais on ne connaît pas la valeur maximale  $a$ . On cherche à l'estimer à partir d'un échantillon de 5 tirages indépendants de valeurs de  $X$ .

Une première méthode (dite des moments) propose de prendre comme estimateur de  $a$  le double de la moyenne :  $\hat{a}_1 = 2\bar{X}$ . Une seconde méthode (dite du maximum de vraisemblance) propose de prendre comme estimateur de  $a$  le maximum des valeurs de l'échantillon :  $\hat{a}_2 = \text{Max}(X_i)$

Quelle méthode faut-il préférer ? Elles ne donnent pas en général la même estimation :

. Si l'échantillon fortuit observé est (3, 7, 2, 5, 3) on obtient  $\hat{a}_1 = 8$  et  $\hat{a}_2 = 7$ .

- . Si l'échantillon fortuit observé est (3, 7, 2, 10, 3) on obtient  $\hat{a}_1 = 10$  et  $\hat{a}_2 = 10$ .
- . Si l'échantillon fortuit observé est (5, 8, 14, 2, 1) on obtient  $\hat{a}_1 = 12$  et  $\hat{a}_2 = 14$ .

Pour trancher, il faut étudier les propriétés de ces deux estimateurs. On peut montrer que le premier est sans biais (Cf. 5.3.1) alors que le second n'est qu'asymptotiquement sans biais. Ils sont tous les deux convergents mais le second a une plus petite variance. Le choix est donc Cornélien. Mais si l'on remarque que très souvent (par exemple dans le troisième échantillon) le premier donne une estimation de  $a$  inadmissible parce que inférieure à une valeur observée, alors le second estimateur est à préférer.

Cet exemple simple mais assez tordu montre que le problème de l'estimation d'un paramètre n'est pas toujours évident. On se contentera donc dans la suite de traiter 3 cas usuels : l'estimation d'une moyenne, d'une proportion, et d'une variance.

### 5.3 Jugement sur la moyenne d'une population

#### 5.3.1 Distribution d'une moyenne empirique $\bar{X}$

Nous nous plaçons dans le cas d'un modèle d'échantillonnage très général.  $X$  est une variable de loi inconnue  $L$  de paramètre  $E(X) = m$  et  $V(X) = \sigma^2$ .

$\bar{X}$  est la moyenne arithmétique de l'échantillon. Tandis que  $m$  est la moyenne théorique (celle de  $X$  dans la population).

1.  $\bar{X}$  est une variable aléatoire, c'est à dire qu'elle varie quand on tire un nouvel échantillon. Mais bien sûr pour un échantillon fortuit donné elle prend une valeur numérique  $\bar{x}$  qui est sa réalisation et qui n'est plus aléatoire.
2.  $\bar{X}$  converge en probabilité vers  $m$  : c'est la **loi des grands nombres** dans le cas particulier  $L = \text{Loi Bernoulli (1713)}$  déjà vu, le théorème de Khinchine 1929 dans le cas général.

$$\forall \varepsilon > 0, \text{prob}\{|\bar{X} - m| < \varepsilon\} \xrightarrow{n \rightarrow \infty} 1$$

3. Notons la nature complexe de cette convergence : on peut trouver une taille  $n$  d'échantillon telle que l'écart entre moyenne et Espérance soit inférieur à un nombre donné  $\varepsilon$  avec une probabilité aussi grande que l'on veut. Donc  $\bar{X}$  varie d'un échantillon à l'autre, mais en restant d'autant plus proche de la moyenne  $m$  de la population que  $n$  est grand.
4. Calculons l'espérance mathématique de  $\bar{X}$ .

$$E(\bar{X}) = E\left(\frac{X_1 + X_2 + \dots + X_n}{n}\right) = \frac{1}{n} E(X_1 + X_2 + \dots + X_n) = \frac{1}{n} nm = m :$$

$E(\bar{X}) = m$ . La moyenne empirique dans l'échantillon « tourne autour » de la vraie valeur dans la population. Nous dirons que  $\bar{X}$  est un estimateur sans biais de  $m$ .

5. Calculons maintenant la variance de  $\bar{X}$

$$V(\bar{X}) = V\left(\frac{X_1 + X_2 + \dots + X_n}{n}\right) = \frac{1}{n^2} V(X_1 + X_2 + \dots + X_n) = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}$$



$$V(\bar{X}) = \sigma^2/n \Leftrightarrow \sigma(\bar{X}) = \sigma(X)/\sqrt{n}$$

La variabilité de  $\bar{X}$  est réduite d'un facteur  $\sqrt{n}$  par rapport à celle de  $X$ . Une moyenne de  $n$  quantité c'est beaucoup moins variable que chacune d'entre elles. Et de plus si  $n$  augmente à l'infini cette variance diminue et tend vers zéro. Nous dirons que  $\bar{X}$  est un estimateur convergent de  $m$ .

6. Quelle loi de probabilité suit  $\bar{X}$  ?

Le **théorème de la limite centrale** de Laplace (1810) et Lindeberg-Lévy (1922) nous dit que quelle que soit la loi  $L$  suivie par  $X$ , pourvu que  $X$  admette un moment d'ordre 2 fini,  $\bar{X}$  suit une loi approximativement normale si  $n$  est assez grand ( $n \geq 30$ ) :

$$\bar{X} \mapsto N(m, \sigma/\sqrt{n}) \Leftrightarrow \frac{\bar{X} - m}{\sigma/\sqrt{n}} \mapsto N(0,1)$$

On observera sur les simulations de la leçon 4 la rapidité de cette convergence qui permet en général d'utiliser ce théorème à partir de  $n=15$  avec une bonne approximation.

6. Puisque la moyenne empirique  $\bar{X}_n$  dans un échantillon de taille  $n$  a pour Espérance  $m$  et pour variance  $\sigma^2/n$ , on en déduit que  $\bar{X}$  est un estimateur de  $m$  sans biais et convergent de  $m$ .

### 5.3.2 Cas particulier d'un échantillon de variables normales

a)  $X_i \mapsto N(m, \sigma) \Leftrightarrow Y_i = \frac{X_i - m}{\sigma} \mapsto N(0,1)$

b) Même si  $n$  est petit,  $\bar{X} \mapsto N(m, \sigma/\sqrt{n}) \Leftrightarrow \sqrt{n}\bar{Y} = \frac{\bar{X} - m}{\sigma/\sqrt{n}} \mapsto N(0,1)$ .

c) La variable  $W = \sum_i Y_i^2 = \sum_{i=1}^n \frac{(X_i - m)^2}{\sigma^2} \mapsto \chi^2(n)$ .

d) Le théorème de Fisher dit que les deux variables :

$$n\bar{Y}^2 = n \frac{(\bar{X} - m)^2}{\sigma^2} \text{ et } U^2 = \sum_{i=1}^n (Y_i - \bar{Y})^2 = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{\sigma^2} = \frac{nS^2}{\sigma^2}$$

sont *indépendantes* et suivent respectivement les lois  $\chi^2(1)$  et  $\chi^2(n-1)$ .

e) On en déduit la loi suivie par la variable  $\bar{X}$  centrée et réduite par l'écart-type corrigé suit une loi de Student :

$$\frac{\sqrt{n}\bar{Y}}{\sqrt{U^2/(n-1)}} = \frac{\bar{X} - m}{S/\sqrt{n-1}} = \frac{\bar{X} - m}{S'/\sqrt{n}} \mapsto Student(n-1)$$

En résumé, quand  $\sigma^2$  est inconnu on ne peut pas utiliser  $\frac{\bar{X} - m}{\sigma/\sqrt{n}} \mapsto N(0,1)$ , mais on lui

substitue l'expression analogue dans laquelle  $\sigma$  est remplacé par son estimation  $S'$ , puis on remplace la table de la loi normale par celle de la loi de Student.( $n-1$ ).

### 5.3.3 Intervalle pour une moyenne m

Nous partons de la **moyenne empirique**  $\bar{X}$  qui est un estimateur sans biais et convergent pour une moyenne théorique  $m$ . De plus on sait d'après le théorème central-limite, que  $\bar{X}$  suit approximativement la loi normale  $LG(m, \sigma/\sqrt{n})$  pour  $n$  suffisamment grand ( $n \geq 30$ ). Donc on peut écrire :

$$\text{prob}\left(-t < \frac{\bar{X} - E(\bar{X})}{\sigma(\bar{X})} \leq t\right) = \text{prob}\left(-t \frac{\sigma}{\sqrt{n}} \leq \bar{X} - m \leq t \frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha$$

avec  $t$  fractile d'ordre  $1 - \alpha/2$  de la loi  $N(0,1)$  et  $-t$  fractile d'ordre  $\alpha/2$ .

De cette relation nous pourrions en tirer un encadrement de  $\bar{X}$  pour  $m$  connu (question d'échantillonnage) mais c'est le problème inverse qui nous intéresse. Donc nous isolons  $m$  au centre de l'inégalité :

L'intervalle de confiance bilatéral au niveau  $1 - \alpha$  pour  $m$  est donc :

$$\left[ \bar{X} - t_{\alpha} \frac{\sigma}{\sqrt{n}} < m < \bar{X} + t_{\alpha} \frac{\sigma}{\sqrt{n}} \right]$$

Mais cette formule n'est applicable que si  $\sigma^2$  est connue.

Dans le cas inverse -  $\sigma^2$  est inconnue - on la remplace par son estimation  $S^2$  ou  $S'^2$  et on utilise la statistique  $\frac{\bar{X} - m}{S'/\sqrt{n}} = \frac{\bar{X} - m}{S/\sqrt{n-1}}$  qui suit la même loi  $N(0,1)$  si  $n$  est grand, et la loi de Student à  $n-1$  degrés de liberté sinon (en supposant que  $X$  est une variable normale) :

$$\text{prob}\left(\left\{-t \leq \frac{\bar{X} - m}{S'/\sqrt{n}} \leq t\right\}\right) = \text{prob}\left(\left\{-t \frac{S'}{\sqrt{n}} \leq \bar{X} - m \leq t \frac{S'}{\sqrt{n}}\right\}\right) = 1 - \alpha$$

ce qui définit un intervalle de confiance bilatéral de niveau  $1 - \alpha$  pour  $m$  :

$$\left[ \bar{X} - t \frac{S'}{\sqrt{n}}, \bar{X} + t \frac{S'}{\sqrt{n}} \right] = \left[ \bar{X} - t \frac{S}{\sqrt{n-1}}, \bar{X} + t \frac{S}{\sqrt{n-1}} \right]$$

#### 5.3.3.1 Exemple

On cherche à estimer le prix moyen  $m$  d'un produit dans les magasins de la région parisienne à partir des résultats d'une enquête dans 20 magasins choisis au hasard (avec remise) dans la région parisienne : prix moyen  $\bar{x} = 285$  F avec un écart-type  $s = 30,2$  F.

L'intervalle de confiance bilatéral à 95 % pour  $m$  est :

$$\left[ \bar{X} - t_{n-1; 1-\alpha/2} \frac{S}{\sqrt{n-1}}, \bar{X} + t_{n-1; 1-\alpha/2} \frac{S}{\sqrt{n-1}} \right]$$

La lecture de la table de la loi de Student avec  $\nu = n - 1 = 19$  donne  $t_{19; 0,975} = 2,093 \approx 2,1$  et la réalisation de l'intervalle de confiance pour l'échantillon fortuit est [270,4 F, 299,5 F].

*Remarque* : il n'est pas indispensable de connaître la série de données individuelles obtenue lors d'une enquête, il suffit d'en connaître deux résumés statistiques, la moyenne arithmétique et l'écart-type (ou la variance) pour estimer la moyenne théorique  $m$ .

## 5.4 Jugement sur une proportion

### 5.4.1 Distribution d'une fréquence F

Si une épreuve aléatoire a deux issues  $A$  et  $B$  de probabilités  $p$  et  $(1-p)$ , alors nous pouvons définir  $X = \{\text{nombre de fois où le résultat est } A\} = \{\text{variable indicatrice de } A\}$  comme une variable de Bernoulli dont la loi est :

x	0	1
$P(X = x)$	$1 - p$	$p$

$E(X) = p, \quad E(X^2) = p$   
 et  $V(x) = p(1-p)$ .

Pour un échantillon  $(X_1, X_2, \dots, X_n)$  de cette loi, la somme des variables est le nombre de fois où l'on a obtenu  $A$  (fréquence absolue), et la moyenne des variables n'est rien d'autre que la fréquence relative de  $A$  dans l'échantillon :  $\bar{X} = F$ . Nous pouvons donc utiliser les propriétés précédentes de  $\bar{X}$  dans le cas où  $L$  n'est pas une loi quelconque mais une loi de Bernoulli.

1.  $F$  est une variable aléatoire (elle varie d'un échantillon à un autre).
2.  $F$  converge en probabilité vers  $p$ , **loi des grands nombres** (Bernoulli 1713) :

$$\forall \varepsilon > 0, \quad \text{prob}\{|F - p| < \varepsilon\} \xrightarrow{n \rightarrow \infty} 1$$

3.  $E(F) = p$ , donc  $F$  est un estimateur sans biais de  $p$ .
4.  $V(F) = p(1-p)/n$ , donc  $F$  fluctue autour de  $p$  avec un écart-type en  $1/\sqrt{n}$ .
5. La loi de  $\sum_{i=1}^n X_i = nF$  est une Binomiale  $(n, p)$  qui peut être approchée.
  - par une loi normale  $N(np, \sqrt{np(1-p)})$  si  $p$  est moyen et  $n$  assez grand,
  - par une loi de Poisson  $(\lambda = np)$  si  $p$  est faible et  $n$  assez grand, mais cette dernière elle-même est très proche de la loi normale si  $\lambda = np > 15$ .

Il en résulte que  $F$  suit approximativement une loi normale ( $n > 30, np > 15$ ) :

$$F \mapsto N(p, \sqrt{p(1-p)/n}) \Leftrightarrow \frac{F - p}{\sqrt{p(1-p)/n}} \mapsto N(0,1)$$

Si par exemple je sonde au hasard  $n = 1000$  personnes dans une population dont une proportion  $p = 30\%$  a la propriété  $A$ , alors la fréquence des personnes qui ont cette même propriété dans l'échantillon est une variable aléatoire normale d'espérance  $E(F) = 0,3$  de variance  $0,00021$  (écart-type  $0,0145$ ).

6. Comme  $V(F) = p(1-p)/n$  tend vers 0 quand  $n$  augmente infiniment,  $F$  est un estimateur sans biais et convergent de  $p$ . D'ailleurs la loi faible des grands nombres (ou théorème de Bernoulli) exprime que la suite des fréquences empiriques  $F_n$  converge en probabilité vers la proportion  $p$ . Tout cela justifie donc le choix de la **fréquence empirique**  $F$  comme estimateur de  $p$ .

### 5.4.2 Intervalle pour une proportion p

On part de la **fréquence empirique**  $F$  qui est un estimateur sans biais et convergent pour une proportion  $p$ . On sait que  $F$  suit aussi approximativement une loi de Laplace-Gauss. Soit  $t$  le fractile d'ordre  $1 - \alpha/2$  de la loi  $LG(0,1)$ , alors on peut écrire :

$$\text{prob}\left(\left|\frac{F - E(F)}{\sigma(F)}\right| \leq t\right) = \text{prob}\left(-t\sqrt{\frac{p(1-p)}{n}} \leq F - p \leq t\sqrt{\frac{p(1-p)}{n}}\right) = 1 - \alpha$$

Ce qui donne comme précédemment un intervalle de confiance pour  $p$  de la forme :

$$\left\{ F - t\sqrt{\frac{p(1-p)}{n}} \leq p \leq F + t\sqrt{\frac{p(1-p)}{n}} \right\}$$

Mais cet intervalle est inutilisable puisque  $p$ , inconnu, est encadré par deux quantités qui dépendent aussi de  $p$ . On a trois solutions pour s'en sortir :

a) On estime la variance théorique  $p(1-p)$  par  $F(1-F)$  et on obtient l'intervalle **approché** suivant :

$$\left\{ F - t\sqrt{\frac{F(1-F)}{n}} \leq p \leq F + t\sqrt{\frac{F(1-F)}{n}} \right\}$$

b) On remarque que  $p(1-p)$  est maximum pour  $p=1/2$  et vaut alors  $1/4$ . On peut en déduire un **intervalle par excès** (bien utile pour un calcul à la louche si  $p$  n'est pas trop faible) :

$$\left\{ F - t\sqrt{\frac{1}{4n}} \leq p \leq F + t\sqrt{\frac{1}{4n}} \right\}$$

c) On peut chercher à résoudre précisément l'équation

$$\begin{aligned} \left\{ -t\sqrt{\frac{p(1-p)}{n}} \leq F - p \leq t\sqrt{\frac{p(1-p)}{n}} \right\} &= \left\{ (F - p)^2 \leq t^2 \frac{p(1-p)}{n} \right\} \\ &= \left\{ \left(1 + \frac{t^2}{n}\right)p^2 - 2\left(F + \frac{t^2}{2n}\right)p + F^2 \leq 0 \right\} \end{aligned}$$

qui définit une ellipse et un

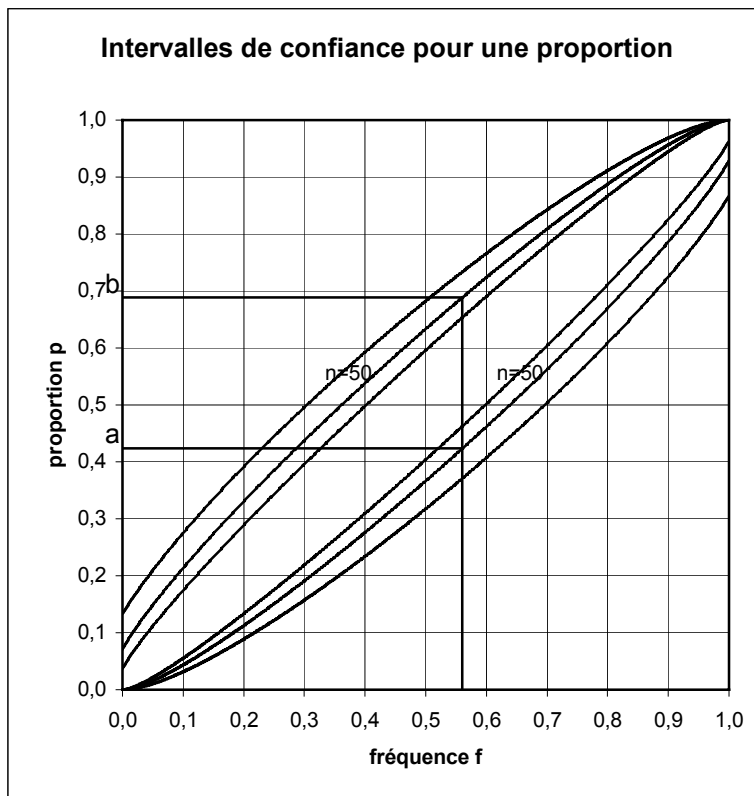
intervalle de la forme

$$\left\{ \frac{F + \frac{t^2}{2n} - t\sqrt{\frac{F(1-F)}{n} + \frac{t^2}{4n^2}}}{1 + \frac{t^2}{n}} \leq p \leq \frac{F + \frac{t^2}{2n} + t\sqrt{\frac{F(1-F)}{n} + \frac{t^2}{4n^2}}}{1 + \frac{t^2}{n}} \right\}$$

Ce résultat, dû au statisticien français Stanislas Millot (1925), a donné lieu à la confection d'abaques (courbes pour différentes valeurs de  $n$  et de  $1 - \alpha$ ) que l'on peut aujourd'hui tracer avec Excel.

*Remarque :* On trouve l'intervalle de confiance précédent en négligeant les termes en  $t^2/n$ .

Exemple de lecture d'un abaque pour un intervalle symétrique en probabilité, pour  $f = 0,56$  et  $n = 50$  :



Dans cet exemple on a  $t=1,96$ ,  $t\sqrt{f(1-f)/n} = 0,138$ ,  $t\sqrt{1/4n} = 0,1386$ . L'intervalle de confiance à 95 % obtenu par la première méthode est  $[0,422, 0,688]$ , par la seconde  $[0,4214, 0,6886]$  alors que par la méthode de l'ellipse on trouve  $[0,423, 0,688]$ . Les écarts sont ici très faibles. Ils seraient très importants pour  $p$  faible.

## 5.5 Jugement sur une variance

### 5.5.1 Distribution d'une variance empirique $S^2$

Nous nous plaçons encore dans le cas d'un modèle d'échantillonnage très général.  $X$  est une variable de loi inconnue  $L$  de paramètre  $E(X) = m$  et  $V(X) = \sigma^2$ .

On appelle  $S^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2$  la variance empirique de  $X$  dans l'échantillon. C'est une statistique que l'on a utilisé pour mesurer la variabilité de  $X$  en statistique descriptive.  $S$ , l'écart-type, représente une moyenne (quadratique) des écarts à la moyenne. On peut aussi calculer  $S^2$  par la formule équivalente  $S^2 = \frac{1}{n} \sum_i X_i^2 - \bar{X}^2$  (moyenne des carrés moins carré de la moyenne). Si les valeurs  $X_i$  ne sont pas observées 1 fois mais  $n_i$  fois on peut bien sûr les regrouper et utiliser la formule plus générale avec pondérations  $S^2 = \frac{1}{n} \sum_i n_i X_i^2 - \bar{X}^2$

1.  $S^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2$  est une variable aléatoire.

2. Calculons l'espérance de  $S^2$ .

$$E(S^2) = E\left(\frac{\sum_i X_i^2}{n}\right) - E(\bar{X}^2) = V(X) + (E(X))^2 - [V(\bar{X}) + (E(\bar{X}))^2] = \sigma^2 + m^2 - \frac{\sigma^2}{n} - m^2 = \frac{n-1}{n}\sigma^2$$

$E(S^2) = \frac{n-1}{n}\sigma^2$ , donc  $S^2$  est un estimateur de  $\sigma^2$  biaisé, mais comme ce biais tend vers 0 quand  $n \rightarrow \infty$ , ce n'est pas très grave, et on dit que  $S^2$  est asymptotiquement sans biais.

On lui préfère quand même  $S'^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2 = \frac{n}{n-1} S^2$  appelée **variance corrigée** dont on peut vérifier qu'elle est sans biais.

3. On admettra que  $V(S^2) \cong \frac{\mu_4 - \mu_2^2}{n}$ . Donc  $S^2$  est convergent et  $S'^2$  l'est aussi.

4.  $Cov(\bar{X}, S^2) = \frac{\mu_3}{n} (1 - \frac{1}{n}) \cong \frac{\mu_3}{n}$  et  $Cov(\bar{X}, S'^2) = \frac{\mu_3}{n}$ , ce qui montre que les deux statistiques sont généralement corrélées. Si  $\mu_3 = 0$  (loi de  $X$  symétrique) elles sont non corrélées (mais pas forcément indépendantes, sauf dans le cas de la loi normale).

5. On ne peut rien dire de la loi de  $S^2$  en général car elle dépend de la loi de  $X$ . Mais **si on peut faire l'hypothèse que  $X$  suit elle même une loi normale**, alors il est possible d'appliquer les résultats de la section 5.3.2 :

$$\frac{\sum_{i=1}^n (X_i - \bar{X})^2}{\sigma^2} = \frac{nS^2}{\sigma^2} = \frac{(n-1)S'^2}{\sigma^2} \mapsto \chi^2(n-1)$$

### 5.5.2 Estimateur d'une variance $\sigma^2$

Le **moment empirique d'ordre 2 autour de  $m$** ,  $M = \frac{1}{n} \sum_{i=1}^n (X_i - m)^2$ ,

la **variance empirique**  $S^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2$  (fonction VAR.P dans Excel)

et la **variance empirique corrigée**  $S'^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$  (VAR dans Excel)

peuvent tous les trois être pris comme estimateurs de  $\sigma^2$ .

a) **Si la moyenne  $m$  est connue**, on utilisera le premier  $M$  qui est calculable et tel que :

Le biais pour  $M$  est  $B = E(M - \sigma^2) = E(M) - \sigma^2 = \sigma^2 - \sigma^2 = 0$

L'erreur quadratique moyenne est  $E((M - \sigma^2)^2) = V(M) = \frac{\mu_4 - \sigma^4}{n} \xrightarrow{n \rightarrow \infty} 0$

$M$  est un estimateur sans biais et convergent pour  $\sigma^2$ .

b) Si la moyenne  $m$  est inconnue, on ne peut pas calculer  $M$ , et  $\sigma^2$  sera estimé par la variance  $S^2$  ou la variance corrigée  $S'^2$ . D'après la section précédente on a :

$$\text{biais pour } S^2 : B = E(S^2 - \sigma^2) = E(S^2) - \sigma^2 = \frac{n-1}{n}\sigma^2 - \sigma^2 = -\frac{\sigma^2}{n} \xrightarrow{n \rightarrow \infty} 0$$

erreur quadratique moyenne :

$$E((S^2 - \sigma^2)^2) = V(S^2) + B^2 = \frac{(n-1)^2(\mu_4 - \sigma^4)}{n^3} + \frac{2(n-1)\sigma^4}{n^3} + \frac{\sigma^4}{n^2} \xrightarrow{n \rightarrow \infty} 0$$

$S^2$  est donc un estimateur de  $\sigma^2$  asymptotiquement sans biais et convergent. Le biais est négatif et la variance empirique sous-estime  $\sigma^2$ .

$$\text{biais pour } S'^2 : B = E(S'^2 - \sigma^2) = E(S'^2) - \sigma^2 = \sigma^2 - \sigma^2 = 0$$

erreur quadratique moyenne :

$$E((S'^2 - \sigma^2)^2) = V(S'^2) = \frac{(\mu_4 - \sigma^4)}{n} + \frac{2\sigma^4}{n(n-1)} \xrightarrow{n \rightarrow \infty} 0$$

$S'^2$  est un estimateur de  $\sigma^2$  sans biais et convergent.

Si  $n$  est grand l'on s'en fiche, mais s'il est petit il faudra préférer  $S'^2$  à  $S^2$ , bien qu'il puisse être légèrement moins précis. Si la moyenne théorique  $m$  est connue, le moment empirique d'ordre 2 autour de  $m$ ,  $M$  est un estimateur sans biais de  $\sigma^2$ , plus précis que  $S^2$  et  $S'^2$ .

### 5.5.3 Exemple :

On ne sait rien de la loi de  $X$  dont on a observé 5 valeurs : 3, 7, 10, 2, 8. Estimez la variance de  $X$ .

Puisque l'on ne connaît pas  $m$ , et bien que l'on puisse l'estimer par  $\bar{X} = 6$ , l'estimation de la variance se fera par le calcul de  $S'^2 = [(-3)^2 + 1^2 + 4^2 + (-4)^2 + 2^2] / 4 = 11,5$ .

Si l'on savait par exemple que la moyenne théorique de  $X$  est  $m = 5$ , alors l'estimation de la variance de  $X$  se ferait par le calcul de  $M = [(-2)^2 + 2^2 + 5^2 + (-3)^2 + 3^2] / 5 = 10,2$ .