Estimateurs

Frédéric Bertrand et Myriam Bertrand[1]

[1]IRMA, Université Louis Pasteur
Strasbourg, France

École Doctorale 19-09-2007

Ce cours s'appuie sur

- "Probabilités, Statistique inférentielle, Fiabilité, Outils pour l'ingénieur" de G. Demengel, P. Bénichou, R. Bénichou, N. Roy, J-P Pouget, Ellipses.
- "Éléments de statistique" de J.J. Droesbeke, Ellipses.

Définition :
La population

est l'ensemble de tous les éléments qui constituent cet ensemble.

Exemples : Un ensemble d'êtres humains, un ensemble d'objets ou de faits.

Définition : Chacun des membres de cette population est appelé **individu**.

Définition : Un sous-ensemble de la population est appelé **échantillon**.

Définition : La propriété qui fait l'objet de l'étude statistique doit être également précisée. On l'appelle **variable statistique**.

Définition : Un caractère est **quantitatif** s'il est mesurable et la liste des nombres exprimant ces mesures est appelée **série statistique**.

Exemples : La taille des habitants d'un pays, le poids des habitants d'un pays, les températures d'une capitale d'un pays.

Définition : Un caractère est **qualitatif** si les modalités échappent à la mesure.

Exemples : Les goûts pour un film, les goûts pour un aliment, l'appréciation d'un livre.

Dans ce dernier cas, la nature du **caractère qualitatif** suggère souvent une répartition de la population en classes que l'on peut numéroter.

On attache alors à chaque individu le numéro de la classe à laquelle il appartient.

On est ainsi ramené à traiter une liste de nombres, c'est-à-dire **une série statistique**.

Une autre distinction peut être faite sur la nature des nombres d'une série statistique.

Définition : Un caractère est dit **discret** si les valeurs qu'il prend sont des nombres isolés, par exemple entières mais pas nécessairement.

Exemples : Le caractère "nombre de pièces dans un appartement", le caractère "nombre d'étudiants en Ecole Doctorale SVS", les températures en dixième degrés sont tous des nombres isolés.

Définition :

Un caractère est dit **continu** si entre deux de ces valeurs on peut toujours trouver une troisième.

Exemples : La taille des habitants d'un pays, exprimée en cm par exemple, se traduit par des nombres réels, le poids des habitants d'un pays, exprimé en kg, s'exprime par des nombres réels.

C'est la nature de l'instrument de mesure qui permet seulement d'en obtenir une valeur approchée. Il est logique alors de considérer que les valeurs du caractère appartiennent à des intervalles appelés **classes**.

Dans ce paragraphe, nous énoncerons les définitions de deux paramètres de position. Nous ne sommes pas exhaustifs. Simplement nous rappelons les définitions de ceux qui seront les plus utilisés dans ce cours.

Définition : **Le mode ou les valeurs modales** d'une série statistique d'une variable discrète est la (ou les) valeur(s) de la variable dont l'effectif est maximum.

Introduisons maintenant les notations dont nous aurons besoin pour définir le paramètre de position suivant : **la moyenne**.

Le tableau des effectifs fournit les effectifs partiels n_j. On dispose donc des couples $(x_j; n_j)$ où j varie de 1 à p.

Dans le cas d'une variable continue, le nombre des classes est encore noté p. On a convenu d'attribuer globalement l'effectif d'une classe au centre c_j de cette classe.

Définition :

Dans le cas d'une variable continue, **la moyenne de la série statistique** composée des couples (c_j, n_j) est définie par :

$$\overline{x} = \frac{1}{N}\sum_{j=1}^{p}(c_j \times n_j)\,,$$

où N représente l'effectif total de la série statistique et p le nombre de classes de la série statistique.

Définition :

Dans le cas d'une variable discrète, **la moyenne de la série statistique** composée des couples (x_j, n_j) est définie par :

$$\overline{x} = \frac{1}{N}\sum_{j=1}^{p}(x_j \times n_j)\,,$$

où N représente l'effectif total de la série statistique et p le nombre de classes de la série statistique.

Dans ce paragraphe, nous énoncerons les définitions de trois paramètres de dispersion. Nous ne sommes pas exhaustifs. Simplement nous rappelons les définitions de ceux qui seront les plus utilisés dans ce cours.

Définition : L'étendue d'une série statistique est la différence entre la plus petite et la plus grande des valeurs du caractère.

Définition :
La variance d'une série statistique

est la moyenne des carrés des écarts des valeurs du caractère x à la moyenne de la série statistique :

$$\sigma^2 = Var(x) = \frac{1}{N} \sum_{j=1}^{p} n_j (x_j - \overline{x})^2,$$

x_j étant remplacé par c_j dans le cas d'une variable continue, N l'effectif total de la série statistique et p le nombre de classes de la série statistique.

Définition :
L'écart-type d'une série statistique

est la racine carrée de la variance de cette série statistique :

$$\sigma = \sqrt{Var(x)}.$$

Remarque : Ce paramètre de dispersion est beaucoup plus intéressant que **la variance** car il a l'intérêt de s'exprimer dans la même unité de mesure que la moyenne. Ainsi on peut comparer des échantillons entre eux.

Une fois ces quelques définitions rappelées, nous pouvons aborder une des questions majeures en statistique :

l'estimation.

Le problème de **l'estimation** est l'impossibilité de connaître exactement la valeur d'un **paramètre inconnu** noté θ. Ce problème est très général et a des aspects distincts.

Les observations permettent de construire une estimation de θ.

Ainsi chaque observation est la valeur d'une variable aléatoire (v.a.) X dont la loi dépend de θ.

Cela revient à doter l'*état de la nature* inconnu d'un **modèle probabiliste**.

Ce dernier est complété par un **modèle d'échantillonnage** décrivant la manière dont les observations sont recueillies.

On se place dans le cas le plus simple : les n observations constituent un **échantillon aléatoire simple** (EAS) composé de n variables aléatoires indépendantes et identiquement distribuées (i.i.d.) $\{X_1, \cdots, X_n\}$.

Signalons qu'un modèle probabiliste complété par un modèle d'échantillonnage est appelé un **modèle statistique**.

Le problème s'énonce ainsi :

comment peut-on estimer θ à partir de n observations $\{X_1, \ldots, X_n\}$ formant un échantillon aléatoire simple dont les valeurs sont notées $\{x_1, \ldots, x_n\}$?

Cette question recouvre **deux problèmes** :

- définir un estimateur possédant de bonnes qualités ou encore de bonnes propriétés statistiques
- trouver la manière adéquate de le choisir.

Soient :

- θ un paramètre réel inconnu défini au sein d'une population U
- Θ l'ensemble des valeurs possibles de θ.

Définition : Si $\{X_1, \ldots, X_n\}$ est un échantillon aléatoire simple d'effectif n prélevé dans U, alors on appelle **estimateur** de θ toute fonction des observations, notée $\widehat{\theta}$:

$$\widehat{\theta} = h(X_1, \ldots, X_n). \qquad (1)$$

On se restreint à des valeurs $\widehat{\theta} \in \Theta$. $\widehat{\theta}$ est une variable aléatoire de loi de probabilité qui dépend du paramètre inconnu θ.

Définition :

On appelle **estimation** de θ une valeur $h(x_1, \ldots, x_n)$ d'un estimateur $\widehat{\theta}$ calculée à partir de n valeurs observées $x_1, \ldots, x_n$ dans un échantillon prélevé dans la population U.

Il est souhaitable de ne pas utiliser uniquement le bon sens ou l'intuition pour choisir entre deux estimateurs.

Pour pouvoir effectuer le bon choix, on doit pouvoir les comparer en recourant à des objectifs choisis a priori.

On va établir une liste de plusieurs propriétés que l'on souhaite retrouver dans un bon estimateur, permettant ainsi de mettre en évidence ceux qui en possèdent sinon le plus grand nombre, du moins les plus importantes.

Toute fonction des observations d'un échantillon aléatoire simple, définie par (1), peut permettre d'estimer la valeur de θ. On a appelé **estimation** la valeur observée $h(x_1, \ldots, x_n)$ de l'**estimateur** $h(X_1, \ldots, X_n)$.

La notation utilisée "majuscule pour une variable aléatoire et minuscule pour sa valeur" est en contradiction avec (1) et veut que l'on distingue aussi la variable aléatoire $\widehat{\theta}$ de sa valeur observée.

Pour ne pas alourdir la notation, et vu la correspondance entre variable aléatoire et valeur observée de cette dernière, on utilise la simplification suivante selon laquelle l'échantillon de taille n est désormais désigné par $\{x_1, \ldots, x_n\}$ sans distinguer explicitement les variables aléatoires X_i, avant observation, de la valeur x_i, après observation.

Attention : toujours vérifier d'après le contexte, si on est dans le premier cas ou dans le second cas.

Le choix d'un estimateur va reposer sur ses qualités. Comme l'on a souligné, il est habituel de comparer des estimateurs entre eux sur la base de propriétés plus ou moins intéressantes qu'ils possèdent ou non.

La première concerne la possibilité de comporter un **biais**. Il est souvent judicieux que la distribution d'un estimateur soit centrée sur le paramètre inconnu, c'est-à-dire qu'il possède la propriété suivante.

Propriété :

$\widehat{\theta}$ est un **estimateur sans biais** (ou non biaisé) du paramètre θ si

$$\mathbb{E}\left[\widehat{\theta}\right] = \theta. \tag{2}$$

Définition : Si la relation (2) n'est pas satisfaite, le **biais** de $\widehat{\theta}$ est alors défini par

$$B\left(\widehat{\theta}\right) = \mathbb{E}\left[\widehat{\theta}\right] - \theta. \tag{3}$$

Remarque :

Pour aborder le caractère biaisé ou non d'un estimateur il faudra utiliser des propriétés de l'espérance mathématique étudiée dans les années précédentes.

Comme nous l'avons vu auparavant, la variance d'un estimateur joue un rôle important dans la mesure de précision.

Propriété : Si $\widehat{\theta}$ est un estimateur sans biais du paramètre θ, on utilise comme mesure de précision sa variance $Var\left[\widehat{\theta}\right]$.

Plus $Var\left[\widehat{\theta}\right]$ est petite, plus l'estimateur $\widehat{\theta}$ est précis. Entre deux estimateurs non biaisés, on choisira le plus précis des deux, c'est-à-dire celui de plus petite variance.

Exemple :

Soit un échantillon aléatoire simple de taille relativement élevée, prélevé dans une population normale $\mathcal{N}(\mu; \sigma^2)$. Si $\overline{x}$ (moyenne de l'échantillon) et $x_{1/2}$ (médiane de l'échantillon) alors, on a

$$\mathbb{E}\left[\overline{x}\right] = \mu, \quad Var\left[\overline{x}\right] = \frac{\sigma^2}{n}$$

et

$$\mathbb{E}[x_{1/2}] = \mu, \quad Var[x_{1/2}] = 1.57\frac{\sigma^2}{n}.$$

On a donc $Var[\overline{x}] < Var[x_{1/2}]$. Donc $\overline{x}$ est un estimateur plus précis que $x_{1/2}$ dans ce cas présent.

Propriété :

Si $\widehat{\theta}$ est un estimateur du paramètre θ, on mesure la précision de l'estimateur $\widehat{\theta}$ par l'**écart quadratique moyen** (EQM) :

$$EQM\left(\widehat{\theta}\right) = Var\left[\widehat{\theta}\right] + B\left(\widehat{\theta}\right)^2.$$

Si $\widehat{\theta}$ est un estimateur sans biais du paramètre θ, ce qui se traduit par

$$B\left(\widehat{\theta}\right) = 0,$$

alors on retrouve la propriété précédente.

Entre deux estimateurs de θ, on choisit celui dont l'écart quadratique moyen est le plus faible.

Pour estimer la moyenne μ d'une population de taille N on utilise la moyenne $\overline{x}$ d'un échantillon de taille n ($n < N$) du type PEAR (probabilités égales avec remise).

Rappel : En tant que variable aléatoire, $\overline{x}$ constitue un **estimateur** de la moyenne μ.

Définition : Toute valeur observée de $\overline{x}$ à partir d'un échantillon est appelée une **estimation** de la moyenne μ.

Le mode de prélèvement retenu est tel que : les observations x_i sont des variables aléatoires indépendantes et identiquement distribuées et sont telles que, pour tout $i = 1, \ldots, n$:

$$\mathbb{E}[x_i] = \mu$$

et

$$Var[x_i] = \sigma^2.$$

Soit $\overline{x}$ définie par

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i, \qquad (4)$$

la moyenne d'un échantillon aléatoire simple prélevé dans une population de moyenne μ.

Calculons l'espérance de $\overline{x}$:

$$\begin{aligned}\mathbb{E}[\overline{x}] &= \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} x_i\right] = \frac{1}{n}\mathbb{E}\left[\sum_{i=1}^{n} x_i\right] = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[x_i] \\ &= \frac{1}{n}\times\mu\times n = \mu.\end{aligned}$$

On trouve ainsi le résultat suivant :

Théorème : En moyenne, l'estimateur $\overline{x}$ est égal à la moyenne μ, ou encore $\overline{x}$ est un estimateur sans biais de la moyenne μ.

En général en statistique, on calcule l'espérance et la variance d'une variable aléatoire. Donc nous allons calculer maintenant la variance.

Calculons la variance de $\overline{x}$:

$$\begin{aligned}Var[\overline{x}] &= Var\left[\frac{1}{n}\sum_{i=1}^{n} x_i\right] \\ &= \frac{1}{n^2}\sum_{i=1}^{n} Var[x_i], \quad \text{car les v.a. st indépendantes} \\ &= \frac{1}{n^2}\sum_{i=1}^{n}\sigma^2, \quad \text{car les v.a. st i.d.} \\ &= \frac{1}{n^2}\times n\times\sigma^2 = \frac{\sigma^2}{n}.\end{aligned}$$

La question :

Que siginifie statistiquement ce dernier résultat ? Comment le statisticien l'interprète-t-il ?

La réponse : Ce paramètre, destiné à connaître la dispersion des valeurs de $\overline{x}$ autour de la moyenne μ, permet de mesurer **l'erreur d'échantillonnage**. Plus $Var[\overline{x}]$ est faible, plus il est probable que l'erreur sera petite et l'estimateur précis. $Var[\overline{x}]$ est faible si la variance σ^2 de la population est petite, ce qui correspond à une population homogène, et/ou si n est grand, c'est-à-dire si la taille de l'échantillon est grande.

Remarque : Cette erreur ne dépend pas de la taille N de la population, ce qui n'est pas intuitif !

Soit s^2_{nc} définie par

$$s^2_{nc} = \frac{1}{n}\sum_1^n (x_i - \overline{x})^2\,, \qquad (5)$$

la variance d'un échantillon aléatoire simple prélevé dans une population U de variance σ^2.

Calculons l'espérance afin de savoir si s^2_{nc} est un estimateur sans biais de σ^2.

$$\begin{aligned}
\mathbb{E}[s^2_{nc}] &= \mathbb{E}\left[\frac{1}{n}\left(\sum_{i=1}^n (x_i-\overline{x})^2\right)\right]\\
&= \mathbb{E}\left[\frac{1}{n}\sum_i x_i^2 - \overline{x}^2\right]\\
&= \frac{1}{n}\mathbb{E}\left[\sum_i x_i^2\right] - \mathbb{E}\left[\overline{x}^2\right]\\
&= \frac{1}{n}\sum_i \left(Var[x_i]+\mu^2\right) - \left(Var[\overline{x}]+\mu^2\right)\\
&= \sigma^2+\mu^2-\frac{\sigma^2}{n}-\mu^2
\end{aligned}$$

On constate ainsi que

$$\mathbb{E}[s^2_{nc}] \neq \sigma^2.$$

Théorème : s^2_{nc} est un estimateur biaisé de la variance σ^2 dont le biais vaut :

$$B(s^2_{nc}) = \sigma^2\left(\frac{n-1}{n}\right) - \sigma^2 = -\frac{\sigma^2}{n}.$$

Remarque : $B(s^2_{nc})$ tend vers 0 quand n tend vers l'infini. On dit dans ce cas que s^2_{nc} est un **estimateur asymptotiquement sans biais**.

À partir de cette remarque, on construit un autre estimateur de la variance σ^2, que l'on appelle la **variance corrigée** de l'échantillon aléatoire simple défini par : **Définition :**

$$s^2_c = \frac{n}{n-1}s^2_{nc} = \frac{1}{n-1}\sum_i (x_i-\overline{x})^2.$$

On vérifie aisément que

$$\mathbb{E}[s^2_c] = \sigma^2.$$

Donc on peut établir le résultat suivant : **Théorème :** La variance corrigée s^2_c est un estimateur sans biais de la variance σ^2.

Si π_A est une proportion d'individus qui possèdent une caractéristique A dans une population U, on peut estimer ce paramètre par la proportion observée dans un échantillon aléatoire simple de taille n prélevé dans cette population U :

$$\widehat{\pi_A} = \frac{n_A}{n}$$

où n_A est le nombre d'individus de l'échantillon qui possèdent une caractéristique A.

π_A peut-être considérée comme la moyenne d'une loi de Bernoulli en dotant tous les individus de la population d'une valeur 1 ou 0 selon qu'ils possèdent ou non la caractéristique A :

$$\begin{aligned}\mu &= \frac{\text{Somme des valeurs 1 + Somme des valeurs 0}}{\text{Taille de la population}}\\ &= \frac{\text{Nombre de 1}}{\text{Taille de la population}}\\ &= \text{Proportion de 1}\\ &= \pi_A.\end{aligned}$$

Si on considère la moyenne $\overline{x}$ de l'échantillon composé de 1 et de 0 :

$$\begin{aligned}\overline{x} &= \frac{\text{Somme des 1 observés + Somme des 0 observés}}{\text{Taille de l'échantillon}}\\ &= \frac{\text{Nombre de 1 observés}}{\text{Taille de l'échantillon}}\\ &= \frac{n_A}{n}.\end{aligned}$$

Comme $\mathbb{E}[\overline{x}] = \mu$, on en déduit que

$$\mathbb{E}\left[\overline{x}\right] = \mathbb{E}\left[\frac{n_A}{n}\right] = \mathbb{E}\left[\widehat{\pi}_A\right] = \mu = \pi_A.$$

Par conséquent, $\widehat{\pi}_A$ est un estimateur sans biais de de π_A.