L3 Economie

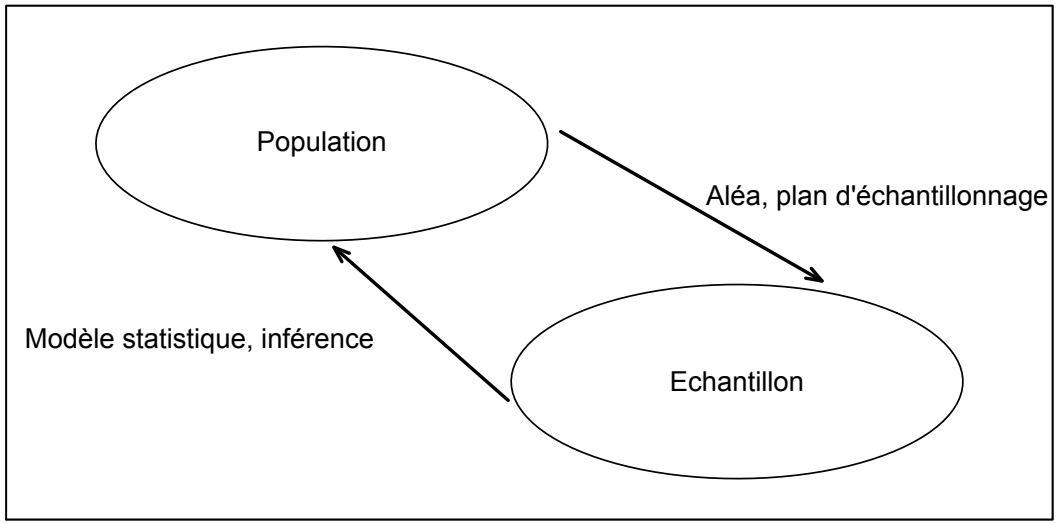
Statistiques inférentielles

Chapitre 2 : Introduction à la statistique inférentielle

Cécile Durot
cecile.durot@gmail.com

Université Paris Nanterre

Erreur de modélisation et d'estimation



Supposant le modèle "vrai", on s'appuie sur l'échantillon pour établir des propriétés de la population en quantifiant l'erreur d'estimation commise. Deux sources d'erreur :

- Erreur de modélisation (aspect simplificateur du modèle)
- Erreur d'estimation (l'échantillon n'est pas la population)

Modèle statistique

Définition : Un modèle statistique consiste en des hypothèses probabilistes formulées sur la loi des observations émanant d'une expérience aléatoire.

Exemples : On choisit au hasard n individus dans une population, avec remise et pour tout $i \in \{1, \ldots, n\}$, on note x_i le salaire net mensuel du i-ème individu choisi. Un exemple de modèle statistique consiste à supposer que $x_1, \ldots, x_n$ sont les réalisations de v.a.r. indépendantes et de même loi $X_1, \ldots, X_n$, définies sur l'univers Ω attaché à l'expérience aléatoire. Un autre modèle statistique consiste à supposer, de plus, que les variables X_i suivent une loi log-gaussienne (i.e. $\log(X_i)$ suit une loi gaussienne).

Remarques :

- Le choix du modèle est guidé par les données, le mode de collecte des données, la connaissance a priori du phénomène.
- Dans ce cours, nous ne travaillerons pas sur la modélisation (ni validation de modèle) : nous nous intéresserons à l'erreur d'estimation dans des modèles classiques.
- Dans ce cours, nous considèrerons uniquement des situations où les données dont des réels $x_1, \ldots, x_n$.

Modèles statistiques étudiés

Le cadre de statistique inférentielle que nous considèrerons dans ce cours est le suivant : nous supposerons que les données collectées sont des nombres réels $x_1, \ldots, x_n$, et que le procédé de collecte est suffisament proche du tirage aléatoire simple (voir plus loin) pour que l'on puisse supposer que

Les données $x_1, \ldots, x_n$ sont les réalisations de v.a.r. $X_1, \ldots, X_n$ indépendantes et de même loi (i.i.d.).

Nous formulerons éventuellement des hypothèses supplémentaires quant à cette loi commune :

- Un modèle est dit paramétrique si cette loi est connue à un nombre fini de paramètres près (e.g. loi de Poisson, loi gaussienne...)
- Un modèle est dit non-paramétrique sinon (e.g. si on ne fait pas d'hypothèse supplémentaire, ou si on suppose seulement que la loi possède une densité...)

Comparaison paramétrique vs non-paramétrique

- Un modèle paramétrique, s'il est correct, est plus facile à ajuster (erreur d'estimation faible).
- Un modèle non-paramétrique est plus flexible (erreur de modélisation faible).

Exemple : Ayant observé les réalisations de $X_1, \ldots, X_n$ v.a.r. i.i.d. de fonction de répartition F_X, on s'intéresse à $F_X(2)$.

- Si $X_1, \ldots, X_n$ suivent une loi $\mathcal{E}(\lambda)$ avec $\lambda > 0$ inconnu, alors $F_X(2) = 1 - \exp(-2\lambda)$ et d'après la LGN, $\bar{X}_n \xrightarrow[n\to\infty]{P} 1/\lambda$ donc $1 - \exp(-2/\bar{X}_n)$ fournit une valeur approchée de (converge en probabilité vers) $F_X(2)$. Ici, $\bar{X}_n = \sum_{i=1}^n X_i/n$.
- Dans tous les cas, avec $Y_i(\omega) = 1$ si $X_i(\omega) \leqslant 2$ et $Y_i(\omega) = 0$ sinon, pour $i = 1, \ldots, n$, les v.a.r. $Y_1, \ldots, Y_n$ sont i.i.d. d'espérance $F_X(2)$ donc par la LGN, $\sum_{i=1}^n Y_i/n$ fournit une valeur approchée de $F_X(2)$.

Quelques exemples de modes d'échantillonnage

- Recensement : non aléatoire, lorsqu'on observe toute la population (cadre de la statistique descriptive)
- Echantillonnage aléatoire simple : consiste à extraire un échantillon de taille n d'une population de taille $N > n$ par des tirages aléatoires équiprobables indépendants (avec remise).
  - ▷ Modèle raisonnable : les données sont les réalisations de v.a. indépendantes et de même loi.
  - ▷ Cadre théorique le plus simple.
  - ▷ Difficulté de mise en oeuvre : nécessite la liste exhaustive des individus de la population.
- Echantillonnage stratifié (méthode des quotas) : La population est découpée selon des caractéristiques connues (e.g. CSP), puis des sous-échantillons sont tirés de sorte que l'échantillon global respecte la composition de la population totale selon ces caractéristiques.

- Tirages aléatoires équiprobables sans remise. Ce mode d'échantillonnage est asymptotiquement équivalent (quand $n \ll N$ et $n \to \infty$) à un échantillonnage aléatoire simple.
- Echantillonnage par grappes : on choisit un échantillon aléatoire de "grappes" (sous-ensembles de la population) et on retient dans l'échantillon tous les individus constituant ces grappes.
  - ▷ Exemple : on divise la ville en quartiers ; on choisit au hasard certains quartiers puis on iterroge toutes les familles résidant dans ces quartiers.
  - ▷ Ne nécessite pas de liste exhaustive.
  - ▷ Difficulté : constitution des grappes (taille des grappes).

Les données

- Nous disposons de données, qui sont des mesures réalisées sur des individus dont l'ensemble constitue l'échantillon étudié.
- Les données décrivent une caractéristique (taille, CSP,...), aussi appelée variable statistique, des individus. On se limite ici au cas d'une variable statistique unidimensionnelle et on notera typiquement $x_1, \ldots, x_n$ les données.
- Une variable statistique peut être qualitative, quantitative discrète ou quantitative continue.

Indicateurs statistiques

Dans le cas de variables quantitatives, des indicateurs numériques permettent de caractériser les données.

Indicateurs de localisation

Le but est de donner un ordre de grandeur général. L'indicateur le plus courant est la moyenne

$$\bar{x}_n = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

Autre indicateur intéressant : la médiane.

Exemple : En 2007 en France, pour un emploi à temps plein, le salaire net mensuel moyen était de 1997 €, et le salaire net mensuel médian de 1594 €. Distribution asymétrique (quelques hauts salaires font monter la moyenne, pas la médiane).

Remarque : Si les données sont les réalisations de v.a.r. $X_1, \ldots, X_n$ i.i.d. d'espérance m, alors par la LGN, on a $\overline{X}_n \xrightarrow[n\to\infty]{P} m$ où $\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$.

Indicateurs de dispersion

Les plus courants sont la variance empirique

$$\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x}_n)^2$$

et l'écart-type empirique, défini comme la racine carrée de la variance empirique.

Remarque : Si les données sont les réalisations de v.a.r. $X_1, \ldots, X_n$ i.i.d. de variance σ^2, alors on peut montrer (Exercice 18) que

$$\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2 \xrightarrow[n\to\infty]{P} \sigma^2$$

et l'écart-type empirique converge en probabilité vers σ.

Motivation

Supposons qu'on observe les réalisations $x_1, \ldots, x_n$ de v.a.r. $X_1, \ldots, X_n$ i.i.d. Notons P_X leur loi commune et F_X leur fonction de répartition. A partir de ces observations, on veut obtenir des informations sur P_X. Est-ce raisonnable ?

Loi empirique

Notation : Pour tout A dans la tribu engendrée par les intervalles dans $\mathbb{R}$, on note $\mathbb{1}_{\{X_i \in A\}}$ la variable aléatoire définie par $\mathbb{1}_{\{X_i \in A\}}(\omega) = 1$ si $X_i(\omega) \in A$ et 0 sinon, pour tout $\omega \in \Omega$.

Définition : La loi empirique (ou loi d'échantillonnage) est définie par

$$P_n(A) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{\{X_i \in A\}},$$

pour tout A dans la tribu engendrée par les intervalles dans $\mathbb{R}$.

Remarque : Pour chaque ω, il s'agit d'une mesure de probabilité, qui pose un poids $1/n$ en chaque $X_i(\omega)$. Le poids total attribué à un réel x est donc k/n s'il y a k indices i pour lesquels $X_i(\omega) = x$:

$$P_n(x) = P_n(\{x\}) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{\{X_i = x\}},$$

Fonction de répartition empirique

Définition : La fonction de répartition associée à la loi empirique est appelée fonction de répartition empirique et est donnée par

$$F_n(t) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{\{X_i \leqslant t\}}$$

pour tout $t \in \mathbb{R}$.

Remarque : Pour chaque ω, il s'agit d'une fonction en escalier puisque la loi empirique est discrète. Les sauts ont lieu en chacun des $X_i(\omega)$ et seulement en ces points.

Propriétés

Pour tout $t \in \mathbb{R}$, on a

- $nF_n(t) \sim \mathcal{B}(n, F_X(t))$.
- $E(F_n(t)) = F_X(t)$.
- $F_n(t) \xrightarrow[n\to\infty]{P} F_X(t)$.
- Soit $X_{(1)}$ la variable aléatoire définie par

$$X_{(1)}(\omega) = \min_{1\leqslant i\leqslant n} X_i(\omega).$$

On définit de même les variables aléatoires $X_{(2)}, \ldots, X_{(n)}$ telles que $X_{(1)} \leqslant \cdots \leqslant X_{(n)}$. Pour tracer la réalisation de F_n, on place d'abord la réalisation de ces v.a.r. sur l'axe des abcisses, ensuite on positionne la valeur de la fonction en ces points puis on prolonge en fonction en escalier. En particulier, si $X_i(\omega) \neq X_j(\omega)$ pour tous $i \neq j$ et tout $\omega \in \Omega$, alors tous les sauts de F_n sont de hauteur $1/n$ et pour tout i, on a $F_n(X_{(i)}) = i/n$.

Théorème de Glivenko-Cantelli

Théorème : Soient $X_1, \ldots, X_n$ des v.a.r. indépendantes et de même loi. Soient F_X leur fonction de répartition et F_n la fonction de répartition empirique. Alors, $\sup_{t\in\mathbb{R}} |F_n(t) - F_X(t)| \xrightarrow[n\to\infty]{P} 0$.

Remarques :

- Ce résultat est plus fort que la convergence $F_n(t) \xrightarrow[n\to\infty]{P} F_X(t)$ pour $t \in \mathbb{R}$: il implique que pour $t_n \in \mathbb{R}$ pouvant dépendre de n, on a $F_n(t_n) - F_X(t_n) \xrightarrow[n\to\infty]{P} 0$.
- Il implique que

$$\sup_{A \text{ intervalles}} |P_n(A) - P_X(A)| \xrightarrow[n\to\infty]{P} 0$$

où P_n est la loi empirique et P_X la loi commune de $X_1, \ldots, X_n$.

Conclusion

Le théorème de Glivenko-Cantelli implique que pour de grands échantillons où l'on peut considérer que les observations sont les réalisations de v.a.r. indépendantes et de même loi, la loi empirique (loi d'échantillonnage) fournit une bonne approximation de la loi des observations :

$$P_n(A) \approx P_X(A),$$

pour tout A de la tribu engendrée par les intervalles dans $\mathbb{R}$, A pouvant dépendre de n. Cela justifie la démarche statistique.