E. Théorème de la Limite Centrale ou Central-Limite (T.C.L)

Soit X_1, X_2,....,X_n une suite de variables aléatoires **indépendantes** et **de même loi**. Notons $\mu = E(X_i)$ et $\sigma^2 = V(X_i)$. Si on pose $S_n = X_1 + ... + X_n$ alors,

$$\sqrt{n}\left(\frac{S_n}{n} - \mu\right) \longrightarrow \mathcal{N}(0, \sigma^2)$$

en loi. $\mathcal{N}(0, \sigma^2)$ est la loi normale centrée de variance σ^2.

Autre écriture :

$$\frac{S_n - n\mu}{\sigma\sqrt{n}} \longrightarrow \mathcal{N}(0, 1)$$

Remarque

- La loi normale apparaît comme loi limite quelle que soit la loi des X_i (il faut juste que μ et σ^2 existent). Ceci montre le caractère universel des lois normales et leur importance.

- La loi normale est la loi limite quand $n \longrightarrow \infty$. En pratique, l'approximation de la loi de la somme par une loi normale est bonne lorsque $n \geq 30$.

Exercice (E.5)

On lance un dé honnête 100 fois, de façons indépendante.
Quelle est la probabilité que la somme totale des points obtenus soit comprise entre 300 et 400 ?

solution

Si X_i représente le nombre de points obtenus au i-ème lancer, alors

$$S_{100} = \sum_{i=1}^{100} X_i$$

est la somme totale des points obtenus après 100 lancers. On sait que $E(X_i) = 3.5$ et $V(X_i) = 35/12$. Par le TCL, comme $n \geq 30$ on sait que l'on peut approcher la loi de la variable S_n centrée et réduite par :

$$\frac{S_{100} - 3.5 \times 100}{10 \times \sqrt{35/12}} \sim \mathcal{N}(0,1).$$

$$\begin{aligned} P(300 \leq S_{100} \leq 400) &= P\left(\frac{300-350}{10\sqrt{35/12}} \leq \frac{S_{100}-350}{10\sqrt{35/12}} \leq \frac{400-350}{10\sqrt{35/12}}\right) \\ &= P(-2.93 \leq Z \leq 2.93) \text{ où } Z \sim \mathcal{N}(0,1) \\ &= 0.9966 \end{aligned}$$

Application

Si $X_i \sim \mathcal{B}(1, p)$, on a vu que $\sum X_i \sim \mathcal{B}(n, p)$.
Une simple application du TCL, nous dit que l'on peut approcher la loi binomiale $\mathcal{B}(n, p)$ par la loi normale $\mathcal{N}(np, np(1-p))$.

Exercice (E.6)

(Style Exercice 8-Feuille 5)

A Lille, des enregistrements climatiques indiquent qu'en moyenne 17 des 31 jours du mois d'Octobre sont pluvieux. On considère les épisodes de pluie journaliers comme des épreuves indépendantes. On note N le nombre de jours pluvieux au cours du mois d'Octobre.

1. Quelle est la loi de N ?
2. Peut-on utiliser l'approximation de la loi de Poisson ?
3. En utilisant l'approximation normale, quelle est la probabilité d'avoir entre 15 et 20 jours pluvieux au mois d'Octobre ?

Théorie de l'estimation

On considère n v. a. $X_1, \cdots, X_n$ indépendantes et de même loi qui constitue un échantillon.
La théorie de l'estimation cherche à exploiter au mieux l'information d'un échantillon pour obtenir des estimateurs des caractéristiques d'une population (moyenne, variance, probabilité qu'un événement se produise, etc).
On distingue :

- la théorie de l'estimation ponctuelle qui cherche à déterminer un estimateur d'une caractéristique de la population.
- la théorie de l'estimation par intervalle qui à partir d'un estimateur donné cherche à donner des fourchettes de variations pour la caractéristique.

1. Généralités sur l'échantillonage

La statistique descriptive a permis de calculer des caractéristiques pour décrire l'échantillon comme la moyenne $\bar{x}$ et la variance s^2. (par exemple : $\bar{x} = 3466\,g$ poids des 100 bébés d'une maternité)
Les caractéristiques correspondantes de la population globale sont l'espérance μ et la variance σ^2. On souhaiterait connaître ces caractéristiques théoriques mais comme la population entière est inobservable, on doit se contenter des caractéristiques de l'échantillon.

Exemple du poids moyen des bébés

Remarque

Question : En quoi la valeur de $\bar{x} = 3466\,g$ observée sur l'échantillon des 100 bébés nous renseigne t'elle sur la valeur de la caractéristique du poids de toute la population des bébés ?

• Un autre échantillon de 100 bébés donnerait sans doute une valeur différente de $3466\,g$.

Lien entre la statistique descriptive et l'estimation : On suppose que les valeurs $x_1, \cdots, x_n$ du poids des 100 bébés sont les réalisations de $n = 100$ v.a. $X_1, \cdots, X_n$ indépendantes et de même espérance μ inconnue. Ainsi la valeur $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = 3466\,g$ est la réalisation pour l'échantillon des 100 bébés de la v.a.

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

Ainsi, un autre échantillon de 100 bébés donnerait une autre valeur de $\bar{X}$.

$\bar{X}$ est une variable aléatoire dont nous allons préciser la loi de probabilité.

2. Loi de $\bar{X}$

Comme $X_1, \cdots, X_n$ sont indépendantes et de même loi, d'espérance μ et de variance σ^2, on a :

$$E(\bar{X}) = E\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n}\sum_{i=1}^{n} E(X_i) = \frac{1}{n}(\mu + \cdots + \mu) = \mu.$$

et

$$V(\bar{X}) = V\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\sum_{i=1}^{n} V(X_i) = \frac{\sigma^2}{n}.$$

Si $\sigma^2 = 1$ avec un échantillon de taille $n = 100$, la variance de $\bar{X}$ vaut $0,01$. Les valeurs $\bar{x}$ prises par la v.a. $\bar{X}$ vont donc se disperser autour de μ avec une faible dispersion (écart-type égal à $\sqrt{0.01} = 0.1$)

2. Loi de $\bar{X}$ (suite)

Que peut-on dire de la loi de $\bar{X}$?

On a 2 résultats très importants que nous admettrons :

Théorème (moyenne de v.a. normales indépendantes)

Si les v.a. $X_1, \cdots, X_n$ sont indépendantes et suivent la même loi normale $\mathcal{N}(\mu, \sigma^2)$ alors pour tout $n \geq 1$

$$\bar{X} \sim \mathcal{N}(\mu, \frac{\sigma^2}{n})$$

Théorème ("central-limite" T.C.L.)

Si les v.a. $X_1, \cdots, X_n$ sont indépendantes et suivent la même loi quelconque d'espérance μ et de variance σ^2 alors pour n assez grand ($n \geq 30$)

$$\bar{X} \sim \mathcal{N}(\mu, \frac{\sigma^2}{n})$$

Théorème (T.C.L. pour n v.a. de Bernoulli indépendantes)

Soient $X_1, \cdots, X_n$ n v.a. indépendantes de Bernoulli de paramètre p (d'espérance p et de variance pq alors pour n assez grand ($n \geq 30$)

$$\bar{X} \sim \mathcal{N}(p, \frac{pq}{n})$$

Interprétation : La proportion de succès $\bar{X} = \frac{\text{nbre de succès}}{n}$ suit une loi normale $\mathcal{N}(p, \frac{pq}{n})$.
On peut formuler ce résultat de la façon suivante :
Le nombre de succès dans le schéma de Bernoulli

$$\sum_{i=1}^{n} X_i \sim \mathcal{B}(n, p) \text{ peut être approchée par une loi } \mathcal{N}(np, npq)$$

De manière général, si on a un échantillon X_1,...,X_n issus d'une loi dépendant d'un paramètre θ, alors un estimateur du paramètre θ est une fonction de X_1,...,X_n $g(X_1, ..., X_n)$ dont la loi dépend de θ. On contrôle alors la qualité de l'estimateur en regardant son biais et sa variance.

- Si $E(g(X_1, ..., X_n)) = \theta$ on dit que l'estimateur est sans biais.
- La variance $V(g(X_1, ..., X_n))$ permet de contrôler la qualité de l'estimateur : plus elle est petite, plus l'estimateur sera précis.

Exercice (E.1)

(difficile) Montrer que si $X_1,\ldots,X_n$ est un échantillon de variables aléatoires telles que $V(X_1) = \sigma^2$ alors

$$\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2$$

est un estimateur biaisé de σ^2. Par contre

$$\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$$

est un estimateur sans biais de σ^2.

1. Estimation par intervalle de confiance

Un intervalle de confiance est un intervalle $[B_I; B_S]$ dont les bornes sont des fonctions de X_1,...,X_n. Il est construit pour contenir avec une probabilité égale à $1 - \alpha$ (fixé à l'avance) le paramètre θ que l'on cherche à estimer.
En d'autres termes, on cherche $[B_I; B_S]$ tels que :

$$P(\theta \in [B_I; B_S]) = P(B_I \leq \theta \leq B_S) = 1 - \alpha$$

Loi normale de variance connue

Supposons $X_1, ..., X_n$ un échantillon de loi normale $\mathcal{N}(\mu, \sigma^2)$ et que l'on veuille construire un intervalle de confiance pour μ.
D'après le cours précédent, on sait que :

$$\sum_{i=1}^{n} X_i \sim \mathcal{N}(n \times \mu, n \times \sigma^2)$$

Ce qui signifie, si on note $z_{1-\alpha/2}$ la quantité telle que $\phi(1-\alpha/2) = 1-\alpha/2$, alors :

$$P\left(z_{1-\alpha/2} \leq \frac{\frac{1}{n}\sum_{i=1}^{n} X_i - \mu}{\sigma/\sqrt{n}} \leq z_{1-\alpha/2}\right) = 1-\alpha$$

Loi normale de variance connue

En re-organisant, on obtient :

$$P\left(\overline{X} - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}} \leq \mu \leq \overline{X} + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha$$

On est donc sûr avec une probabilité de $1 - \alpha$ que la vraie moyenne se trouve entre $\overline{X} - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}$ et $\overline{X} + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}$.
Pour obtenir une estimation de cet intervalle de confiance, on remplace les variables aléatoires X_i par leur réalisation x_i.

Exercice (E.3)

On a observé la taille de 200 hommes américains adultes. Après calcul, on a obtenu une moyenne de 68.8 inches. Si on suppose que la taille d'un américain suit une loi normale, donnez un intervalle de confiance à 95% de la vraie moyenne.

2. Estimation par intervalle de confiance pour l'espérance μ

On suppose maintenant que l'on dispose d'un échantillon $X_1, ..., X_n$ de loi quelconque avec $E(X_1) = \mu$ et $V(X_1) = \sigma^2$. Par le théorème centrale limite, on sait que

$$\overline{X} \longrightarrow \mathcal{N}(\mu, \sigma^2/n)$$

On peut donc dire que, si n est suffisamment grand que

$$P\left(\overline{X} - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}} \leq \mu \leq \overline{X} + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha$$

Le problème c'est qu'il y a σ^2 qui est inconnu dans les bornes ! !
On remplace alors les X_i par les réalisations x_i, et σ^2 par un estimateur sans biais : $\widehat{\sigma^2} = \frac{1}{n-1} \sum i = 1^n (X_i - \overline{X})^2$.
On obtient alors l'intervalle de confiance au niveau $1 - \alpha$:

$$\overline{x} - z_{1-\alpha/2} \frac{\widehat{\sigma}}{\sqrt{n}} ; \overline{x} + z_{1-\alpha/2} \frac{\widehat{\sigma}}{\sqrt{n}}$$

Exercice (E.3)

On a mesuré la taille de 200 femmes américaines. On a obtenu les valeurs suivantes :

$$\overline{x} = 61.2 \quad \text{et} \quad \widehat{\sigma^2} = 16.5$$

Donner un intervalle de confiance à 95% de la vraie moyenne des femmes américaines. Avant l'étude, un spécialiste avait dit que la taille moyenne des américaines était de 65. Que peut-on conclure de l'étude ?

3. Estimation par intervalle de confiance pour une proportion

On suppose maintenant que l'on dispose d'un échantillon $X_1, ..., X_n$ de loi de Bernoulli avec $E(X_1) = p$ et $V(X_1) = np(1-p)$. Par l'application théorème centrale limite qui nous donne une approximation de la binomiale, on sait que

$$\sum X_i \longrightarrow \mathcal{N}(np, np(1-p))$$

On peut donc dire que, si n est suffisamment grand que

$$P\left(\widehat{p} - z_{1-\alpha/2}\frac{\sqrt{p(1-p)}}{\sqrt{n}} \leq \mu \leq \widehat{p} + z_{1-\alpha/2}\frac{\sqrt{p(1-p)}}{\sqrt{n}}\right) = 1-\alpha$$

où $\widehat{p}$ est la proportion observée, c'est à dire $\frac{1}{n}\sum_{i=1} X_i$

Le problème c'est qu'il y a p qui est inconnu dans les bornes ! !
On remplace alors le $p(1-p)$ inconnu par son estimation et on obtient :
On obtient alors l'intervalle de confiance au niveau $1-\alpha$:

$$\widehat{p} - z_{1-\alpha/2}\frac{\widehat{p}(1-\widehat{p})}{\sqrt{n}}; \widehat{p} + z_{1-\alpha/2}\frac{\widehat{p}(1-\widehat{p}}{\sqrt{n}}$$

Exercice (E.4)

A la veille d'une élection présidentielle qui s'annonce particulièrement serrée entre les candidats A et B, on eeffectue un sondage sur 100 personne, et on obtient 53% d'intentions de vote pour le candidat A contre 47% pour le candidat B.
Donner un intervalle de confiance pour la vraie proportion de personne qui ont l'intention de voter pour A. Commenter. Quelle devrait être la taille de l'échantillon pour que l'on soit sûr à 95% qu'un des deux candidats va gagner ?