IV - Notions de statistique

Échantillonnage

Soit *f(x)* la densité de probabilité des valeurs possibles de la variable aléatoire observable *x* décrivant un processus aléatoire.

***f*(*x*) décrit la population infinie des observations potentielles de *x*.**

Expérience = échantillonnage aléatoire non - biaisé de *f*(*x*) de taille limitée à *n* observations $(x_1, x_2, \ldots x_n)$

L'échantillon ne diffère de la population que par les fluctuations aléatoires (statistiques) liées à sa taille finie.

Si l'échantillonnage est biaisé, le biais doit être ou négligeable par rapport à la précision que l'on veut obtenir, ou alors corrigible pour que l'échantillon soit représentatif de la population.

Taille des fluctuations statistiques : Inégalité de Bienaymé - Chebyshev

Quelle que soit la *fdp*, la fréquence d'occurence de x hors de $[\mu-\lambda\sigma,\mu+\lambda\sigma]$ est $\leq \frac{1}{\lambda^2}$

$$P\left(|x-\mu| \geq \lambda\sigma\right) \leq \frac{1}{\lambda^2} \qquad P\left(|x-\mu| \geq 3\sigma\right) \leq \frac{1}{9}$$

Les écarts à la moyenne μ sont mesurés en nombre λ d'écarts types σ.

$$P\left(|x-\mu| \geq \varepsilon\right) \leq \left(\frac{\sigma}{\varepsilon}\right)^2$$ **si les écarts ε sont mesurés de manière absolue.**

$$\int_{-\infty}^{\infty}\left(1-\frac{(x-\mu)^2}{(\lambda\sigma)^2}\right)f(x)dx = 1-\frac{\sigma^2}{(\lambda\sigma)^2} = 1-\frac{1}{\lambda^2} = \int_{-\infty}^{\mu-\lambda\sigma} + \int_{\mu+\lambda\sigma}^{\infty} + \int_{\mu-\lambda\sigma}^{\mu+\lambda\sigma}\left\{\left(1-\frac{(x-\mu)^2}{(\lambda\sigma)^2}\right)f(x)dx\right\}$$

$$\overbrace{\frac{(x-\mu)^2}{(\lambda\sigma)^2} \geq 1}^{\uparrow \quad \uparrow}$$

$$\Rightarrow 1-\frac{1}{\lambda^2} \leq \int_{\mu-\lambda\sigma}^{\mu+\lambda\sigma}\left(1-\frac{(x-\mu)^2}{(\lambda\sigma)^2}\right)f(x)dx \leq \int_{\mu-\lambda\sigma}^{\mu+\lambda\sigma} f(x)dx$$

Concepts de statistique et d'estimateur

Une statistique : une variable aléatoire ne dépendant que de l'échantillon d'observations et de paramètres de valeur connue.

Un estimateur : une statistique permettant de faire une estimation $\hat{\theta}$ d'un paramètre θ de valeur inconnue θ_0 et que l'on veut mesurer.

Un estimateur non biaisé : un estimateur conduisant à une mesure $\hat{\theta}$ d'un paramètre θ qui ne diffère de la vraie valeur θ_0 inconnue que par les fluctuations statistiques dues à la dimension finie de l'échantillon. La moyenne de toutes les valeurs possibles de $\hat{\theta}$ est égale à θ_0

$$E\left[\hat{\theta}\right] = \theta_0$$

Un estimateur cohérent : un estimateur tel que, à la limite des grands échantillons, toute mesure $\hat{\theta}$ tend vers la vraie valeur θ_0

$$lim_{n\to\infty}\ \hat{\theta} = \theta_0$$

• Estimateur de la moyenne de l'échantillon

$$\hat{\mu} = \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

$\bar{x}$ est une statistique

$\bar{x}$ est une un estimateur (une mesure) $\hat{\mu}$ de la moyenne μ de la population

$\bar{x}$ est un estimateur non biaisé: $E\left[\bar{x}\right] = E\left[\frac{1}{n}\sum_{i=1}^{n} x_i\right] = \frac{1}{n}\sum_{i=1}^{n} E\left[x_i\right] = \mu$

$\bar{x}$ est un estimateur cohérent: $lim_{n\to\infty}\ \bar{x} = \mu$

démonstration intuitive: $\bar{x} = \sum_{i=1}^{n} a_i x_i$, $a_i = \frac{1}{n} \rightarrow \sigma_{\bar{x}}^2 = \frac{\sigma^2}{n}$

$\rightarrow$ la dispersion sur $\bar{x}$ décroit comme $1/\sqrt{n}$

démonstration exacte :

résulte directement de l'inégalité de Bienaymé-Chebyshev

• Estimateur de la variance de l'échantillon si la moyenne μ est inconnue

$$\hat{\sigma}^2 = s^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2$$

s^2 est une **statistique**: une variable aléatoire ne dépendant que des mesures $\underline{x}$

s^2 est un **estimateur** (une mesure) $\hat{\sigma}^2$ de la variance σ^2 de la population

s^2 est un **estimateur cohérent**: $lim_{n\to\infty}\, s^2 = \sigma^2$

démonstration intuitive: on peut montrer que $\sigma^2_{\sigma^2} = \frac{\left(m_4 - \sigma^{2^2}\right)}{n} + \frac{2\sigma^{2^2}}{(n-1)n}$

$\rightarrow$ dispertion sur σ^2 décroit avec n

démonstration exacte: théorèmes de convergences similaires à l'inégalité de Bienaymé-Chebyshev.

s^2 est un estimateur **non biaisé** : $E\left[s^2\right] = E\left[\frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2\right] = \sigma^2$

Démonstration intuitive:

La moyenne s'obtient en divisant par $(n-1)$ et non par n.

Seules $(n-1)$ valeurs sont indépendantes puisque $x_n = n\,\bar{x} - \sum_{i=1}^{n-1} x_i$

Démonstration exacte: soit la transformation de Helmert

$$\begin{cases} y_i = \dfrac{\sum_{j=1}^{i-1} x_j - (i-1)\,x_i}{\sqrt{(i-1)\,i}} \quad i = 1, n-1 \begin{cases} \mu_{y_i} = E\left[y_i\right] = \dfrac{(i-1)\,\mu - (i-1)\,\mu}{\sqrt{(i-1)\,i}} = 0 \\ \sigma^2_{y_i} = E\left[(y_i)^2\right] = \dfrac{(i-1)\,\sigma^2 + (i-1)^2\,\sigma^2}{(i-1)\,i} = \sigma^2 \end{cases} \\ y_n = \dfrac{x_1 + x_2 + \ldots + x_n}{\sqrt{n}} = \sqrt{n}\,\bar{x} \end{cases}$$

$$\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2 = \sum_{i=1}^{n} x_i^2 - n\,\bar{x}^2 = \sum_{i=1}^{n} y_i^2 - n\,\bar{x}^2 = \sum_{i=1}^{n} y_i^2 - y_n = \sum_{i=1}^{n-1} y_i^2$$

On a $\sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} y_i^2$ parce que la transformation est orthonormée:

$$y_i = \sum_{j=1}^{n} a_{ij}\,x_j \quad \text{avec} \quad \sum_{i=1}^{n} a_{ij}\,a_{ik} = \delta_{jk}$$

$$E\left[\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2\right] = E\left[\sum_{i=1}^{n-1} y_i^2\right] = (n-1)\,\sigma^2 \Rightarrow E\left[s^2\right] = \frac{(n-1)\,\sigma^2}{(n-1)} = \sigma^2$$

• Estimation de la variance de l'échantillon si la moyenne μ est connue

$$\hat{\sigma}^2 = S^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2$$

S^2 est un estimateur **non biaisé**: $E\left[S^2\right] = \sigma^2$

S^2 est un estimateur **cohérent**: $lim_{n\to\infty}\, S^2 = \sigma^2$

• Relation entre les estimateurs $\bar{x}$, s^2 et S^2

On vérifie facilement que $nS^2 = (n-1)s^2 + n(\bar{x} - \mu)^2$

Loi des Grands Nombres

Application de l'inégalité de Bienaymé-Chebyshev $P\left(|x-\mu| \geq \lambda\sigma\right) \leq \frac{1}{\lambda^2}$

à la moyenne $\bar{x}$ **de l'échantillon de variance** $\sigma_{\bar{x}}^2 = \frac{\sigma^2}{n}$

$$P\left(|\bar{x}-\mu| \geq \lambda\sigma\right) = P\left(|\bar{x}-\mu| \geq \lambda\sqrt{n}\sigma_{\bar{x}}\right) = \leq \frac{1}{n\lambda^2} \quad \Rightarrow \quad P\left(|\bar{x}-\mu| \geq \varepsilon\right) = \frac{\sigma^2}{n\varepsilon^2}$$

Etant donnée la largeur de la fonction de densité de probabilité de la population, mesurée par sa variance σ^2 **:**
la probabilité d'observer une valeur de $\bar{x}$ **éloignée de** μ **de plus de** ε **peut être rendue arbitrairement petite en choissant une taille** n **d'échantillon suffisamment grande.**

La précision avec laquelle on détermine $\bar{x}$ **est proportionnelle à l'écart type de la poulation** σ **et à l' inverse de la racine carrée de la taille de l'échantillon** n**.**

Version faible de la Loi des Grands Nombres:

Si σ^2 **n'existe pas parce que** $E\left[\left(x-\mu\right)^2\right]$ **diverge :** $lim_{n\to\infty}\ P\left(|\bar{x}-\mu| \geq \varepsilon\right) = 0$