Statistique Inférentielle

[L2] [2017/2018]

1. Echantillon

On a un ensemble d'**individus** Ω, appelé **population**. On observe sur celle-ci des variable aléatoire $X \to \mathbb{N}, \mathbb{Z}, \mathbb{R}$. La variable est dite **quantitative** si l'ensemble d'arrivée est numérique, elle est **qualitative** dans le cas contraire.

On considère n individus $\omega_1, \cdots, \omega_n \in \Omega$ formant un **échantillon**. On **observe** alors les valeurs $x_1, \cdots, x_n = X(\omega_1), \cdots, X(\omega_n)$.

Une variable quantitative X est **discrète** si $X(\Omega)$ est inclus dans un ensemble E discret, c'est-à-dire un ensemble pour laquelle la seule topologie est $\mathcal{P}(E)$. Elle est **continue** dans le cas contraire. $\mathbb{N}, \mathbb{Z}$ sont discrets, à l'inverse de $\mathbb{Q}, \mathbb{R}, \mathbb{C}$.

Dans le cas de valeurs discrètes, on peut regrouper les observations faites sur l'échantillon. On trouve k valeurs distinctes observées $(\tilde{x}_1, \cdots, \tilde{x}_k)$, chacune étant observée n_j fois, de telle sorte que $\sum_{j=1}^{k} n_j = n$, c'est-à-dire la taille de l'échantillon.

Il arrive qu'on traite des variables continues comme des variables discrètes, en groupant les observations dans des **classes**, typiquement des intervalles. Exemple : tranches d'âge.

On appelle **moyenne de l'échantillon** $\bar{x} := \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{n}\sum_{j=1}^{k} n_j \tilde{x}_j$

On appelle **variance de l'échantillon**

$$s_x^2 := \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{n}\sum_{j=1}^{k} n_j(\tilde{x}_j - \bar{x})^2$$

$$= \frac{1}{n}\left(\sum_{i=1}^{n} x_i^2\right) - \bar{x}^2 = \frac{1}{n}\left(\sum_{j=1}^{n} n_j x_j^2\right) - \bar{x}^2$$

La variance est nulle ssi tous les x_i sont égales, auquel cas on aurait $\bar{x} = x_i$.

On appelle **écart-type de l'échantillon** $s_x := \sqrt{s_x^2}$

On prend pour hypothèse que les x_i sont les observations d'une même variable aléatoire et que ces observations sont mutuellement indépendantes.
On va supposer que l'on connaît la nature de cette variable aléatoire, (e.g. on va supposer qu'elle est gaussienne), et on essaie alors d'en estimer les paramètres.

△ On appelle **échantillon théorique de taille** n de la variable aléatoire X un n-uplet (X_n) de variables aléatoires mutuellement indépendantes et de même loi que X. Cet échantillon théorique a une moyenne

$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ qui est une variable aléatoire.

2. Intervalle de confiance pour la moyenne d'un échantillon Gaussien

La moyenne $\bar{x}$ de cet échantillon étant supposée connue.
On sait que la moyenne et variance de la somme de deux variable aléatoire est la somme de l'une ou l'autre. Si les lois sont Gaussiennes, la somme des lois est encore Gaussienne.
Si les X_i suivent une loi $\mathcal{N}(\mu, \sigma^2)$ alors $\bar{X} \sim \mathcal{N}(\mu, \sigma^2/n)$.

On recherche un intervalle I centré en μ tel que $\mathbb{P}(\bar{X} \in I) \geq 1 - \alpha$ où $1 - \alpha$ est le **niveau de confiance**.
On pose $Z := \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$ qui est la variable centrée réduite associée à $\bar{X}$. Alors $Z \sim \mathcal{N}(0, 1^2)$.

On considère $F_Z(t) = \Pi(t) = 1 - \Pi(-t)$. On cherche β tel que

$$1 - \alpha = \mathbb{P}(Z \in [-\beta, \beta]) = \Pi(\beta) - \Pi(-\beta) \quad \blacktriangleright\blacktriangleright \quad 2\Pi(\beta) - 1 = 1 - \alpha \quad \blacktriangleright\blacktriangleright \quad \Pi(\beta) = 1 - \frac{\alpha}{2}$$

On trouve β en utilisant les tables de Π, puis :

$$P(Z \in [-\beta, \beta]) = P\left(\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \in [-\beta, \beta]\right) = P\left(\bar{X} \in \left[\mu - \beta\frac{\sigma}{\sqrt{n}}, \mu + \beta\frac{\sigma}{\sqrt{n}}\right] =: I\right)$$

On suppose alors, au niveau de confiance $1 - \alpha$, que la moyenne observée $\bar{x}$, en tant que réalisation de $\bar{X}$, appartient à cet intervalle I, d'où l'estimation de μ :

$$\bar{x} \in I \iff \mu \in \left[\bar{x} - \beta\frac{\sigma}{\sqrt{n}}, \bar{x} + \beta\frac{\sigma}{\sqrt{n}}\right]$$

Le niveau de confiance $1 - \alpha$ n'intervient qu'à travers la valeur de $\beta = \Pi^{-1}\left(1 - \frac{\alpha}{2}\right)$.

Last updated 2018-04-10 23:13:07 CEST