STT 1000 - STATISTIQVES

ARThVR ChARPENTIER

Fréquence

Considérons un échantillon $\{x_1, \cdots, x_n\}$, prenant des valeurs A ou B (voire davantage). Supposons que l'on s'intéresse à la fréquence d'apparition de la modalité A.

Notons $y_i = \mathbf{1}_A(x_i)$, et $\{y_1, \cdots, y_n\}$ l'échantillon prenant les valeurs 0 ou 1. La fréquence (d'apparition de A) est

$$f = \frac{1}{n}\sum_{i=1}^{n} y_i = \overline{y} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_A(x_i)$$

(on parle aussi parfois de proportion)
Considérons maintenant une collection de variables aléatoires indépendantes et identiquement distribuées, $Y_1, \cdots, Y_n$, de loi $\mathcal{B}(p)$. Posons

$$F = \frac{1}{n}\sum_{i=1}^{n} Y_i = \overline{Y}$$

Fréquence

Si les variables $Y_1, \cdots, Y_n$ sont i.i.d. de loi $\mathcal{B}(p)$

$$\mathbb{E}[F] = p \text{ et } \text{Var}[F] = \frac{p(1-p)}{n}$$

Plus précisément, comme $nF \sim \mathcal{B}(n, p)$,

$$\mathbb{P}\left(F = \frac{k}{n}\right) = \binom{n}{k} p^k (1-p)^{n-k}$$

Si n est suffisemment grand, d'après le théorème central limite

$$Z_n = \sqrt{n}\frac{F-p}{\sqrt{p(1-p)}} \xrightarrow{\mathcal{L}} \mathcal{N}(0,1)$$

En pratique, on suppose l'approximation normale valide si $n \geq 30$, $np \geq 15$ et $n(1-p) \geq 15$