

Option Statistique 6. Normalité et droite de Henry

Variable aléatoire normale centrée réduite : On dit que T est une variable aléatoire (v.a.) normale centrée réduite et on note $T \rightsquigarrow \mathcal{N}(0, 1)$ si et seulement si

$$\mathbb{P}(T \leq t) = N(t) \text{ , où } N(t) := \int_{-\infty}^t n(s)ds \text{ , avec } n(t) := \frac{1}{\sqrt{2\pi}}e^{-\frac{t^2}{2}}.$$

On vérifie que si $T \rightsquigarrow \mathcal{N}(0, 1)$, alors $\mathbb{E}T = 0$ et $\text{Var } T = 1$.

Sous Excel, la valeur de $N(t)$ est donnée par `LOI.NORMALE.STANDARD(t)`.

v.a. normale : On dit que X est une v.a. normale et on note $X \rightsquigarrow \mathcal{N}(\mu, \sigma)$ si et seulement si $T := \frac{X-\mu}{\sigma} \rightsquigarrow \mathcal{N}(0, 1)$. Comme dans ce cas $X = \sigma T + \mu$, on voit que $\mathbb{E}X = (\sigma T + \mu) = \sigma \mathbb{E}T + \mu = \mu$, et $\text{Var } X = \text{Var}(\sigma T + \mu) = \text{Var}(\sigma T) = \sigma^2 \text{Var } T = \sigma^2$. Les paramètres μ et σ ne sont donc autres que l'espérance et l'écart-type de la v.a. X . Le changement de variable $t = \frac{x-\mu}{\sigma}$ montre que

$$N_{\mu\sigma}(x) := \mathbb{P}(X \leq x) = \int_{-\infty}^x n_{\mu\sigma}(s)ds \text{ , avec } n_{\mu\sigma}(x) := \frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(x-\mu)^2}{2\sigma^2}}.$$

Sous Excel, la valeur de $N_{\mu\sigma}(x)$ est donnée par `LOI.NORMALE(x, μ, σ, VRAI)`, et la valeur de $n_{\mu\sigma}(x)$ est donnée par `LOI.NORMALE(x, μ, σ, FAUX)` ; le paramètre booléen `VRAI-FAUX` indique qu'on veut la fonction de répartition (ou de "probabilité cumulative") $N_{\mu\sigma}$ ou "au contraire" sa densité $n_{\mu\sigma}$.

Droite de Henry : Pour une suite finie de valeurs $x_1 < x_2 < \dots < x_N$ notons $y_i := F(x_i)$ les fréquences cumulées d'un échantillon suivant la loi X . La loi des grands-nombres assure que, pour un échantillon suffisamment grand de tirages indépendants de X on a $F(x_i) \simeq \mathbb{P}(X \leq x_i)$ et donc, en posant $y_i := F(x_i)$,

$$y_i = F(x_i) \simeq \mathbb{P}(X \leq x_i) = \mathbb{P}\left(\frac{X-\mu}{\sigma} \leq \frac{x_i-\mu}{\sigma}\right) = \mathbb{P}(T \leq t_i) \stackrel{?}{=} N(t_i),$$

où $\stackrel{?}{=}$ est une égalité si on a bien $X \rightsquigarrow \mathcal{N}(\mu, \sigma)$. On calcule donc les $t_i := N^{-1}(y_i)$, et si les $\simeq$ et $\stackrel{?}{=}$ étaient des égalités, on devrait donc avoir $t_i = \frac{x_i-\mu}{\sigma}$; en pratique on aura

$$t_i = ax_i + b + \varepsilon_i \text{ , avec } a = \frac{1}{\sigma} \text{ , et } b = -\frac{\mu}{\sigma},$$

et on choisit a et b par la méthode des MCO. La droite ainsi obtenue est appelée *droite de Henry*.

Rappelons qu'on teste la qualité du modèle par le calcul de la corrélation empirique des deux vecteurs $x = (x_1, \dots, x_N)$ et $y = (y_1, \dots, y_N)$:

$$\text{Cor}(x, t) := \frac{\text{Cov}(x, t)}{\sqrt{\text{Var } x \text{Var } t}}.$$

On a toujours $|\text{Cor}(x, t)| \leq 1$; le modèle est d'autant meilleur que $|\text{Cor}(x, t)|$ est proche de 1.

Sous Excel, la valeur de $N_{\mu\sigma}^{-1}(y)$ est donnée par `LOI.NORMALE.INVERSE(y, μ, σ)`, et donc $N^{-1}(y)$ est donnée par `LOI.NORMALE.INVERSE(y, 0, 1)`.