Calcul de la variance : Pourquoi (n – 1) ?

Comme je vous l'ai dit en cours, nous calculons la variance sur (n – 1) plutôt que sur n en statistiques inférentielles, c'est-à-dire pour estimer la variance de la population à partir de celle d'un échantillon. Vous trouverez ci-dessous la démonstration mathématique aboutissant à la conclusion que le calcul de la variance sur (n – 1) permet d'avoir un estimateur non biaisé, c'est-à-dire permettant de mieux estimer la variance de la population.

Selon le théorème de König-Huygens (voir la démonstration dans l'autre pdf), on peut affirmer que la formule de la variance peut s'écrire ainsi :

$$s_n^2 = \left(\frac{1}{n}\sum_{i=1}^{n} x_i^2\right) - \overline{x}^2$$

En statistiques inférentielles, nous allons chercher à estimer la variance de la population, autrement dit, nous allons chercher à déterminer l'espérance de la variance de notre échantillon, qui s'écrit $\mathbb{E}$. Pour calculer $\mathbb{E}(s_n^2)$, on va chercher à déterminer :

$$\mathbb{E}\left(\left(\frac{1}{n}\sum_{i=1}^{n} x_i^2\right) - (\overline{x}^2)\right)$$

$\mathbb{E}$ étant une fonction linéaire, pour toutes variables X1 et X2, $\mathbb{E}(X1 + X2) = \mathbb{E}(X1) + \mathbb{E}(X2)$. On a donc :

$$\mathbb{E}(s_n^2) = \mathbb{E}\left(\frac{1}{n}\sum_{i=1}^{n} x_i^2\right) - \mathbb{E}(\overline{x}^2)$$

L'étape suivante va consister à traduire autrement chacun des deux termes de l'équation. Nous allons commencer par le premier terme : $\mathbb{E}\left(\frac{1}{n}\sum_{i=1}^{n} x_i^2\right)$

Etant donné que l'espérance est une fonction linéaire, pour toute variable aléatoire X et pour toute constante a, on a $\mathbb{E}(aX) = a\mathbb{E}(X)$. Nous pouvons donc passer la constante $\frac{1}{n}$ à gauche de $\mathbb{E}$ et dire que :

$$\mathbb{E}\left(\frac{1}{n}\sum_{i=1}^{n} x_i^2\right) = \frac{1}{n}\mathbb{E}\left(\sum_{i=1}^{n} x_i^2\right)$$

Comme dit plus haut, $\mathbb{E}$ est linéaire, donc pour toutes variables X1 et X2, $\mathbb{E}(X1 + X2) = \mathbb{E}(X1) + \mathbb{E}(X2)$. On peut donc passer la somme à gauche de $\mathbb{E}$, puisque faire $\mathbb{E}(\sum_{i=1}^{n} x_i^2)$ revient à faire $\mathbb{E}(x_1^2 + x_2^2 + \cdots + x_n^2)$. On a donc : $\frac{1}{n}\mathbb{E}(\sum_{i=1}^{n} x_i^2) = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}(x_i^2)$

Ici, on va supposer que toutes les observations x_i ont la même espérance, autrement dit que toutes les estimations de la variance sont équivalentes et correspondent à l'espérance de la variable aléatoire X. On peut donc dire que $\mathbb{E}(x_i^2) = \mathbb{E}(X^2)$. Sur la base de cette hypothèse, on peut simplifier le symbole de la somme en disant qu'on additionne n fois $\mathbb{E}(X^2)$, étant donné que $\mathbb{E}(X^2)$ a toujours la même valeur, ce qui donne :

$$\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}(X^2) = \frac{1}{n}n\mathbb{E}(X^2)$$

On peut simplifier cette dernière équation en $\mathbb{E}(X^2)$ étant donné que $\frac{1}{n}n$ = 1.

Enfin, on va se servir du théorème de König-Huygens en probabilités, qui nous dit que :

$$\mathrm{Var}(X) = E[X^2] - E[X]^2$$

Sur la base de ce théorème, on peut passer $\mathbb{E}[X]^2$ du côté de Var(X), ce qui donne :

$$\mathbb{E}(X^2) = \mathbb{E}(X)^2 + V(X)$$

On a donc au final :

$$\mathbb{E}\left(\frac{1}{n}\sum_{i=1}^{n} x_i^2\right) = \mathbb{E}(X)^2 + V(X)$$

Passons maintenant au second terme de notre équation de l'espérance de la variance : $\mathbb{E}(\overline{x}^2)$

Si nous reprenons le théorème de König-Huygens, nous savons que :

$$\mathbb{E}(\overline{x}^2) = \mathbb{E}(\overline{x})^2 + V(\overline{x})$$

On admet que la moyenne $\bar{x}$ d'un échantillon est une variable aléatoire d'espérance $\mathbb{E}(\bar{x}) = \mathbb{E}(X)$ et de variance $V(\bar{x}) = \frac{1}{n}V(X)$.

On a donc : $\mathbb{E}(\overline{x}^2) = \mathbb{E}(\overline{x})^2 + V(\overline{x}) = \mathbb{E}(X)^2 + \frac{1}{n}V(X)$

Il nous suffit maintenant de reprendre notre équation de départ de l'espérance de la variance, et de remplacer les deux termes par ceux que nous avons démontrés :

$$\mathbb{E}(s_n^2) = \mathbb{E}\left(\frac{1}{n}\sum_{i=1}^{n} x_i^2\right) - \mathbb{E}(\overline{x}^2) \ = \mathbb{E}(X)^2 + V(X) - \mathbb{E}(X)^2 - \frac{1}{n}V(X)$$

En simplifiant, on obtient $V(X) - \frac{1}{n}V(X)$, qu'on peut transformer en $\frac{n}{n}V(X) - \frac{1}{n}V(X)$ et simplifier enfin en $\frac{n-1}{n}V(X)$. On a donc :

$$\mathbb{E}(s_n^2) = \ \frac{n-1}{n}V(X)$$

Cette forme finale de l'espérance de la variance nous permet de voir que la variance de l'échantillon tourne autour de $\frac{n-1}{n}V(X)$, et non autour de $V(X)$, qui correspond à la variance de la population que l'on cherche à déterminer. Par conséquent, pour corriger ce biais, nous devons calculer la variance de nos échantillons sur (n – 1), ce qui nous permet d'avoir une estimation de la variance non biaisée.