Sur la variance empirique

J'ai peut-être dit une **grosse bêtise** lors du dernier TD. Ce document est là pour, j'espère, clarifier les choses.

Définition 1. Soit X une variable aléatoire, n un entier naturel non nul et X_1, X_2, ..., X_n, n variables aléatoires indépendantes ayant la même loi que X.

La moyenne empirique de l'échantillon $(X_1, X_2, \ldots, X_n)$ est le réel

$$\overline{X}_n \coloneqq \frac{1}{n}\sum_{i=1}^{n} X_i.$$

La variance empirique biaisée de l'échantillon $(X_1, X_2, \ldots, X_n)$ est le réel positif

$$\widehat{s}_n^2 \coloneqq \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \overline{X}_n\right)^2.$$

Si $n \geq 2$, la variance empirique non biaisée de cet échantillon est le réel positif

$$\widehat{\sigma}_n^2 \coloneqq \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X}_n\right)^2.$$

Pour passer de l'une des variances empiriques à l'autre, il suffit bien sûr d'écrire

$$\widehat{\sigma}_n^2 = \frac{n}{n-1}\widehat{s}_n^2$$

et

$$\widehat{s}_n^2 = \frac{n-1}{n}\widehat{\sigma}_n^2.$$

Proposition 2. *Avec les mêmes notations que dans la définition précédente, on a*

$$\widehat{s}_n^2 = \left(\frac{1}{n}\sum_{i=1}^{n} X_i^2\right) - \overline{X}_n^2. \tag{1}$$

Démonstration. On a, pour tout i entre 1 et n,

$$\left(X_i - \overline{X}_n\right)^2 = X_i^2 - 2X_i\overline{X}_n + \overline{X}_n^2.$$

En sommant pour tous les i entre 1 et n, on obtient

$$\sum_{i=1}^{n}\left(X_i - \overline{X}_n\right)^2 = \sum_{i=1}^{n}\left(X_i^2 - 2X_i\overline{X}_n + \overline{X}_n^2\right) = \left(\sum_{i=1}^{n} X_i^2\right) - 2\overline{X}_n\left(\sum_{i=1}^{n} X_i\right) + n\overline{X}_n^2.$$

On remarque que

$$\sum_{i=1}^{n} X_i = n\overline{X}_n$$

donc on a

$$\sum_{i=1}^{n}\left(X_i - \overline{X}_n\right)^2 = \left(\sum_{i=1}^{n} X_i^2\right) - 2n\overline{X}_n^2 + n\overline{X}_n^2 = \left(\sum_{i=1}^{n} X_i^2\right) - n\overline{X}_n^2.$$

En en déduit le résultat, en divisant par n. □

Pour la version non biaisée, malheureusement, on a seulement

$$\widehat{\sigma}_n^2 = \frac{1}{n-1}\left(\sum_{i=1}^n X_i^2\right) - \frac{n}{n-1}\overline{X}_n^2.$$

La **grosse bêtise** (j'espère que je n'ai pas écrit ça au tableau...) est qu'on n'a surtout pas

$$\text{FAUX, FAUX, FAUX !}\ \widehat{s}_n^2 = \frac{1}{n}\left[\left(\sum_{i=1}^n X_i^2\right) - \overline{X}_n^2\right]\ \text{FAUX, FAUX, FAUX !}$$

Par contre, on a

$$\widehat{s}_n^2 = \frac{1}{n}\sum_{i=1}^n \left(X_i^2 - \overline{X}_n^2\right),$$

et aussi

$$\widehat{\sigma}_n^2 = \frac{1}{n-1}\sum_{i=1}^n \left(X_i^2 - \overline{X}_n^2\right),$$

parce que

$$\sum_{i=1}^n \overline{X}_n^2 = n\overline{X}_n^2.$$

Tant qu'on y est, calculons les biais. On suppose maintenant que

$$\mathbb{E}\left[X^2\right] < \infty.$$

On va d'abord calculer $\mathbb{E}\left[\overline{X}_n^2\right]$. On commence par faire le développement suivant :

$$\left(\sum_{i=1}^n X_i\right)^2 = \sum_{i=1}^n \sum_{j=1}^n X_i X_j.$$

Il y a n^2 termes dans cette somme et parmi ces n^2 termes, il y en a n qui sont tels que $i = j$. On a la décomposition suivante

$$\left(\sum_{i=1}^n X_i\right)^2 = \sum_{i=1}^n X_i^2 + \sum_{\substack{j=1 \\ j\neq i}}^n X_i X_j. \tag{2}$$

Avant de passer à l'espérance, on souligne le fait très important que, pour tout i entre 1 et n,

$$\mathbb{E}\left[X_i^2\right] = \mathbb{E}\left[X^2\right],$$

car les X_i ont la même loi que X. Et pour tous i et j entre 1 et n *avec* $i \neq j$, par indépendance

$$\mathbb{E}\left[X_i X_j\right] = \mathbb{E}\left[X_i\right]\mathbb{E}\left[X_j\right] = \mathbb{E}\left[X\right]\mathbb{E}\left[X\right] = \mathbb{E}\left[X\right]^2.$$

Maintenant, on applique l'espérance au développement (2). Par linéarité de l'espérance et le fait que les X_i ont tous la même loi que X et sont indépendants,

$$\mathbb{E}\left[\left(\sum_{i=1}^n X_i\right)^2\right] = n\mathbb{E}\left[X^2\right] + (n^2-n)\mathbb{E}\left[X\right]^2.$$

On en déduit, en divisant par n^2,

$$\mathbb{E}\left[\overline{X}_n^2\right] = \frac{1}{n}\mathbb{E}\left[X^2\right] + \mathbb{E}\left[X\right]^2 - \frac{1}{n}\mathbb{E}\left[X\right]^2 = \mathbb{E}\left[X\right]^2 + \frac{1}{n}\,\mathbb{V}\text{ar}\left[X\right].$$

Finalement, en utilisant la formule (1),

$$\mathbb{E}\left[\widehat{s}_n^2\right] = \frac{1}{n} \times n\mathbb{E}\left[X^2\right] - \mathbb{E}\left[X\right]^2 - \frac{1}{n}\,\mathbb{V}\mathrm{ar}\left[X\right] = \mathbb{V}\mathrm{ar}\left[X\right] - \frac{1}{n}\,\mathbb{V}\mathrm{ar}\left[X\right].$$

Donc le biais de $\widehat{s}_n^2$ est

$$b\left(\widehat{s}_n^2\right) = \mathbb{E}\left[sn^2\right] - \mathbb{V}\mathrm{ar}\left[X\right] = -\frac{1}{n}\,\mathbb{V}\mathrm{ar}\left[X\right].$$

Ce biais est strictement négatif (sauf si X est presque sûrement égale à une constante). Donc l'estimateur $\widehat{s}_n^2$ de la variance est biaisé.

En revanche,

$$\mathbb{E}\left[\widehat{\sigma}_n^2\right] = \mathbb{E}\left[\frac{n}{n-1}\widehat{s}_n^2\right] = \frac{n}{n-1}\,\mathbb{V}\mathrm{ar}\left[X\right] - \frac{1}{n-1}\,\mathbb{V}\mathrm{ar}\left[X\right] = \mathbb{V}\mathrm{ar}\left[X\right],$$

donc $\widehat{\sigma}_n^2$ est non biaisé ($b(\widehat{\sigma}_n^2) = 0$).