Quelle est la « bonne » formule de l'écart-type ?

Emmanuel Grenier

Reims Management School
emmanuel.grenier@reims-ms.fr

Relu par Jacques Goupy et Henry P. Aubert

Il suffit de consulter les normes ou un bon manuel de statistique pour avoir la réponse. Alors pourquoi cette notule ? C'est que la réponse diffère d'un auteur à l'autre. Examinons ces formules si familières qu'on n'y prête plus guère attention.

1. Ecart-type s et écart-type σ

1.1. L'écart-type s des valeurs prises par une variable

On considère un ensemble de valeurs prises par une grandeur numérique. L'écart-type est une mesure de la dispersion des valeurs autour de leur moyenne arithmétique.

Prenons par exemple les tailles suivantes relevées sur 7 personnes :

152 158 164 168 168 169 176

Calculons la moyenne arithmétique des tailles, $\bar{x} = \frac{1}{n}\sum_i x_i$, avec ici $n = 7$:

$$\bar{x} = \frac{1}{7}\left[152 + 158 + 164 + 2\times 168 + 169 + 176\right] = 165{,}0$$

Par définition, l'écart-type est la moyenne quadratique des écarts à la moyenne $\bar{x}$. On le note habituellement *s* (de l'anglais ***s****tandard deviation*) :

$$\sqrt{\frac{1}{n}\sum_i (x_i - \bar{x})^2} \qquad \{1\}$$

Soit, pour l'exemple,

$$\sqrt{\frac{1}{7}\left[(152-165)^2 + (158-165)^2 + (164-165)^2 + 2\times(168-165)^2 + (169-165)^2 + (176-165)^2\right]}$$

$$= 7{,}3$$

Le carré de l'écart-type, s^2, est appelé la variance. La variance est par conséquent la moyenne arithmétique des carrés des écarts à la moyenne $\bar{x}$.

1.2. L'écart-type σ des valeurs possibles d'une variable aléatoire

On peut également calculer l'écart-type sur les valeurs possibles d'une variable aléatoire numérique.

Prenons par exemple le résultat d'un lancer de dé. Les valeurs possibles sont les entiers de 1 à 6, chacune ayant une probabilité de réalisation égale à 1/6.

La moyenne des valeurs possibles est $\mu = \frac{1}{6}\times 1 + \frac{1}{6}\times 2 + \cdots + \frac{1}{6}\times 6 = 3{,}5$

L'écart-type est $\sigma = \sqrt{\frac{1}{6}\times(1-3{,}5)^2 + \frac{1}{6}\times(2-3{,}5)^2 + \cdots + \frac{1}{6}\times(6-3{,}5)^2} = 1{,}71$

1.3. Cas où σ = s : l'écart-type d'une population

Si on choisit un individu de manière aléatoire dans une population et que l'on relève une valeur numérique sur cet individu, les valeurs possibles sont les valeurs présentes dans la population (et les probabilités associées sont les fréquences dans la population). De ce fait, la moyenne μ et l'écart-type σ des valeurs possibles sont égales à la moyenne $\bar{x}$ et à l'écart-type *s* des valeurs prises par les individus de la population.

2. Estimation de σ par l'écart-type s d'un échantillon : le problème du biais d'estimation

On dispose d'un échantillon constitué par des réalisations d'une variable aléatoire.

L'écart-type *s* des valeurs de l'échantillon donne une estimation de l'écart-type σ des valeurs possibles de la variable. L'écart-type de l'échantillon peut prendre diverses valeurs *s*, qui tantôt sous-estiment, tantôt surestiment σ. On pourrait penser que ces valeurs sont centrées sur σ. Ce n'est pas le cas : il existe un écart entre la moyenne des valeurs possibles *s* de l'écart-type de l'échantillon et la valeur σ à estimer.

Ce phénomène de biais apparaît également lorsqu'on estime la variance σ^2 de la variable par la variance s^2 de l'échantillon. Le biais est plus simple à exprimer dans le cas de la variance parce qu'il ne dépend que de la taille de l'échantillon, *n*, et de σ^2. En effet, on montre (voir par exemple la référence [3]) que la moyenne des valeurs possibles s^2 de la variance de l'échantillon est égale à

$$\frac{n-1}{n}\sigma^2$$

Ceci se vérifie par simulation (voir [2]) :

Reprenons l'exemple du lancer de dé. La variance des valeurs possibles est égale au carré de l'écart-type : $\sigma^2 = 1{,}71^2 = 2{,}92$.

Produisons un échantillon, de petite taille pour que le biais soit appréciable, par exemple de taille $n = 5$. On peut lancer 5 dés mais, pour la suite, il vaut mieux simuler l'expérience sur ordinateur (avec Excel, il suffit de recopier dans 5 cellules la formule =ALEA.ENTRE.BORNES(1;6)). Admettons qu'on ait obtenu les valeurs suivantes :

3 4 5 2 5

La variance de l'échantillon est le carré de l'écart-type *s* calculé par la formule {1} (avec Excel la fonction VAR.P, carré de la fonction ECARTYPEP) : $s^2 = 1{,}36$

Ici la variance de l'échantillon sous-estime la variance $\sigma^2 = 2{,}92$. Produisons un deuxième échantillon :

2 1 5 5 1

$s^2 = 3{,}36$; on surestime σ^2.

Répétons cette opération un très grand nombre de fois (avec Excel, il suffit de recopier les cellules donnant les valeurs d'un échantillon et de sa variance) et calculons la moyenne des variances des échantillons. Nous observons alors un décalage par rapport à σ^2 : la moyenne

des variances des échantillons est proche de $\sigma^2 (n-1)/n$, pour l'exemple proche de $2{,}92 \times 4/5 = 2{,}19$ et non de $\sigma^2 = 2{,}92$.

La moyenne des valeurs possibles de la variance étant égale à σ^2 au facteur $(n-1)/n$ près, on élimine le biais en multipliant la variance de l'échantillon par l'inverse de ce facteur, c'est-à-dire par $n/(n-1)$. On obtient ainsi la « variance en n-1 », somme des carrés des écarts à la moyenne divisée, non par *n* comme dans le cas de la variance s^2, mais par $n - 1$:

$$s_{n-1}^2 = \frac{1}{n-1}\sum (x_i - \bar{x})^2 \qquad \{2\}$$

Remplaçons la variance s^2 de nos échantillons par la variance en n-1 (fonction VAR à la place de la fonction VAR.P). Nous observons que la moyenne est maintenant proche de σ^2 = 2,92.

Notons que le biais n'est pas nul quand on estime σ par l'écart-type en n-1. Il est cependant plus faible en général qu'avec l'écart-type *s*.

3. La racine carrée du carré moyen, ou « écart-type corrigé »

3.1. Définition

On appelle carré moyen la variance de l'échantillon (ou une composante de cette variance comme, par exemple, la variance résiduelle de l'analyse de la variance), corrigée de manière à obtenir une estimation non biaisée de la variance d'une variable aléatoire. La variance en n-1, s_{n-1}^2, définie au paragraphe précédent (formule {2}) est le carré moyen associé à la variance de l'échantillon s^2 dans le cas où on estime la variance de la variable aléatoire qui a produit l'échantillon (ou variance de la population).

3.2. Avantage et inconvénient de l'usage du carré moyen

Formules plus agréables…

Prenons par exemple l'intervalle de confiance de la moyenne. La demie amplitude de l'intervalle est égale à $1{,}96 \times \sigma/\sqrt{n}$ (pour un niveau confiance de 95%).

Dans le cas où σ est inconnu, la demie amplitude est égale à $t \times s/\sqrt{n-1}$, où *s* est l'écart-type de l'échantillon et où *t* est le fractile d'ordre 0,975 de la loi de Student à $n - 1$ degrés de liberté.

Remplaçons dans la formule l'écart-type *s* par l'écart-type en n-1, s_{n-1}. La demie amplitude s'écrit $t \times s_{n-1}/\sqrt{n}$

On retrouve l'expression utilisée dans le cas où σ est connu : le fractile 1,96 de la loi de Gauss est remplacé par le fractile *t* de la loi de Student et l'écart-type σ est remplacé par la racine carrée du carré moyen, s_{n-1}.

mais risque de confusion

- Non biaisé ne veut pas dire précis

Revenons aux échantillons simulés au § 2. Sur chacun des échantillons, calculons l'erreur d'estimation, c'est-à-dire la différence entre la variance de l'échantillon et σ^2. Par exemple, pour le premier échantillon, l'erreur d'estimation est égale à $s^2 - \sigma^2$ = 1,36 – 2,92 = -1,56. Calculons la moyenne des erreurs, les erreurs étant prises en valeurs absolues ou mises au carré. Remplaçons maintenant sur chaque échantillon la variance s^2 par le carré moyen s_{n-1}^2. L'erreur (absolue ou quadratique) moyenne est plus importante lorsqu'on utilise le carré moyen. Le carré moyen apparaît également moins précis lorsqu'on compte la proportion des échantillons où l'erreur dépasse une limite fixée.

- De quoi parle-t-on ?

Un carré moyen est souvent appelé « variance » et sa racine carrée « écart-type ». Par exemple, les normes AFNOR [1] appellent variance et écart-type « d'échantillon » la variance et l'écart-type en n-1.

4. Conclusion

L'écart-type devrait toujours être défini comme la moyenne quadratique des écarts à la moyenne {1}, aussi bien sur un échantillon que sur une variable aléatoire ou une population. On ne peut appeler « écart-type » la racine carrée d'un carré moyen sans que ceci n'introduise des confusions, même si l'objectif est de simplifier l'expression de calculs.

5. Références

[1] AFNOR - Statistiques – Vocabulaire et symboles – Partie 1 : Probabilité et termes statistiques généraux. ISO TC 69/SC 1 N26, août 2002.

[2] Morineau A., Chatelin Y.-M. (coordinateurs) - L'analyse statistique des données. Apprendre, comprendre et réaliser avec Excel. Ellipses, 2005. 407 pages.

[3] Saporta G. - Probabilités, analyse des données et statistique. 2e édition. Editions Technip, 2006. 656 pages.