Piège de la variance.

On a pu lire :
J'ouvre ce fil pour simplement vous faire part d'une de mes "découvertes" par rapport à la calculatrice XX, ou en tout cas la calculatrice xxxxxxx pour moi.
Quand on utilise la commande "variance" avec une série qu'on a entrée, la valeur que celle-ci retournée n'est pas celle qu'on calcule à la main.

Ceci est naturellement traumatisant pour un professeur de mathématique qui enseigne les probabilités. De quoi s'agit-il ?

On a une liste de notes. On peut en calculer la moyenne. On peut calculer la somme des carrés des écarts à cette moyenne, diviser par le nombre de notes, puis en prendre la racine carrée. On obtient ainsi la moyenne quadratique des écarts. De la même façon, on pourrait calculer la moyenne arithmétique de ces écarts (moyenne des valeurs absolues) ou la moyenne géométrique.

Parallèlement, on connaît la formule de König-Huygens

Théorème — Pour toute variable aléatoire réelle X, on a :

$$\operatorname{Var}(X) \equiv E\left[(X - E[X])^2\right] = E[X^2] - E[X]^2$$

La fonction Var(X) désigne la variance, E[…] désigne l'espérance.
La variance est définie comme le carré de l'écart moyen quadratique. Ce terme est très utilisé en probabilité, mais il est indispensable de rappeler que cette notion n'est que théorique, puisqu'il s'agit d'une valeur réelle (au sens de réalité) qu'on utilise au carré.

L'espérance est plus difficile à définir. On peut lire la définition suivante :
En théorie des probabilités, l'**espérance mathématique** d'une variable aléatoire réelle est, intuitivement, la valeur que l'on s'attend à trouver, en moyenne, si l'on répète un grand nombre de fois la même expérience aléatoire. Elle se note $\mathbb{E}(X)$ et se lit « espérance de X ». (Wikipédia)
Elle paraît suffisamment claire : on y trouve les termes "variable aléatoire réelle", "intuitivement", "un grand nombre de fois".

Toues ces hypothèses font partie intrinsèquement du théorème de König-Huygens.
Si on remplace E[…] par M[…], M étant la moyenne arithmétique, alors ce théorème est mathématiquement vrai, quant à l'égalité des deux membres. Par contre, l'identité avec Var(X) n'est vrai que dans les hypothèses strictes de la définition de l'espérance mathématique.

Dans la plupart des cas, on ne connaît pas la valeur de l'espérance, on parle généralement de "valeur vraie". Alors, l'expression de la variance sera :
$\mathrm{Var}(x) = 1/(n-1)\ \Sigma(x_i-\mu)^2$
μ étant la moyenne arithmétique des x_i et n leur nombre.

La calculette utilisée par le professeur traumatisé respecte cela.

Et l'écart type ?

Autre exemple de question posée sur un forum.
Comment procéderiez vous pour répondre à cette question:
"115 enfants ont bénéficié de séances de soutien en lecture. les scores de lecture obtenus par les 115 enfants après le soutien avaient pour moyenne 7,2 avec un écart type de 1,7. Sachant que le score moyen au même test des élèves de 6° sans difficulté est de 7,5, peut on dire à l'issue du soutien, que les élèves ont toujours des scores significativement en dessous de la normale"

Nous avons 3 informations :

1- le score des élèves de 6è sans difficulté pour un certain test : 7.5
2- le score de 115 élèves de 6è après avoir bénéficié de soutien en lecture : 7.2
3- l'écart type de cette moyenne est 1.7.

D'abord d'un point de vue strictement mathématique.

L'expression "écart type" sous-entend forcément que la répartition des notes obtenues est celle de la loi normale, autrement dit que cette répartition a la représentation graphique de la courbe de Gauss. Etant donné le nombre d'élèves concernés (115), on peut effectivement espérer que cette sélection a un caractère aléatoire. Admettons cette hypothèse.

La loi normale est basée sur la moyenne arithmétique des résultats.
La répartition de la loi normale est telle que 67% des résultats sont compris dans la tranche $[\mu-\sigma ; \mu+\sigma]$ et 95% dans la tranche $[\mu-2\sigma ; \mu+2\sigma]$. Si on appelle "écart probable" ep = 2/3 σ, alors 50% des résultats sont compris dans la tranche $[\mu-ep ; \mu+ep]$. Une simple lecture d'une table de répartition confirme cela.

Exploitons les résultats :
Ecart type = 1.7 ; ep = écart probable = 1.1
Pour les élèves après les séances de soutien en lecture,
50% des élèves ont un résultat compris entre 6.1 et 8.3
67% des élèves ont un résultat compris entre 5.5 et 8.9
On peut admettre que l'écart type pour l'ensemble des élèves de 6è sur ce test de lecture est aussi 1.7, alors pour l'ensemble des élèves de 6è
50% des élèves ont un résultat compris entre 6.4 et 8.6
67% des élèves ont un résultat compris entre 5.8 et 9.2
On constate que les écarts entre ces deux tranches n'est pas significatif, ceci répond donc à la question posée : "Non, à l'issue du soutien, les élèves n'ont plus des scores significativement en-dessous de la normale".
Par des jeux calculatoires on peut même donner une estimation chiffrée du risque de cette réponse.

Critique de la méthode.

On a fait plusieurs hypothèses :

1- Les 115 élèves ayant profité du soutien ET qui on subi le test de contrôle, résulte d'un choix aléatoire. Si cette hypothèse n'est pas admise cela voudrait dire que le choix des élèves concernés par le soutien de lecture a été fait avec une arrière pensée.

2- La note reçue par chaque élève correspond à une expérience aléatoire. Ce point mérite d'être détaillé. Aléatoire signifie ne dépend que du hasard. On peut dire aussi, que le résultat dépend d'un très grand nombre de données, toutes aussi inconnues les unes que les autres, comme le tir sur cible. Peut-on dire cela d'une note obtenue par élève, personnellement, je ne le pense pas. Les cause de l'attribution d'une note sont assez bien connues et ne suivent pas les mêmes lois pour chacun des élèves. Donc, si il ne s'agit pas de la même loi, le résultat ne correspond pas à la loi normale.
3- Les progrès en lecture dépendent uniquement des séances de soutien. Rien dans l'énoncé de la question n'aborde ce point. On constate simplement que ces élèves lisent à peu près comme la moyenne des élèves de 6è. Que s'est-il passé pendant ces séances, une plus grande motivation, une amélioration de l'acuité visuelle, un passage d'une période difficile, n'ayant rien à voir avec la lecture, ou tout simplement l'apprentissage des tests. Le but de la question est de montrer que les enfants avaient du mal à lire avant les séances et que ces séances ont aidé à l'apprentissage de la lecture. La seule conclusion qu'on puisse en tirer est que le jour du test, les notes de ces 115 élèves correspondent à la moyenne.
4- L'écart type de la répartition "normale" des élèves de 6è est 1.7. Cette hypothèse est sous-entendue. La question n'évoque pas cette valeur. Pourtant, c'est fondamental pour pouvoir comparer 2 scores. En matière de probabilité, le terme "normal" a une signification précise. La valeur 7.5 n'est pas une valeur "normale" mais une valeur "moyenne".
5- Les élèves ont progressé. Cela est aussi sous-entendu dans l'énoncé. Comment pourrait-on le savoir, puisqu'il n'y a aucune information sur un test éventuel réalisé avant les séances de soutien.

En conclusion, il semble que l'auteur de cette question n'est pas un étudiant et que cette question n'est pas l'énoncé d'un exercice. Elle est donc posée dans un but précis et certainement pas dans un contexte de théorie mathématique. Chacun sait qu'on peut faire ce qu'on veut avec des chiffres, mais ce n'est pas toujours souhaitable, surtout lorsqu'il s'agit de méthodes éducatives.