Un étudiant a posté le sujet suivant sur un forum d’aide aux mathématiques, niveau post-bac. Il n’y a pas d’indication concernant la source de cet exercice.

<u>ENONCE</u>

Un évènement peut se produire à tout instant X (loi uniforme) dans un intervalle I = [0, b], où b est inconnu. Pour estimer la valeur de b inconnue, on va considérer un n-échantillon (X1, X2, • • • , Xn).

1. Rappeler la densité, l'espérance, et la variance de X en fonction de b.
2. Soit l'estimateur $T_1 = \overline{X} = \frac{1}{n} \sum Xi$ pour i allant de 1 à n.

. Calculer $E(T_1)$. T_1 est-il sans biais ?

3. À l'aide de T_1, construire un estimateur sans biais T_2 de b. Calculer $V(T_2)$.

ET la réponse dans le livre:

1. Si x ∈ [0, b] alors $f(x) = \frac{1}{b}$sinon f(x) = 0, $E(X) = \frac{b}{2}$, $V(X) = \frac{b\hat{A}^2}{12}$.
2. $E(T_1) = \frac{b}{2}$

Ainsi T_1 est un estimateur biaisé de b.

Ma question pourquoi il est biaisé alors que son espérance est égale à l'espérance de X?

3. On peut donc prendre $T_2 = 2T_1 = 2\overline{X}$ et $V(T_2) = 4\,V(T_1) = \frac{b\hat{A}^2}{3n}$

Plusieurs points méritent discussion :
1- la signification de « biaisé »
2- l’utilisation de la loi de probabilité dite uniforme.

Ce terme « biais » est actuellement très utilisé. On peut se poser la question de savoir s’il a une signification précise, surtout dans son utilisation en mathématiques ou si au contraire, c’est un mot passe-partout, bien utile quand on ne veut pas rentrer dans les détails.
Pratiquement toutes les utilisations convergent vers le sens « pas bon ». Mais en mathématiques on aime généralement être précis, donc cette signification n’est pas suffisante.
On peut observer que ce terme est particulièrement souvent observé à l’occasion du calcul de l’écart-type. C’est un cas extrême, puisqu’en l’occurrence, cela signifie « calcul faux parce qu’on a utilisé une formule fausse ». En d’autres termes il ne s’agit pas d’une « erreur » qui est notion précise, mais d’une faute.

Dans le présent exercice, ce terme est donc à considérer comme synonyme de « erreur ».
Il y a deux types d’erreurs, les erreurs systématiques et les erreurs accidentelles. On appelle erreur systématique une erreur due à une cause constante et permanente, par exemple un défaut d’étalonnage. On appelle erreur accidentelle une erreur due à un défaut de précision, contrairement à l’erreur systématique, cette erreur est aléatoire.
On peut corriger les erreurs systématiques par des modes opératoires particuliers, on peut réduire les erreurs accidentelles, en tout cas leur influence sur le résultat définitif en multipliant le nombre de mesures.

Le but de l’exercice est de déterminer la valeur de ‘b’.

L'énoncé appelle T_1 la moyenne arithmétique des valeurs Xi observées. Ceci est justifié par le postulat de la moyenne. T_1 est appelé « estimateur » et l'énoncé propose d'exprimer $E(T_1) = b/2$ appelée « espérance ».
On a donc d'une part T_1 qui est égal à la moyenne arithmétique de n observations. Cette valeur est « la plus probable compte tenu des informations disponibles ». D'autre part on écrit que l'espérance de T_1 est égale à b/2 qui est une valeur inconnue (rappelons que le but est d'évaluer 'b').
La question posée est « T_1 est-il sans biais ? ». Si « sans biais » signifie « exact » alors évidemment T_1 est biaisé. Si, comme on le lit assez souvent, le biais correspond à une erreur systématique, alors T_1 n'est pas biaisé, comme le comprend l'étudiant qui pose la question.
Il reste une dernière possibilité que je note pour mémoire : $E(T_1)$ a une définition mathématiquement précise mais sa valeur est parfaitement inconnue. Comme l'espérance de T_1 n'est pas égale à l'estimateur T_1, alors on dit que T_1 est biaisé.
Le corrigé de la question 3) nécessiterait une explication : si T_1 est biaisé, comment $T_2 = 2T_1$ peut être non-biaisé ?

Naturellement, je n'ai pas la réponse, et je constate, une fois de plus, que ce terme « biais » a le sens qu'on veut bien lui donner.

Étant donné cet énoncé, je ne voudrais pas manquer l'occasion d'évoquer la notion de « loi de probabilité ».
Dire qu'une expérience suit une loi uniforme signifie que tous les éléments ont la même probabilité de sortir. Par exemple au tirage d'un dé ordinaire équilibré, chaque face a la même probabilité de sortir, quel que soit le nombre de taches qui servent à la repérer. Si on compte le nombre de sorties 'n' de chaque face, on constate que la moyenne arithmétique de ce nombre de sorties est proche de n=N/6. Mais on observe aussi que ces 6 nombres de sorties ont une répartition normale, conformément au TCL.
En d'autres termes, une expérience suivant la loi uniforme a toujours une répartition normale et on peut ajouter qu'il est difficile d'obtenir une répartition uniforme. La seule méthode que je connaisse est de diviser l'espace d'étude, appelé souvent « univers », en espaces de même taille et de réaliser l'expérience successivement dans chacun de ces espaces.

Pour confirmer que les nombreuses ambiguïtés et imprécisions sur le sens des termes posent de réels problèmes aux étudiants et aux enseignants, il y a eu une autre question concernant l'expression « intervalle de fluctuation ». Je ne m'étendrai pas sur ce point, mais dans le cours de la discussion, il a été question de « biais » et l'on peut lire [G.] :
« La notation qui devient classique pour une loi Normale est de donner sa moyenne et sa variance : $\mathcal{N}(m,s2)$, surtout lorsqu'il s'agit d'une approximation, car on a un estimateur simple, sans biais et convergent de la variance, pas de l'écart type. », comme réponse à la question :
« Toujours à propos de statistiques quelle est la notation usuelle de la loi normale de moyenne m et d'écart-type s, $\mathcal{N}(m,s)$ ou $\mathcal{N}(m,s2)$? Personnellement je pencherais pour la première, qui est celle que j'ai rencontré à peu près dans tous les cours de proba-stats que j'ai pu parcourir et qui n'est étonnement pas celle du programme officiel de TS. »
Le professeur donne le lien sur le B,O, :
http://cache.media.education.gouv.fr/file/special_8_men/98/4/mathematiques_S_195984.pdf

Je tiens d'abord à préciser la signification de « Loi Normale ». C'est une loi de la nature, au même titre que la loi de gravitation universelle, par exemple. Ceci a été formalisé sous le nom de

Théorème Central Limite. La notation que l’on trouve met en évidence la translation et l’affinité, en d’autres termes la « mise à l’échelle » qui permet de se ramener à la formule de Laplace bien connue. Naturellement le gros avantage de préciser, comme second paramètre, la variance plutôt que l’écart type, évite le doute sur la signification de ce paramètre.

Revenons sur la réponse de [G.] : « [,,,] surtout lorsqu'il s'agit d'une approximation, car on a un estimateur simple, sans biais et convergent de la variance, pas de l'écart type. »
Mathématiquement, la définition de l’écart-type est parfaitement claire : c’est l’écart moyen quadratique, appelé aussi « moyenne de second ordre ». La discussion sur le dénominateur est hors-sujet. Pour des motifs de normalisation internationale, au lieu de parler d’écart moyen quadratique, on parle généralement d’écart-type, ce qui sous-entend que cet écart est type de quelque-chose, c’est à dire qu’on est dans le contexte de la loi normale.
La variance est le carré de l’écart-type. En d’autres termes, la seule différence entre la variance et l’écart-type, c’est que l’un est le carré de l’autre, c’est à dire que ces deux notions représentent la même chose, à part leur valeur numérique. Il y a lieu de préciser que la valeur numérique de l’écart-type est dans la même unité que la moyenne, ce qui n’est pas le cas de la variance. Exemple : on mesure une distance N fois, la moyenne donnera 1235.25 et un écart type de 1.00. Là l’unité employée est le mètre. La variance est donc 1.00. Si on employait comme unité le Km, alors la moyenne serait 1.23525, l’écart-type 0.001 et la variance 10^{-6}.
Je tiens aussi à ajouter que le choix de l’écart moyen quadratique plutôt que l’écart moyen arithmétique ou que l’écart probable est simplement conventionnel et qu’il s’agit seulement d’une unité de calcul. Ces trois types d’écarts sont parfaitement proportionnels. La variance n’est rien d’autre que le carré de l’écart-type.
Alors ma question à [G.] : comment peut-il se faire que la valeur de la variance puisse résulter d’un estimateur simple, sans biais et convergent, et pas l’écart-type, d’autant que le calcul de la valeur numérique de la variance est une opération arithmétique simple : élévation au carré ?
Je ne sais pas comment ce pauvre professeur qui s’adresse à un forum de math renommé, pourra faire un cours crédible à ses élèves.

Rappel : il est toujours possible de réagir en me contactant.