Courrier des lecteurs

À propos de la page 517 du bulletin 447

Je voudrais faire quelques remarques sur l'article de R. Ferréol « Quelques calculs de probabilité de la vie courante » paru dans le Bulletin n° 447, page 517 à propos du problème 1 : Consultation d'un dictionnaire en 15 volumes.
Rappel du problème : Combien doit-on ouvrir de volumes en moyenne lorsqu'on doit consulter n mots dans un dictionnaire de p volumes ?

La solution donnée peut s'exprimer de la manière suivante :
Soient $m_1, m_2, \ldots, m_n$ les n mots à chercher, $v_1, \ldots, v_p$ les p volumes, X la variable aléatoire : « Nombre de volumes à ouvrir pour consulter les n mots », $X_1, \ldots, X_p$ les variables aléatoires définies par :

$$X_i = \begin{cases} 1 & \text{si au moins un des } n \text{ mots est dans le volume } v_i, \\ 0 & \text{si aucun des } n \text{ mots n'est dans le volume } v_i. \end{cases}$$

On a : $X = X_1 + X_1 + \ldots + X_p$.
Soit M_j l'événement : « le mot m_j n'est pas dans le volume v_1 », $j = 1, 2, \ldots, n$.

$$P(X_1 = 0) = P(M_1 \cap \ldots \cap M_n) = \left(1 - \frac{1}{p}\right)^n$$

en supposant l'indépendance des événements $M_1, \ldots, M_n$. On a alors :

$$E(X_1) = P(X_1 = 1) = 1 - \left(1 - \frac{1}{p}\right)^n.$$

Le même raisonnement peut être fait pour les variables $X_2, \ldots, X_p$ et on obtient :

$$E(X) = p\left(1 - \left(1 - \frac{1}{p}\right)^n\right)$$

(une erreur de frappe s'est glissée dans l'article du bulletin où il faut remplacer le premier n par p). Si on applique la formule pour $n = 2$ et $p = 2$, on obtient :

$$E(X) = 2\left(1 - \left(1 - \frac{1}{2}\right)^2\right) = \frac{3}{2}.$$

Or dans le cas particulier où il n'y a que 2 mots dans un dictionnaire en 2 volumes, comme par hypothèse les 2 volumes ont le même nombre de mots, chaque volume contiendra un seul mot et pour consulter les 2 mots il faudra obligatoirement ouvrir les 2 volumes. On devrait donc trouver :

$$E(X) = 2.$$

Où est l'erreur ?
Je pense que c'est l'hypothèse : « Tous les volumes ont le même nombre de mots » qui doit être changée. En effet cette hypothèse est contradictoire avec l'indépendance des événements $M_1, \ldots, M_n$. Dans le cas particulier précédent on a par exemple :

$$P(M_1 \cap M_2) = 0$$

et

$$P(M_1)P(M_2) = \frac{1}{2} \times \frac{1}{2} = \frac{1}{4}.$$

On ne peut donc pas écrire :

$$P(M_1 \cap \ldots \cap M_n) = P(M_1) \times \ldots P(M_n).$$

Pour que les résultats soient valables, il faut supposer que chacun des mots de la langue a été mis « au hasard » dans l'un des p volumes. Les volumes n'ont pas alors obligatoirement le même nombre de mots et il peut même théoriquement exister des volumes « vides » (la recherche des mots dans un dictionnaire construit de cette manière poserait quelques problèmes). Cependant si le nombre de mots de la langue est « grand », tous les volumes auront approximativement le même nombre de mots. On peut alors conjecturer que l'espérance calculée ci-dessus est la limite lorsque le nombre de mots tend vers l'infini de l'espérance calculée dans l'hypothèse où tous les volumes ont le même nombre de mots. Essayons de le démontrer :

Supposons que chaque volume contienne k mots. Notons $E_k(X)$ l'espérance du nombre de volumes à ouvrir pour consulter n mots. En reprenant les notations précédentes, on a :

$$P(X_1 = 0)$$
$$= \frac{\text{nombre de répartitions de } k \text{ mots dans } v_1 \text{ ne contenant pas les } n \text{ mots à chercher}}{\text{nombre total de répartitions de } k \text{ mots dans } v_1}$$

$$P(X_1 = 0) = \frac{C_{pk-n}^{k}}{C_{pk}^{k}}.$$

D'où

$$E_k(X) = p - p\frac{C_{pk-n}^{k}}{C_{pk}^{k}}.$$

On suppose que $n \leq (p-1)k$ car si $n > (p-1)k$, il faudra ouvrir tous les volumes et on aura :

$$E_k(X) = p.$$

En simplifiant on obtient :

$$E_k(X) = p - p\frac{(pk-n)!(pk-k)!}{(pk)!(pk-k-n)!} = p - p\prod_{i=1}^{k}\left(1 - \frac{n}{(p-1)k+i}\right).$$

On remarque que cette dernière expression est valable pour tout $n \leq pk$ car si n est compris entre $(p-1)k+1$ et pk, l'un des termes du produit sera nul et on aura bien :

$$E_k(X) = p.$$

$$\lim_{k \to +\infty} \frac{k}{pk - n + i} = \frac{1}{p}$$

d'où

$$\lim_{k \to +\infty} \mathrm{E}_k(\mathrm{X}) = p\left(1 - \left(1 - \frac{1}{p}\right)^n\right).$$

Pierre CARRIQUIRY