Coïncidences des dates d'anniversaires

Jean François Kentzel(*)

L'an dernier, j'ai fait le pari avec les élèves d'une classe que deux d'entre eux fêtaient leur anniversaire le même jour. Il y avait trente quatre élèves dans cette classe, mais nous avons constaté qu'il n'y avait, contrairement à mon attente, aucune telle coïncidence de date.

Une question classique

Le calcul effectif de la probabilité d'une coïncidence de dates d'anniversaires dans un groupe de n personnes peut être effectué en classe de Première. Désignant par A_n l'événement « il y a au moins une coïncidence » et par B_n l'événement contraire, on obtient facilement, en supposant notamment l'équiprobabilité des dates d'anniversaires et en laissant de côté le 29 février qui ne change pas grand chose (car la probabilité de son apparition est, en première approximation, $\frac{1}{3\times 365+366}$) :

$$P(B_n)=\frac{365!}{(365-n)!\,365^n}.$$

Pour obtenir toutes ces probabilités, par exemple jusqu'à $n = 50$, sur une feuille de tableur, il suffit d'installer une colonne compteur (1, 2, 3, …, etc.) et d'utiliser la formule

$$P(B_n)=P(B_{n-1})\times\frac{365-(n-1)}{365},$$

formule utilisable aussi avec certaines calculatrices, qui n'est pas justifiable par un calcul de probabilité en Première (les probabilités conditionnelles n'étant pas connues(1)) mais l'est algébriquement.

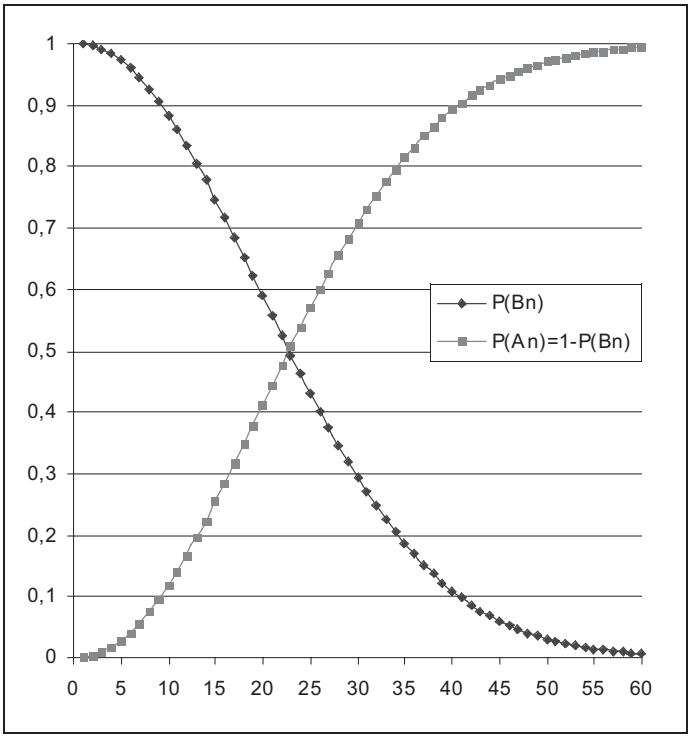
Cet exercice est vraiment intéressant car on obtient des résultats étonnants (on peut insister sur ce point en demandant aux élèves, sans insister pour obtenir une réponse ! car ce n'est pas une question sérieuse, l'idée qu'ils ont de certains résultats avant d'effectuer les calculs).

Par exemple $P(A_n)$ dépasse 0,5 quand n dépasse 22 et vaut environ 0,8 pour $n = 34$ (l'intuition « avec 34 dates, on va utiliser, en gros, un dixième du calendrier, probablement sans coïncidence » est tout à fait fausse).

L'assistant graphique donne le dessin ci-dessous.

(*) Lycée Pardailhan à Auch (32).

(1) Lorsqu'elles le sont, en désignant par M_n l'événement « la n-ième personne a une date d'anniversaire distincte des autres », on peut écrire $P(B_n) = P(B_{n-1})\ P_{B_{n-1}}(M_n)$.



Ce calcul qu'on trouve un peu partout suppose aussi que le modèle choisi pour l'expérience consistant à obtenir n dates est celui de la succession de n expériences identiques et indépendantes consistant à choisir au hasard une date parmi 365 ; en d'autres termes, on suppose qu'on opère n tirages avec remise et on construit un arbre formé de n fois 365 branches.

Cet exercice est proposé dans plusieurs livres de Première S, mais il est si intéressant qu'il peut être traité, malgré sa relative difficulté, dans d'autres séries (par exemple, dans une classe de première STT, une bonne dizaine d'élèves ont trouvé la formule de tableur qui donne le graphique ci-dessus).
Il en est question dans un des articles de la rubrique « Statistiques En Ligne » du CD ROM d'accompagnement du programme de Terminale.

Simulation

J'ai effectué en classe quelques simulations sur ce thème. Cette activité a beaucoup plu aux élèves qui avaient été intrigués par ce pari que j'avais perdu(2).
Le manuel Terracher Hachette de Première S propose, page 278, une procédure pour le faire : dans la cellule A1 on tape « = ENT (365*ALEA()+1) », qui simule une

(2) J'ai pu observer que les élèves sont également très stimulés lorsqu'on a gagné le « pari des anniversaires ». Lorsqu'on fait un pari en classe, ce à quoi les programmes actuels nous invitent, on est donc gagnant à tous les coups !

date d'anniversaire et on le recopie (33 fois dans mon cas) sur une colonne. Les éventuelles coïncidences ne sautent pas aux yeux et le livre propose pour les rendre plus visibles d'ordonner (tri croissant) ces nombres, ce qui nécessite de les « bloquer » avant le tri en cochant l'option *Sur ordre* dans le menu *Outil-Options-Calcul*. Ce blocage doit être effectué avant chaque nouvelle simulation.

La procédure qui suit est plus rapidement exécutable, ce qui peut être utile car les résultats obtenus heurtent l'intuition de la plupart des élèves et un assez grand nombre de simulations peut donc être nécessaire si on veut être convaincant. Elle consiste, après l'installation de la colonne A, à mettre 0 en B1, puis en B2 la formule =NB.SI(A$1:A1;A2) et à la recopier dans les cellules B3 à Bn. On a ainsi dans la colonne B des zéros, sauf si la date a déjà été rencontrée, auquel cas on a un 1 (voire un 2 ou un 3…). On peut même raffiner, en mettant un format d'affichage conditionnel dans la colonne B : par exemple si le contenu est 0 le mettre en couleur blanche, et si le contenu n'est pas 0 le mettre en rouge. Par ailleurs, on peut mettre la colonne A au format « date », ce qui rend l'affichage plus parlant.

Cette feuille de tableur est disponible, ainsi que celle donnant le graphique ci-dessus, sur les pages Classes virtuelles du site http://www.aromath.net.

Cas général (sans équiprobabilité des dates d'anniversaires)

On peut se demander si l'équiprobabilité des dates d'anniversaire est un modèle plausible. Je n'ai pas su trouver cette information sur le site de l'INSEE mais quelques petits sondages m'ont convaincu que ce n'est pas tout à fait le cas. Il me semble qu'il y a, dans mon département, un peu plus de naissances au printemps (exercice concret de terminale sur le paragraphe « adéquation de données statistiques à une loi équirépartie » : traiter les données relatives aux mois de naissance obtenues dans une maternité donnée[3]), ce phénomène étant moins marqué qu'il ne l'a été. Je me suis alors demandé si c'était une des causes de ma mésaventure. Ce qui suit prouve que non.

Notations

Soit n un entier fixé ($1 < n < 365$).
Soit L l'ensemble des 365-uplets de réels positifs dont la somme vaut 1. L est en bijection avec l'ensemble de toutes les lois de probabilités possibles d'une variable aléatoire pouvant prendre 365 valeurs qu'on désigne lui aussi par L. Dans L, il y a (1/365,…..,1/365), noté E, qui est l'équirépartition.

(3) On imagine la taille N de l'échantillon dont il faudrait disposer pour traiter sérieusement la question de l'équirépartition suivant les jours de naissance ! La condition N > 100, parfois rencontrée, est évidemment abusivement simplificatrice. Lorsqu'on considère le vrai test du khi-deux, généralisant le test d'équirépartition de la classe de terminale et permettant de tester la loi de probabilité définie par $(p_1, p_2, \ldots, p_k)$ d'une variable aléatoire (avec les p_i ne valant pas nécessairement $1/k$), on rencontre dans la littérature la condition : les nombres Np_i, qui sont les effectifs « attendus » ou « théoriques », dépassent 5, ce qui donne avec $p_i = 1/365$, $N \geq 1825$.

Soit P = (p_1, p_2, …, p_{365}) un élément de L, p_i désignant la probabilité (quelconque) pour qu'une naissance ait lieu le i-ème jour de l'année (pour tout i entre 1 et 365).

On rappelle que $P(B_n)$ est la probabilité de **non-coïncidence des dates** d'anniversaires pour n personnes quand on a la distribution P des dates pour la population considérée.

Expression de $P(B_n)$ en fonction des p_i :

Pour tout k entre 1 et n et tout i entre 1 et 365, désignons par $E_{k,i}$ l'événement : la k-ème personne choisie est née le i-ème jour de son année de naissance. $P(E_{k,i})$ ne dépend pas de k et vaut p_i.

$$\Omega = \bigcup_{i_1=1}^{365}\left(\bigcup_{i_2=1}^{365}\cdots\left(\bigcup_{i_n=1}^{365}\left(E_{1,i_1}\cap E_{2,i_2}\cap\cdots\cap E_{n,i_n}\right)\right)\right)$$

où les n réunions successives sont disjointes.
B_n s'écrit de la même façon avec la condition supplémentaire que les i_k sont tous distincts.

Les événements E_{k,i_k} sont indépendants(4) donc $P\left(E_{1,i_1}\cap E_{2,i_2}\cap\cdots\cap E_{n,i_n}\right)$ est le produit des $P\left(E_{k,i_k}\right)$ pour k variant de 1 à n, c'est-à-dire le produit des p_{i_k} et la probabilité de non-coïncidence $P(B_n)$ vaut donc

$$\sum_{i_1}\sum_{i_2\neq i_1}\sum_{i_3\neq i_1,i_2}\cdots\sum_{i_n\neq i_1,i_2,\cdots,i_{n-1}}\left(p_{i_1}p_{i_2}\cdots p_{i_n}\right)$$

(les n sommes successives étant prises a priori de 1 à 365).
Cette expression de $P(B_n)$ montre que si s et t sont deux des nombres p_1, p_2, …, p_{365},

$$P(B_n) = (s+t)\,F + st\,G + H. \qquad (1)$$

où F, G et H sont des sommes de produits des probabilités p_1, p_2, …, p_{365} autres que s et t (ces produits étant formés de $n-1$, $n-2$ et n facteurs respectivement pour F, G et H).

On va montrer que ce nombre $P(B_n)$ est nécessairement inférieur ou égal à $E(B_n)$, c'est à dire $\dfrac{365!}{(365-n)!}\times\dfrac{1}{365^n}$.
L'inégalité qu'on obtiendra est une généralisation de l'inégalité arithmético-géométrique(5). Elle sera surtout la preuve que :
c'est dans le cas de l'équiprobabilité qu'il y a le moins de coïncidences.

(4) C'est le choix fait dans tout ce texte : on suppose qu'on obtient les n dates de naissance au terme de n épreuves identiques et indépendantes.
(5) Cette inégalité dit que si x_1, x_2, …, x_n sont des réels positifs de somme S fixée, le produit $x_1\,x_2\,\ldots\,x_n$ est maximum si $x_1 = x_2 = \ldots = x_n = \dfrac{S}{n}$. À un facteur 365 ! près, c'est le cas particulier $n = 365$ de l'inégalité proposée ici car poser S = 1 ne restreint pas la généralité.

Principe de la preuve

Soit f l'application définie sur $\mathbf{R}^{365}$ par

$$f\left(\left(x_1, x_2, \ldots, x_{365}\right)\right) = \sum_{i_1} \sum_{i_2 \neq i_1} \sum_{i_3 \neq i_1, i_2} \cdots \sum_{i_n \neq i_1, i_2, \cdots, i_{n-1}} \left(x_{i_1} x_{i_2} \cdots x_{i_n}\right).$$

f est continue sur $\mathbf{R}^{365}$ donc si on restreint f à L qui est un compact(6) de $\mathbf{R}^{365}$, f atteint son maximum sur L.
On va montrer que pour toute loi P dans L autre que l'équirépartition E, il existe P′ dans L telle que $P'(B_n) > P(B_n)$. On pourra en déduire que $E(B_n)$ est le maximum de f sur L.

Soit P dans L définie par $P = (p_1, p_2, \ldots, p_{365})$. Comme annoncé, on suppose que $P \neq E$. Il existe alors deux entiers s et t compris entre 1 et 365 tels que $p_s \neq p_t$.
L'égalité (1) s'écrit :

$$P(B_n) = (p_s + p_t)\,F + p_s \cdot p_t\,G + H.$$

Soit alors P′ dans L définie par les mêmes probabilités p_i que P à ceci près qu'on remplace p_t et p_t par $\dfrac{p_s + p_t}{2}$ qui est donc pour cette loi P′ la probabilité de naissance le s-ème jour de l'année de naissance et aussi celle relative au t-ème jour.
On a alors

$$P'(B_n) = (p_s + p_t)F + \left(\frac{p_s + p_t}{2}\right)^2 G + H$$

et donc

$$P'(B_n) - P(B_n) = \left[\left(\frac{p_s + p_t}{2}\right)^2 - p_s p_t\right] G = \left(\frac{p_s - p_t}{2}\right)^2 G$$

qui est strictement positif car on a supposé $p_s \neq p_t$.
Donc $P'(B_n) > P(B_n)$.

Notes

1) Cette inégalité est intuitivement « évidente » ou tout au moins compréhensible si on pense à des cas extrêmes dans L du type « toutes les naissances ont eu lieu en deux mois ou même sur une durée encore plus courte » et il me semble qu'on peut donc en parler aux élèves s'ils posent des questions à ce sujet.

2) Ce problème des coïncidences est dû, semble-t-il, à Richard von Mises (1883-1953), dans « Uber Aufleitungs- und Besetzungs- Wahrscheinlichkeiten », in Revue de la faculté des sciences d'Istambul N.S., volume 4, p. 145-163. La page http://www-gap.dcs.st-and.ac.uk/~history/Mathematicians/Mises.html contient une biographie de ce mathématicien.

(6) Cet argument de compacité de L évite la construction, pour une loi P quelconque dans L, d'une suite $(P_k)_{1 \leq k \leq N}$ de L vérifiant $P_1 = P$, $P_N = E$ et $(P_k(B_n))_{(k)}$ est une suite croissante.

3) La page http://mathworld.wolfram.com/BirthdayProblem.html contient des informations sur le problème de k ($k \geq 1$) coïncidences.

Conclusion

Ce qui précède montre que nous pouvons continuer à poser sereinement « l'exercice des anniversaires » en l'agrémentant de sondages ; quelle que soit la classe considérée, la probabilité d'au moins une coïncidence est au moins celle « espérée », c'est-à-dire calculée dans le cas de l'équiprobabilité, même si ça n'empêche pas que, quasi-fatalement, un d'entre nous, ou un de nos successeurs, passera un jour une heure entière à se ridiculiser en donnant des prédictions toutes fausses, éventuellement pimentées de simulations toutes aberrantes(7) ! Il y a des mauvais jours, peu nombreux.

(7) Il n'est même pas nécessaire de penser que les générateurs de « nombres aléatoires » n'engendrent en fait que des nombres pseudo-aléatoires pour imaginer que cette situation est possible.