Le paradoxe des anniversaires et l'Euro 2016

Paul Jolissaint, UniNE

18 octobre 2019

Largement inspiré par le site `blogdemaths.wordpress.com`.

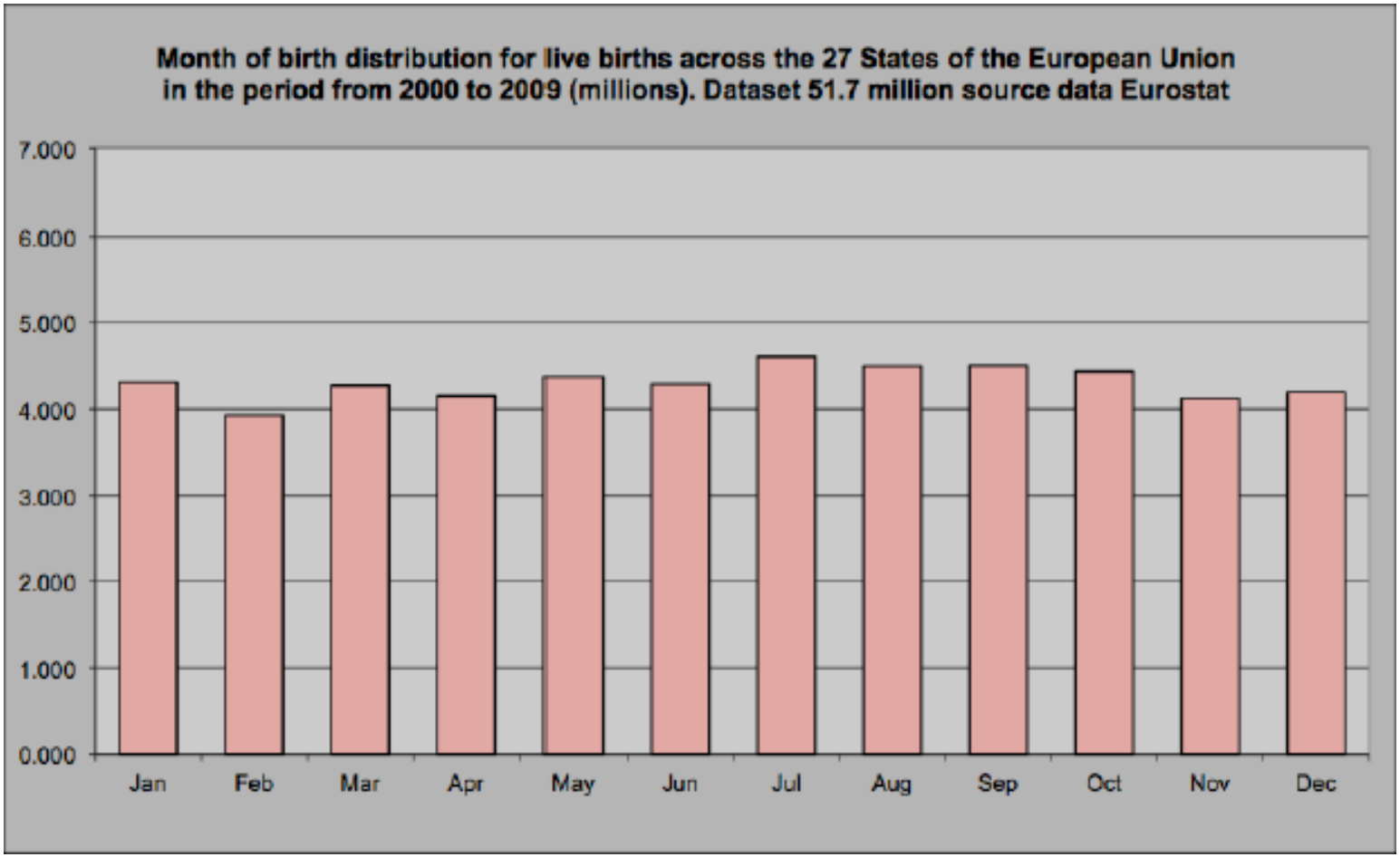
Le **paradoxe des anniversaires** affirme que, dans une population de 23 personnes, la probabilité qu'au moins deux d'entre elles aient leur anniversaire le même jour est approximativement égale à 0.51.

On parle de paradoxe car la probabilité est considérée intuitivement comme particulièrement élevée.

Deux précisions tout d'abord :

On suppose que chaque date entre le 1er janvier et le 31 décembre est équiprobable ;

cela est justifié par exemple par le tableau suivant qui montre la répartition mensuelle des naissances dans 27 pays d'Europe entre 2000 et 2009 et qui concerne plus de 50 millions de naissances.


Month of birth distribution for live births across the 27 States of the European Union
in the period from 2000 to 2009 (millions). Dataset 51.7 million source data Eurostat
7.000
6.000
5.000
4.000
3.000
2.000
1.000
0.000
Jan
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct
Nov
Dec

Ensuite, on néglige la date du 29 février (on l'identifie par exemple avec le 1er mars). Nous reviendrons sur cette hypothèse à la fin.

Petite digression sur les calendriers

Le premier calendrier romain date du VII^{e} siècle avant J.-C. et il comptait 304 jours.

Il commençait le **1er mars**, et des jours étaient ajoutés pour égaler l'année solaire. Cela explique les noms des 4 derniers mois de l'année :

septembre (de *septem*, 7^{e} mois), octobre (de *octo*, 8^{e} mois), novembre (de *novem*, 9^{e} mois) et décembre (de *decem*, 10^{e} mois).

En 46 avant J.-C., Jules César décida d'introduire un nouveau calendrier, appelé depuis *calendrier julien* : il comptait 365 jours, et une année bissextile (deux fois six) tous les quatre ans.

Puisque l'année solaire n'est pas exactement égale à 365.25 jours mais légèrement plus courte, un décalage de 8 jours environ sur un millénaire devenait tout à fait perceptible.

Au XVI^{e} siècle, le pape Grégoire XIII a décidé d'instaurer un nouveau calendrier.

Ainsi, le lendemain du 4 octobre 1582 fut le 15 octobre 1582.

Les années multiples de 100 mais pas de 400 sont non-bissextiles.

Enfin, le mot *calendrier* vient du latin *calendæ* qui désignait le premier jour de chaque mois chez les Romains. (Fin de la digression.)

Démonstration du paradoxe :

Traduisons maintenant le problème des anniversaires mathématiquement dans un groupe de k personnes :

Cela revient à choisir k nombres *au hasard* entre 1 et 365, chacun ayant la même probabilité d'être choisi.

Appelons donc A_k l'événement : *obtenir au moins 2 fois la même valeur parmi k choix dans* $\{1, \ldots, 365\}$.

On cherche la valeur de $P(A_k)$. Nous allons utiliser le théorème bien connu :

$$P(A_k) = 1 - P(\bar{A}_k).$$

Proposition. On a :

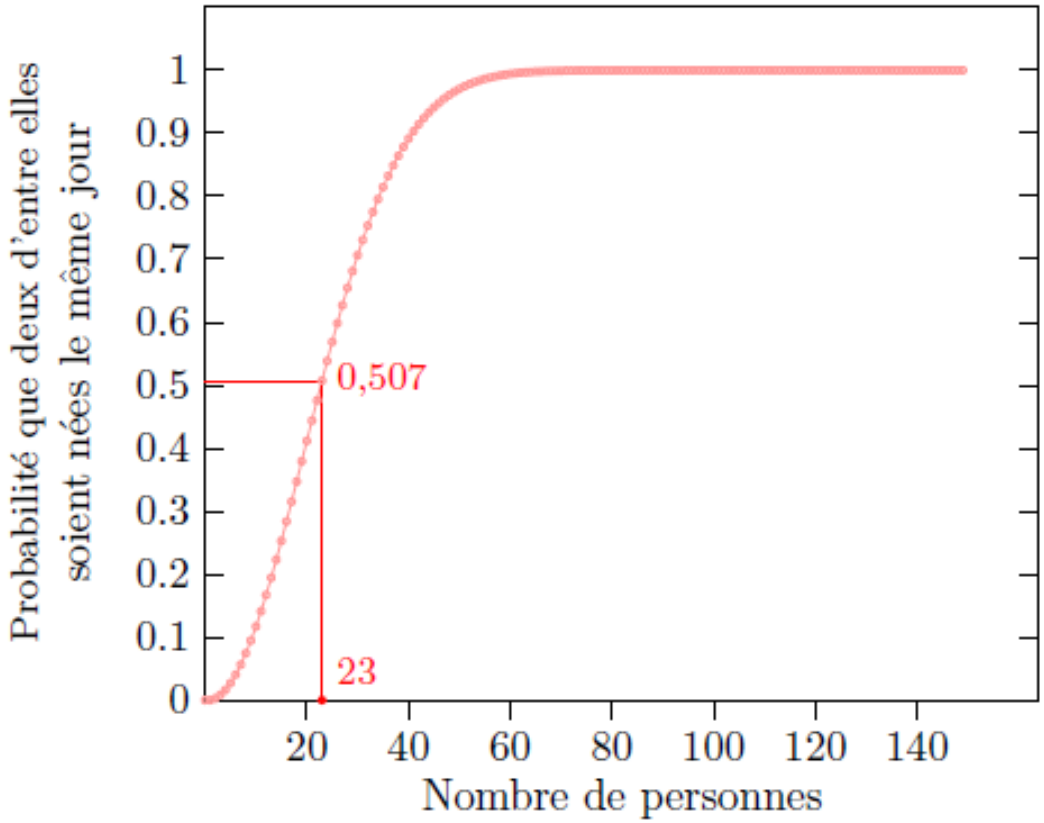
$$P(\bar{A}_k) = \frac{365 \cdot 364 \cdots (365 - k + 1)}{365^k}.$$

Donc

$$P(A_k) = \frac{365^k - 365 \cdot 364 \cdots (365 - k + 1)}{365^k}.$$

Preuve de la proposition. L'univers Ω associé à cette expérience aléatoire est l'ensemble des k-tuples $(a_1, a_2, \ldots, a_k)$ de nombres entiers compris entre 1 et 365. Il y en a en tout 365^k (arrangements avec répétitions). Les issues qui sont dans $\bar{A}_k$ sont les k-tuples $(a_1, a_2, \ldots, a_k)$ pour lesquels $a_i \neq a_j$ pour tous $i \neq j$. Il s'agit donc d'arrangements simples ; il y en a
$A_k^{365} = 365 \cdot 364 \cdots (365 - k + 1)$.
On obtient immédiatement la première affirmation de la proposition. □

Le tableau suivant montre les valeurs de $P(A_k)$ en fonction de k.



On a en particulier $P(A_{23}) = 0.507 \approx 0.51$.

On obtient la représentation graphique ci-dessus en utilisant par exemple les suites $(a(k))$ et $(b(k))$ suivantes :

$$b(1) = 1,\ b(k) = b(k-1) \cdot \left(1 - \frac{k-1}{365}\right)$$

puis en posant $a(k) = 1 - b(k)$.

Cela provient de

$$P(\bar{A}_k) = \frac{365 \cdot 364 \cdots (365 - (k-1))}{365^k} = 1 \cdot \frac{364}{365} \cdots \left(1 - \frac{k-1}{365}\right).$$

Avant de passer à l'Euro 2016, donnons une expression approximative de $P(A_k)$:

On peut réécrire

$$P(A_k) = 1 - \left(1 - \frac{1}{365}\right)\left(1 - \frac{2}{365}\right)\cdots\left(1 - \frac{k-1}{365}\right).$$

Or, pour x proche de 0, on a l'approximation $e^x \approx 1 + x$ qui implique, pour k pas trop proche de 365,

$$P(A_k) \approx 1 - e^{-1/365} \cdot e^{-2/365} \cdots e^{-(k-1)/365} = 1 - \prod_{\ell=1}^{k-1} e^{-\ell/365}$$

puis, comme

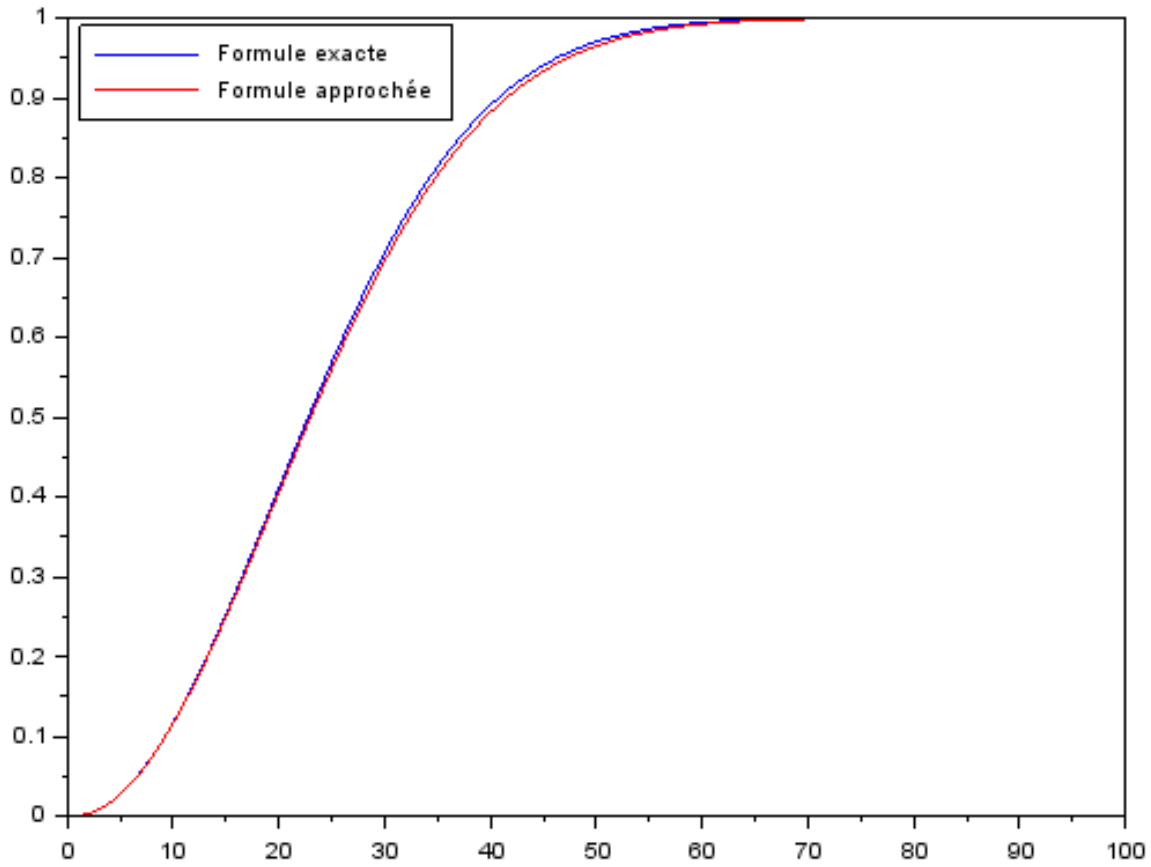
$$\prod_{\ell=1}^{k-1} e^{-\ell/365} = e^{-\frac{1}{365}\sum_{\ell=1}^{k-1}\ell} = e^{-k(k-1)/730},$$

on obtient finalement

$$P(A_k) \approx 1 - e^{-k(k-1)/730}$$

pour tout $k << 365$.

Voici les représentations graphiques des formules exacte et approximative sur l'intervalle $[1, 100]$:



La formule approximative permet également, pour une probabilité $0 < p < 1$ donnée, de déterminer une valeur de k pour laquelle $P(A_k) \approx p$:

On a

$$p = 1 - e^{-k(k-1)/730}$$

donc

$$\ln\left(\frac{1}{1-p}\right) = \frac{k(k-1)}{730}$$

et ainsi, avec $k(k-1) \approx k^2$,

$$k \approx \sqrt{730 \ln\left(\frac{1}{1-p}\right)}.$$

Par exemple, on trouve $k \approx 22.5$ pour $p = 0.5$, et $k \approx 58$ pour $p = 0.99$.

Lors de l'Euro 2016, 24 équipes nationales étaient en compétition, chacune étant formée de 23 joueurs. Voyons lesquelles de ces 24 équipes satisfont le paradoxe.

Pays	**Joueurs nés le même jour**
Albanie	O. Shehi et A. Ajeti (25 septembre)
Allemagne	Aucun
Angleterre	K. Walker et J. Stones (25 mai)
Autriche	Aucun
Belgique	K. De Bruyne et J. Denayer (28 juin)

Croatie	M. Kovacic et M. Pjaca (6 mai)
Espagne	K. et D. Silva (8 janvier)
France	D. Payet et N'golo Kanté (29 mars) et A. Martial et A-P. Gignac (5 décembre)
Hongrie	B. Bese et P. Gulacsi (6 mai)
Irlande du Nord	M. McGovern et S. Ferguson (12 juillet)
Islande	R. Sigurdsson et O. Kristinsson (19 juin)
Italie	Aucun
Pays de Galles	J. Chester et J. Ledley (23 janvier) et A. Williams et J. Collins (23 août)
Pologne	W. Szcsesny et L. Fabiansky (18 avril)

Portugal	A. Lopes et Eliseu (1er octobre)
Rép. d'Irlande	A. McGeady et S. Quinn (4 avril)
Rép. Tchèque	T. Vaclik et M. Suchy (29 mars)
Roumanie	Aucun
Russie	R. Shishkin et D. Glushakov (27 janvier) et A. et V. Berezutski (20 juin, jumeaux)
Slovaquie	J. Mucha et O. Duda (5 décembre)
Suède	R. Olsen et P. Carlgren (8 janvier) et A. Isaksson et Z. Ibrahimovic (3 octobre)
Suisse	Aucun

Turquie	S. Kaya et Y. Malli (24 février) et H. Balta et O. Tufan (23 mars)
Ukraine	Aucun

Il y a donc 18 équipes sur 24 qui possèdent au moins deux joueurs ayant leur anniversaire le même jour, ce qui représente le 75% des équipes. Le paradoxe des anniversaires prévoit qu'environ le 50% des équipes auraient au moins deux joueurs nés le même jour.

Ce résultat est-il acceptable (fruit du hasard) ou y a-t-il un biais ?

Pour y répondre, nous allons faire un test d'hypothèse.

Étape 1 On commence par faire une hypothèse appelée **hypothèse nulle** ; dans notre cas :
Dans un groupe de 23 joueurs de l'Euro, la probabilité p que deux (au moins) soient nés le même jour vaut $p = 0.51$,
autrement dit, les joueurs de l'Euro se comportent comme la population normale.
(Hypothèse alternative : $p > 0.51$ puisqu'on a trouvé les trois quarts des équipes.)

Étape 2 On fixe de manière un peu arbitraire une certaine probabilité de se tromper, appelée **seuil** ; généralement on prend 0.05. Cela signifie que si on arrive à la conclusion que l'hypothèse nulle est fausse, ce sera avec une probabilité de se tromper de 5%, donc de 95% de dire vrai.

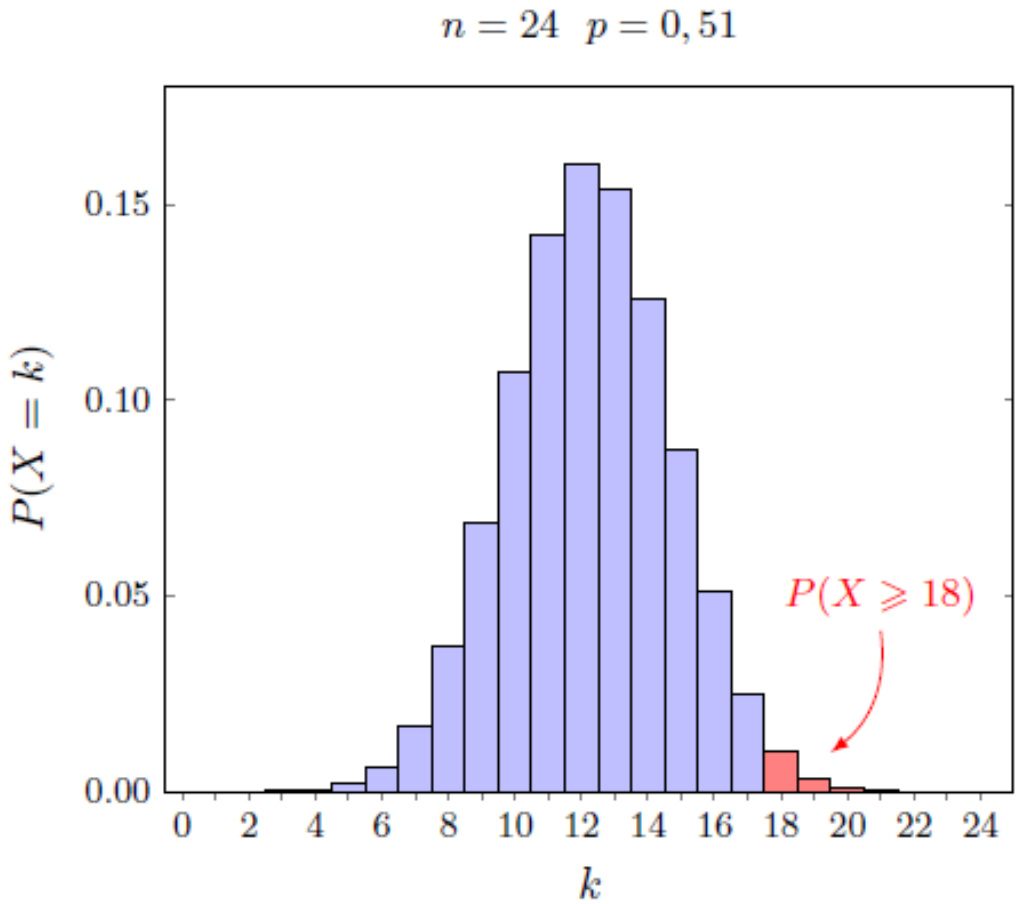
Étape 3 En supposant l'hypothèse nulle réalisée, nous allons calculer la probabilité p' de trouver au moins 18 équipes sur 24 qui contiennent chacune au moins deux joueurs nés le même jour.

Si on trouve $p' \geq 0.05$, alors on ne peut pas rejeter l'hypothèse nulle, mais si $p' < 0.05$, notre hypothèse nulle est fausse et on peut la rejeter.

Les équipes étant indépendantes les unes des autres, nous allons appliquer aux 24 équipes la loi binomiale avec $n = 24$ et la probabilité du "succès" égale à 0.51.
Nous devons calculer la probabilité d'obtenir au moins 18 succès sur 24 répétitions de l'expérience, c'est-à-dire

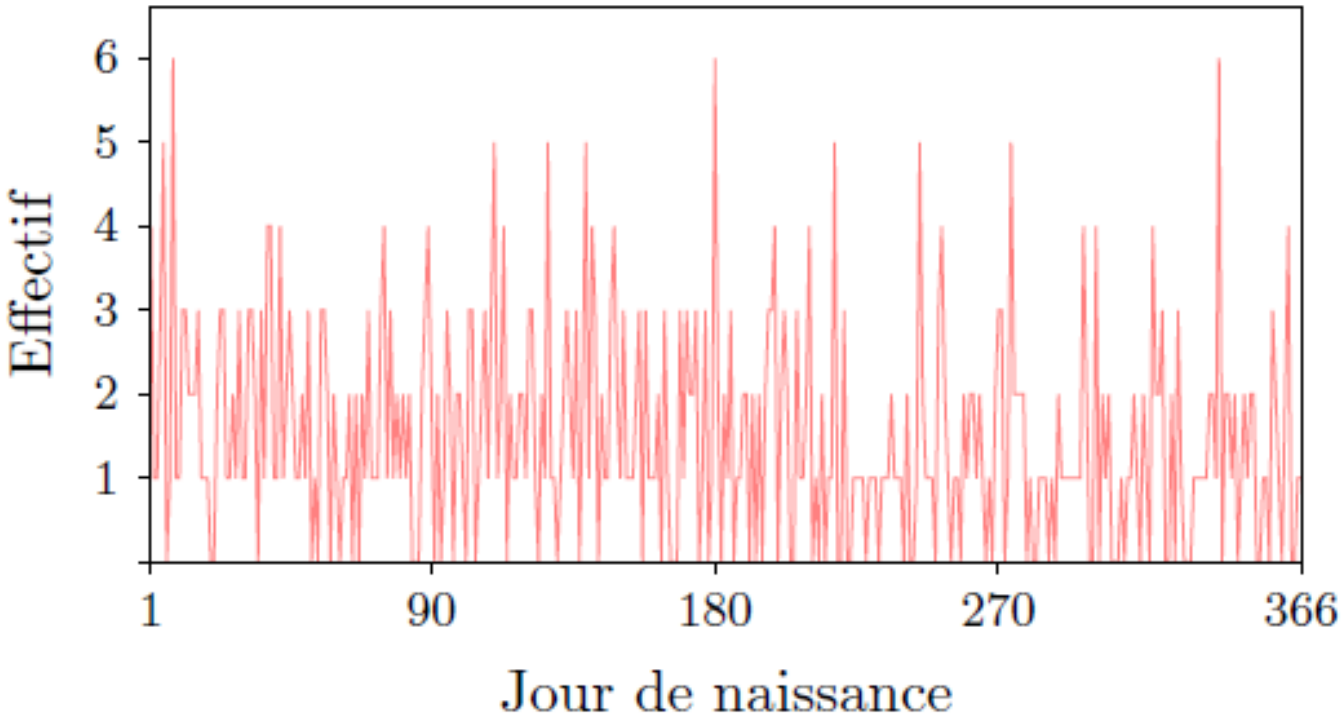
$$p' = C_{18}^{24} \cdot 0.51^{18} 0.49^{6} + C_{19}^{24} \cdot 0.51^{19} 0.49^{5} + \cdots + C_{24}^{24} \cdot 0.51^{24}.$$

On trouve $p' \approx 0.015$.

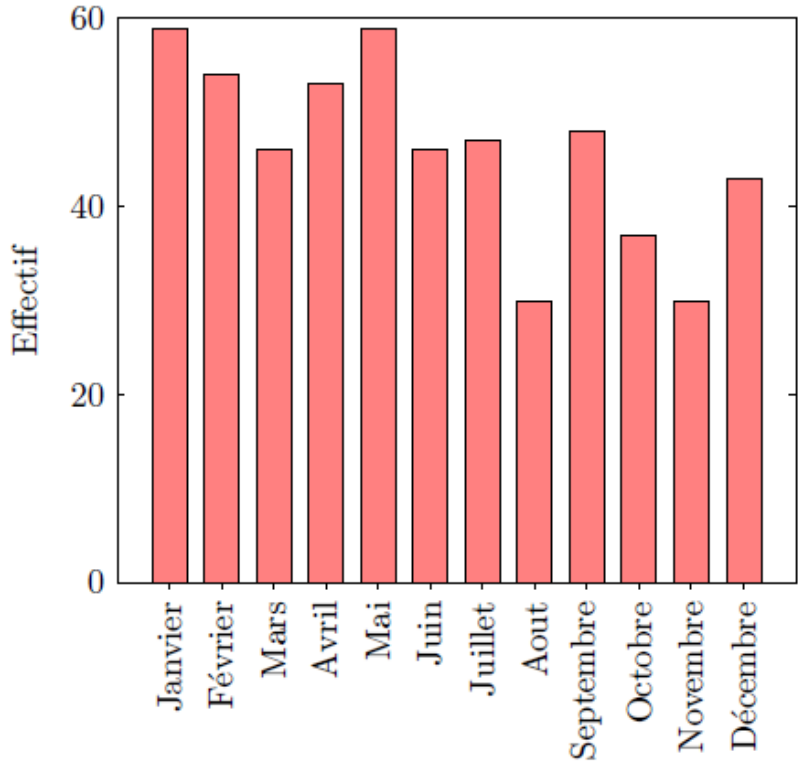

$n = 24 \quad p = 0,51$
$P(X = k)$
0.15
0.10
0.05
0.00
$P(X \geqslant 18)$
0 2 4 6 8 10 12 14 16 18 20 22 24
k

Comme $p' = 0.015 < 0.05$, cela signifie qu'il y a moins de 5% de chances d'avoir au moins 18 équipes qui possèdent au moins deux joueurs nés le même jour.
Ainsi, avec 5% de chances de nous tromper, nous pouvons affirmer que la probabilité de trouver au moins deux joueurs nés le même jour dans les équipes de l'Euro 2016 est supérieure à 0.51.

Nous venons de voir que la répartition des naissances des joueurs de l'Euro 2016 n'est pas conforme à celle de la population.
Pour en comprendre la raison, il faut considérer les statistiques des dates de naissance des 552 joueurs :



Il y a un joueur de l'Euro 2016 qui est né un 29 février : Benedikt Höwedes de l'équipe d'Allemagne ! En regroupant les naissances par mois on a :



On voit que les 5 premiers mois de l'année sont prédominants, même février qui ne compte que 28 ou 29 jours !

Ce phénomène, appelé **relative age effect**, dit que les personnes nées en début d'année ont plus de chance de devenir footballeurs professionnels que celles nées dans la seconde partie de l'année.

Cela est dû au fait que les catégories de jeunes joueurs sont déterminées par l'âge (M15, M16,. . .) et qu'il y a une *date limite* qui est le 1er janvier pour chaque catégorie.

Dans une même catégorie d'âge, un jeune né en janvier aura un écart de développement physique plus important qu'un jeune né en novembre ou en décembre (presque 11 mois d'écart), et il aura plus de chances d'être repéré par les recruteurs !

Finalement, puisque les naissances des joueurs de l'Euro sont plus concentrées sur certains mois, il y a plus de chances que plusieurs soient nés le même jour !

Formule exacte du paradoxe des anniversaires

On considère un ensemble de $2 \leq k \leq 366$ personnes, que l'on numérote de 1 à k.

On numérote également les dates possibles d'anniversaire de $i = 1$ à $i = 366$, en convenant par exemple que 1 correspond au 1^{er} mars, 2 au 2 mars et finalement 366 au 29 février.

Pour tout $1 \leq j \leq k$, soit X_j la variable aléatoire qui donne la date de naissance de j ; on suppose les X_j indépendantes et identiquement distribuées.

On a alors $P(X_j = i) = a_i$ où

$$a_i = \begin{cases} \frac{4}{1461} & i \leq 365 \\ \\ \frac{1}{1461} & i = 366. \end{cases}$$

L'événement $\bar{A}_k$, qui est le fait que les k personnes ont des dates d'anniversaires toutes différentes, correspond à l'ensemble $\mathcal{A}_k$ de tous les arrangements simples $(i_1, \ldots, i_k)$ de k objets parmi $\{1, \ldots, 366\}$,

donc

$$\bar{A}_k = \bigsqcup_{(i_1,\ldots,i_k)\in\mathcal{A}_k} \{(X_1, \ldots, X_k) = (i_1, \ldots, i_k)\}.$$

Ainsi,

$$P(\bar{A}_k) = \sum_{(i_1,\ldots,i_k)\in\mathcal{A}_k} a_{i_1} \cdots a_{i_k}.$$

Or, $\mathcal{A}_k = \mathcal{A}_k(1) \sqcup \mathcal{A}_k(2)$ où

$\mathcal{A}_k(1)$ est l'ensemble des arrangements sans répétitions de k objets parmi $\{1, \ldots, 365\}$, et

$\mathcal{A}_k(2)$ est l'ensemble des arrangements sans répétitions de k objets parmi $\{1, \ldots, 366\}$, dont un (exactement) est égal à 366.

En désignant par A_j^m le nombre d'arrangements sans répétitions de j objets parmi m, on a

$$|\mathcal{A}_k(1)| = A_k^{365} = 365 \cdot 364 \cdots (365 - k + 1)$$

et

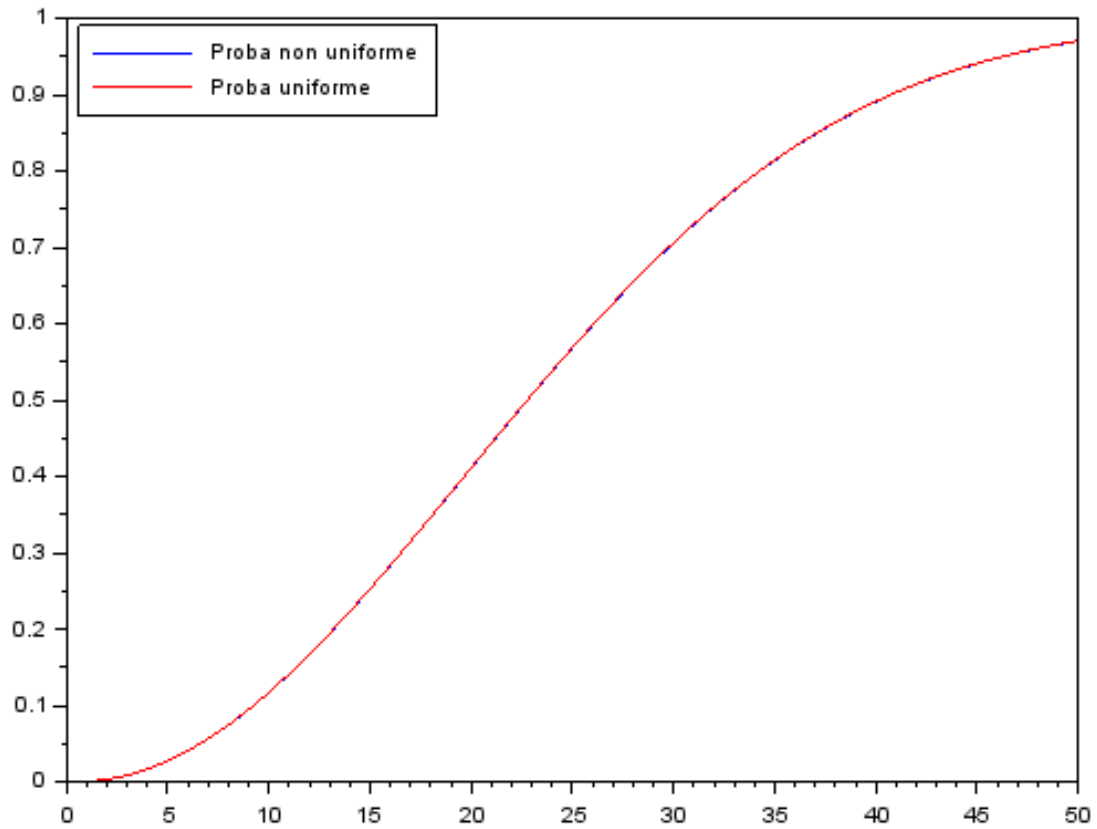
$$|\mathcal{A}_k(2)| = k \cdot A_{k-1}^{365} = 365 \cdot 364 \cdots (365 - k + 2)k.$$

En effet, la première égalité est évidente, et la deuxième suit du fait que l'on a k choix possibles pour la place de 366 et A_{k-1}^{365} pour les $k-1$ dates restantes (cf. la relation de récurrence : $A_j^m = A_j^{m-1} + j \cdot A_{j-1}^{m-1}$).

On obtient finalement

$$\begin{aligned} P(\bar{A}_k) &= \frac{4^k}{1461^k} \cdot 365 \cdot 364 \cdots (365-k+1) \\ &\quad + \frac{4^{k-1}}{1461^k} \cdot 365 \cdot 364 \cdots (365-k+2)k \\ &= \frac{4^{k-1}}{1461^k} \cdot 365 \cdot 364 \cdots (365-k+2)[4(365-k+1)+k] \\ &= \frac{4^{k-1}}{1461^k} \underbrace{\cdot 365 \cdot 364 \cdots (365-k+2)}_{k-1 \text{ termes}} (1464-3k). \end{aligned}$$

Les différences entre le cas de probabilité uniforme et de probabilité non uniforme sont pratiquement imperceptibles (cas de 50 personnes) :



Pour programmer les calculs, on exprime $P(\bar{A}_k)$ ainsi :

$$P(\bar{A}_k) = \frac{4 \cdot 365}{1461} \cdots \frac{4 \cdot (365 - (k-2))}{1461} \cdot \left(1 - \frac{3k-3}{1461}\right)$$

et comme, pour $2 \leq j \leq k$, on a

$$\frac{4 \cdot (365 - j + 2)}{1461} = 1 - \frac{4j-7}{1461},$$

on a

$$P(\bar{A}_k) = \prod_{j=2}^{k} \left(1 - \frac{4k-7}{1461}\right) \cdot \left(1 - \frac{3k-3}{1461}\right).$$