# Le problème du collectionneur.

Robert FERREOL ferreol@mathcurve.com

Un collectionneur doit réunir *n* pièces. Ces pièces lui parviennent une par une, mais il y a malheureusement des doubles, chaque modèle de pièce ayant une probabilité égale à 1/n d'arriver à chaque fois. Il s'agit par exemple d'un enfant qui collectionne une série d'images Panini, ou les magnets des départements français distribués par la marque « le gaulois », d'un amateur d'euros qui veut obtenir toutes les pièces européennes, ou d'un joueur qui lance un dé et attend d'avoir tous les nombres de 1 à 6.

Voici quelques questions auxquelles vous pouvez essayer de répondre, avant que nous ne les résolvions.

1) Quelle est l'espérance du nombre de pièces à recevoir pour obtenir une collection complète ?

2) Quelle est sa variance ?

3 ) Lorsque le collectionneur a reçu $N$ pièces, quelle est la probabilité que sa collection soit complète ?

4) Et quelle est l'espérance du nombre de pièces distinctes qu'il possède ?

**1) Espérance du nombre de pièces à recevoir pour obtenir une collection complète.**

Nous allons pour cela utiliser une propriété très importante pour la compréhension des probabilités et rarement mentionnée sous cette forme : lors d'une succession d'épreuves aléatoires répétées indépendantes, si un événement a une probabilité *p*, cet événement arrive en moyenne tous les 1/*p* coups.
Cette propriété est très naturelle : elle signifie par exemple que si un événement a une chance sur 3 d'arriver, il arrivera en moyenne une fois sur 3… Autre exemple : la probabilité d'avoir 4 as lors d'une distribution de 52 cartes à 4 joueurs est égale à $\frac{\binom{48}{9}}{\binom{52}{13}} \approx 0{,}0027$ ; ce nombre n'est pas très parlant. Mais si l'on dit que cela signifie qu'on aura en moyenne 4 as tous les $\frac{\binom{52}{13}}{\binom{48}{9}} \approx 379$ coups, ça l'est plus !

Voici la démonstration de cette propriété :
Un événement de probabilité $p$ va arriver au bout de $k$ coups avec la probabilité $(1-p)^{k-1}p$ ; le temps d'attente moyen (en nombre de coups) de cet événement est donc

$T_A = \sum_{k=1}^{\infty} k(1-p)^{k-1} p$ ; or on sait que $\sum_{k=1}^{\infty} kx^{k-1} = \frac{d}{dx}\left(\sum_{k=1}^{\infty} x^k\right) = \frac{d}{dx}\left(\frac{1}{1-x}\right) = \frac{1}{(1-x)^2}$ ; donc $T_A = \frac{p}{(1-1+p)^2} = \frac{1}{p}$.

Revenons donc à notre problème ; Supposons que je reçoive une pièce par jour et disons que ce matin j'ai $k$ pièces (distinctes) dans ma collection ; la probabilité que la pièce qui va m'arriver aujourd'hui soit nouvelle vaut $\frac{n-k}{n}$ ; le temps d'attente moyen de cette nouvelle pièce vaut donc, d'après la propriété ci-dessus, $\frac{n}{n-k}$ jours ; le temps d'attente moyen de la collection complète vaut donc $N = \sum_{k=0}^{n-1} \frac{n}{n-k} = \left(1+\frac{1}{2}+....+\frac{1}{n}\right)n = \boxed{nH_n}$ jours ($H_n$ est le $n$-ième nombre harmonique).

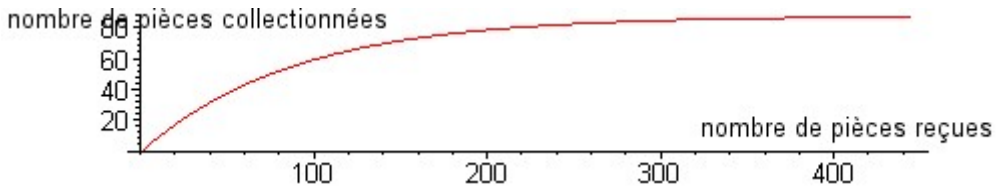
Conclusion : pour terminer une collection de $n$ pièces, je devrai en réunir environ $nH_n$.

Pour $n = 88$ (11 pays à 8 pièces chacun), cela fait 445 pièces à devoir obtenir… Bigre !

Mais ce sont les dernières pièces qui sont les plus longues à obtenir : par exemple, le temps d'attente moyen de la dernière pièce vaut $n$, autant que le nombre de pièces à collectionner ! Et un calcul montre que pour avoir seulement 60 pièces sans double, il suffit d'en collectionner 103 environ.

Voici d'ailleurs la courbe de remplissage de la collection en fonction du temps, obtenue par maple :

```
n:=88:plot([seq([sum(n/(n-k),k=0..q),q],q=0..n-1)]);
```




En face d'une telle courbe discrète, le matheux a envie de savoir s'il y a une courbe continue approchante pour $n$ grand.

Or il est bien connu que le $n$-ième nombre harmonique $H_n$ équivaut à $\ln n$ (car $\ln n < H_n < \ln n + 1$) ; remplaçons froidement le temps $t = \sum_{k=0}^{q-1} \frac{n}{n-k} = n\left(\frac{1}{n-q+1}+...+\frac{1}{n}\right)$ par $n(\ln n - \ln(n-q))$ ; on obtient $\frac{t}{n} = -\ln\left(1-\frac{q}{n}\right)$, soit $\boxed{q = n\left(1-e^{-\frac{t}{n}}\right)}$ où $q$ est le nombre de pièces dans la collection , et $t$ le nombre de pièces réellement nécessaires pour les obtenir.

C'est vraiment étrange comme la courbe continue et dérivable associée ressemble à la courbe discrète qu'elle approche ! Si l'on superpose le tracé précédent avec celui obtenu par `plot(n*(1-exp(-t/n)),t=0..n*ln(n));` on obtient un tracé quasi identique.

**2) Variance et écart-type du nombre de pièces à recevoir pour obtenir une collection complète.**

Commençons par chercher la variance de notre temps d'attente de l'évènement de probabilité $p$, arrivant toujours au bout de $k$ coups avec la probabilité $(1-p)^{k-1}p$ ; d'après la formule $V(X)=E(X^2)-(E(X))^2$, cette variance vaut $V=\sum_{k=1}^{\infty}\left(k^2(1-p)^{k-1}p\right)-\frac{1}{p^2}$ qui est égale, par le calcul, cette fois, de la dérivée seconde de $\sum x^k$, à $\boxed{V=\frac{1-p}{p^2}}$.

Comme, ici, le fait de recevoir une pièce nouvelle ou non est indépendant de l'état de ma collection, la variance totale est la somme des variances partielles d'où : $V=\sum_{k=0}^{n-1}\frac{1-\frac{n-k}{k}}{\left(\frac{n-k}{k}\right)^2}=n\sum_{k=0}^{n-1}\frac{k}{(n-k)^2}=n\sum_{k=1}^{n}\frac{n-k}{k^2}=\boxed{n^2\sum_{k=1}^{n}\frac{1}{k^2}-nH_n}$.

Comme $\sum_{k=1}^{\infty}\frac{1}{k^2}=\frac{\pi^2}{6}$, et $H_n << n$, $V$ est équivalent à $\frac{\pi^2}{6}n^2$ et l'écart-type $\sigma$ à $\frac{\pi}{\sqrt{6}}n\approx 1{,}3.n$.
Amusant que π intervienne dans cette affaire !
Avec les données précédentes, cela fait un écart-type de 110 pièces (pour, rappelons-le, 445 pièces en moyenne), donc de grosses fluctuations possibles !

**3) Probabilité que la collection soit complète lorsque le collectionneur a reçu *N* (≥ *n*) pièces.**

Un tirage de $N$ pièces choisies parmi les $n$ pièces à collectionner équivaut au tirage au hasard d'une application $f$ d'un $N$-ensemble vers un $n$-ensemble ; il y a donc $n^N$ tels tirages.

Le fait que la collection soit alors complète équivaut au fait que $f$ soit surjective.
La probabilité de collection complète vaut donc $p_{N,n}=\frac{S_N^n}{n^N}$ où $S_N^n$ est le nombre de surjections d'un $N$-ensemble vers un $n$-ensemble.
Par exemple $p_{n,n}=\frac{n!}{n^n}$ qui tend très vite vers 0 : pour $n$ = 88, on trouve une probabilité de $1{,}5.10^{-37}$ de constitution de la collection en 88 coups !

Nous n'étudierons pas ici les nombres $S_N^n$ qui sont, à une factorielle près, les *nombres de Stirling de deuxième espèce* (pour obtenir $S_N^n$ avec maple, demander $n!$*Stirling2$(N,n)$).

Cherchons, pour $n$ = 88, combien de pièces il faut collectionner pour avoir au moins 50 % de chances d'obtenir la collection complète (autrement dit résolvons $\frac{S_N^{88}}{88^N}>\frac{1}{2}$) ; la réponse est 423 pièces au moins, nombre pas très différent du 445 de la question 1, mais pas exactement égal…

**4) Espérance du nombre de pièces distinctes lorsque le collectionneur a reçu *N* pièces.**

Il faut d'abord calculer la probabilité qu'il y ait *k* pièces distinctes. Il faut donc dénombrer les applications *f* pour lesquelles l'image de l'ensemble de départ possède *k* éléments. Si on fixe cette image, il y a $S_N^k$ telles *f* ; il faut maintenant multiplier par le nombre d'images possibles, soit le nombre de parties à *k* éléments, soit $\binom{n}{k}$ ; l'espérance du nombre de pièces distinctes vaut donc $\sum_{k=0}^{N} k\frac{\binom{n}{k}S_N^k}{n^N} = \boxed{\frac{1}{n^N}\sum_{k=0}^{N} k\binom{n}{k}S_N^k}$ .

On a évidemment envie de simplifier cette expression, surtout que l'on sait que $\sum_{k=0}^{N}\binom{n}{k}S_N^k = n^N$ (pourquoi, à propos ?).

La formule du pion $\binom{n}{k} = \frac{n}{k}\binom{n-1}{k-1}$, et la relation de Pascal vont être nos amies.

On peut écrire : $\sum_{k=0}^{N} k\binom{n}{k}S_N^k = n\sum_{k=1}^{N}\binom{n-1}{k-1}S_N^k = n\left(\sum_{k=1}^{N}\binom{n}{k}S_N^k - \sum_{k=1}^{N}\binom{n-1}{k}S_N^k\right)$.

Or $\sum_{k=1}^{N}\binom{n}{k}S_N^k = \sum_{k=0}^{N}\binom{n}{k}S_N^k = n^N$ et $\sum_{k=1}^{N}\binom{n-1}{k}S_N^k = \sum_{k=0}^{N}\binom{n-1}{k}S_N^k = (n-1)^N$, donc $\boxed{\sum_{k=0}^{N} k\binom{n}{k}S_N^k = n\left(n^N - (n-1)^N\right)}$, et l'espérance du nombre de pièces distinctes est tout simplement égal à $\boxed{n\left(1-\left(1-\frac{1}{n}\right)^N\right)}$. Cette forme montre tout de suite que, comme on pouvait l'espérer, ce nombre tend bien vers *n* quand *N* tend vers l'infini.

Toujours avec *n* = 88, un calcul numérique dit que pour une espérance de 60 (resp. 80) pièces distinctes, il me faudra recevoir au moins 101 (resp. 210) pièces.

Maintenant, un résultat aussi simple oblige à se demander s'il n'y aurait pas un raisonnement direct pour y arriver.
Le voici :
Lors d'une réception de *N* pièces, la probabilité de ne pas recevoir une pièce donnée vaut $\left(1-\frac{1}{n}\right)^N$ ; donc la probabilité de la recevoir (au moins une fois) vaut $1-\left(1-\frac{1}{n}\right)^N$ .
Mais la probabilité d'un événement est aussi l'espérance de la variable aléatoire valant 1 si l'événement est réalisé et 0 sinon.
Et ici, je suis en présence de *n* variables aléatoires associées à chacune des *n* pièces à collectionner. Et l'espérance du nombre de pièces distinctes est la somme des espérances de chacune de ces variables aléatoires. On retrouve bien le résultat $\boxed{n\left(1-\left(1-\frac{1}{n}\right)^N\right)}$ (aveu : je suis parti du résultat pour trouver ce raisonnement…)

Moralité : raisonner en variables aléatoires donne souvent des résultats plus simples que les raisonnements utilisant les dénombrements.

**5) La consultation d'un dictionnaire en plusieurs volumes.**

Je reconnais que le problème dont nous allons parler maintenant, à l'heure d'Internet et de Wikipedia, est un peu désuet.
Je le mentionne ici car j'ai mis beaucoup de temps (plusieurs années !) à m'apercevoir qu'il est en fait équivalent au précédent.
Mais voilà en quoi il consiste : lorsque vous avez, disons, $N = 10$ mots quelconques à consulter dans un dictionnaire en $n = 15$ volumes, vous avez espoir de ne pas avoir à ouvrir 10 volumes différents ! Mais combien de volumes aurez-vous à ouvrir en moyenne ?

Eh bien la réponse a été donnée dans la partie 4) précédente. En effet, chaque mot a une probabilité $1/n$ de se trouver dans un dictionnaire donné (si l'on suppose que les volumes ont le même nombre de mots). Et chaque nouveau mot à chercher est comme une nouvelle pièce qui arrive dans ma collection. Soit il se trouve dans un dictionnaire que j'ai déjà ouvert, soit il m'oblige à en ouvrir un nouveau…

La réponse numérique est donc $\boxed{15\left(1-\left(1-\frac{1}{15}\right)^{10}\right)}$ soit 7 dictionnaires et demi à ouvrir en moyenne.

Et comment s'interprètent alors dans ce contexte les résultats 1), 2), 3) ci-dessus ?

Le problème 1) est celui du nombre moyen de mots à chercher obligeant l'ouverture de *tous* les volumes. La réponse est $15H_{15}$ soit 50 mots.
Le problème 2) est celui de l'écart-type du nombre précédent. La réponse est $\boxed{15\sqrt{\sum_{k=1}^{15}\frac{1-k/15}{k^2}}}$ soit 18 mots environ.

Et le problème 3) celui de la probabilité d'ouverture de tous les volumes dans une recherche de $N$ mots. Pour $N = 15$ on trouve toujours une probabilité très faible de $\frac{15!}{15^{15}}$ soit 0,0003.

Je n'avais pas vu le lien entre les deux problèmes, car, au départ, dans le cas de la collection, je ne m'étais posé que les questions 1) et 2), et dans le cas du dictionnaire, que la question 4)…

Signalons que ce problème très riche du collectionneur est aussi traité dans « Gilles Pagès, Claude Bouzitat, En passant par hasard…, Vuibert, 1999 », avec d'autres prolongements.