Manuscrit auteur, publié dans "Images des Mathématiques (2010) http://images.math.cnrs.fr/Petite-collection-d-informations.html"

hal-00586348, version 1 - 15 Apr 2011

Petite collection d'informations utiles pour collectionneur compulsif

To view high-resolution math symbols, click the Hi-Res Fonts for Printing button on the jsMath control panel.

Le 11 août 2010, par **Sylvain Sardy** et **Yvan Velenik**

Collectionner les vignettes autocollantes vendues à l'occasion de la coupe du monde de football ou les cartes Pokémon : voici des activités onéreuses auxquelles s'adonnent avec joie de nombreux enfants (et certains grands enfants !). Nous discutons ici de certains aspects mathématiques qu'un collectionneur avisé devrait connaître avant de débuter une collection.

Introduction

UNE stratégie marketing bien connue et fort utilisée consiste à incorporer à chaque unité achetée d'un certain produit (boîte de céréales, etc.) un petit objet (vignette, magnet, figurine, etc.) extrait d'une collection de n objets différents. Le désir de compléter la collection encourage alors l'acheteur à continuer à se procurer ce type de produit chez la même marque. On peut même aller plus loin dans cette logique, et supprimer le produit sous-jacent pour ne conserver que les objets à collectionner, comme l'ont découvert avec profit de nombreuses compagnies : les vignettes autocollantes associées à des événements sportifs, des films, etc., de la compagnie Panini, les cartes à collectionner Pokémon, Magic, etc., et bien d'autres encore. L'étude mathématique, principalement à l'aide d'outils probabilistes et statistiques, de ce type de problèmes est bien connue, et possède également de nombreuses applications très concrètes en ingénierie, par exemple dans des problèmes de télécommunication [1].

Modélisation

Afin de simplifier l'écriture, nous utiliserons la terminologie correspondant à la collecte de vignettes à coller dans un album. Évidemment, d'autres interprétations sont possibles.

Vignettes et pochettes

Il existe de très nombreuses variantes de ce problème. Dans cet article, nous nous restreindrons aux versions les plus simples. En particulier, nous ferons les hypothèses suivantes :

- l'album à remplir est composé de n vignettes différentes, que nous supposerons numérotées de 1 à n ;
- les vignettes s'achètent par pochette, chaque pochette contenant m vignettes distinctes (le contenu d'une pochette est inconnu au moment de l'achat) ;
- chacun des C_n^m ensembles de m vignettes distinctes apparaît dans les pochettes avec la

même fréquence [2]. Nous verrons plus bas comment s'assurer de la validité de cette hypothèse dans la réalité.

Par exemple, l'album vendu par Panini à l'occasion de la coupe du monde de football 2010 contient, en Suisse, $n = 660$ vignettes différentes, vendues par paquets de $m = 5$.

Échanges

Un autre facteur est important : les vignettes acquises en plusieurs exemplaires peuvent être échangées contre celles d'autres collectionneurs. En effet, l'échange avec les copains dans la cour de récréation est sans doute une des composantes majeures du plaisir (et de la pression sociale) ressenti par les enfants concernés. La modélisation des échanges est cependant plus délicate du point de vue mathématique. En particulier, la description de procédures d'échange « réalistes », c'est-à-dire prenant en compte des facteurs tels que : (i) la variabilité de la « valeur » d'une vignette (en particulier, en fonction de sa rareté au sein du groupe dans lequel a lieu l'échange), (ii) la modification des modalités des échanges lorsqu'un des collectionneurs a complété son album, (iii) la variabilité du nombre de vignettes achetées par chacun des collectionneurs, etc. S'il est effectivement facile d'étudier par des simulations numériques des procédures d'échange de complexité arbitraire, en faire une étude mathématique est beaucoup plus difficile. Pour cette raison, dans cet article nous nous restreindrons à une procédure d'échange assez artificielle, mais qui possède deux avantages : elle permet d'une part d'obtenir certains résultats mathématiques, et d'autre part correspond à une procédure d'échange « optimale ». Dans la procédure d'échange optimale, les q collectionneurs achètent tous le même nombre de pochettes, mettent en commun toutes les vignettes obtenues, et remplissent les albums successivement jusqu'à obtention de q albums complets. Ce qui rend cette procédure particulièrement simple est qu'elle est manifestement équivalente à la situation d'un *unique* collectionneur cherchant à remplir q albums (c'est-à-dire à obtenir au moins q copies de chaque vignette).

Problèmes abordés

Dans cette section, nous décrivons quelques-unes des questions discutées de façon plus précise dans la suite de cet article. Afin de rester le plus concret possible, nous nous restreindrons ici au cas de l'album Panini mentionné plus haut, c'est-à-dire un album de $n = 660$ vignettes, avec $m = 5$ vignettes toutes distinctes par pochette.

Combien une personne seule doit-elle acheter de pochettes pour compléter son album ?

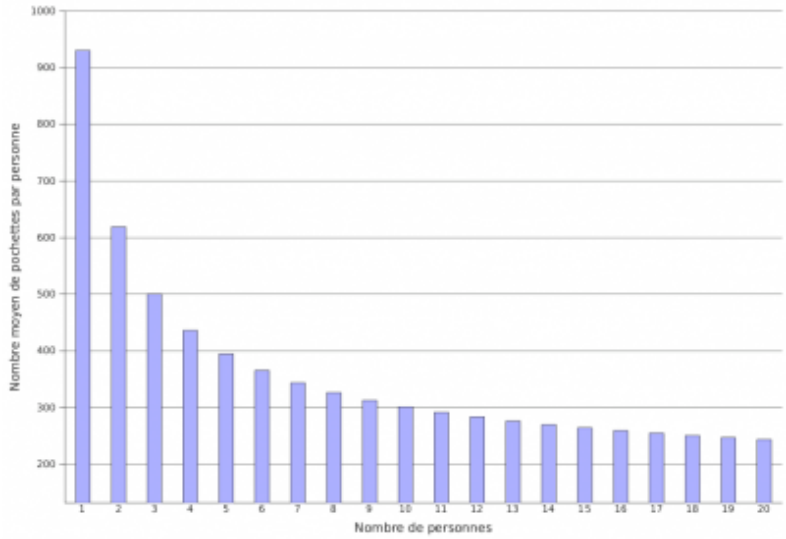
Nous verrons ci-dessous que le nombre de pochettes à acheter, en l'absence d'échanges avec d'autres collectionneurs, croît très rapidement avec le nombre n de vignettes. En particulier, pour $n = 660$, il faut acheter en moyenne 931 pochettes, c'est-à-dire 4655 vignettes ! À un franc suisse la pochette, remplir ainsi un album n'est pas bon marché !

Et en faisant des échanges ?

Si l'on se restreint à la procédure d'échange optimal, le problème peut encore être étudié en

hal-00586348, version 1 - 15 Apr 2011

détails. L'influence des échanges sur le coût d'un album est très importante : pour un groupe de 10 personnes, le nombre de pochettes que chacune des personnes doit acheter en moyenne descend à approximativement 301, au lieu des 931 pochettes nécessaires à un collectionneur isolé.



Certaines vignettes sont-elles plus rares que les autres ?

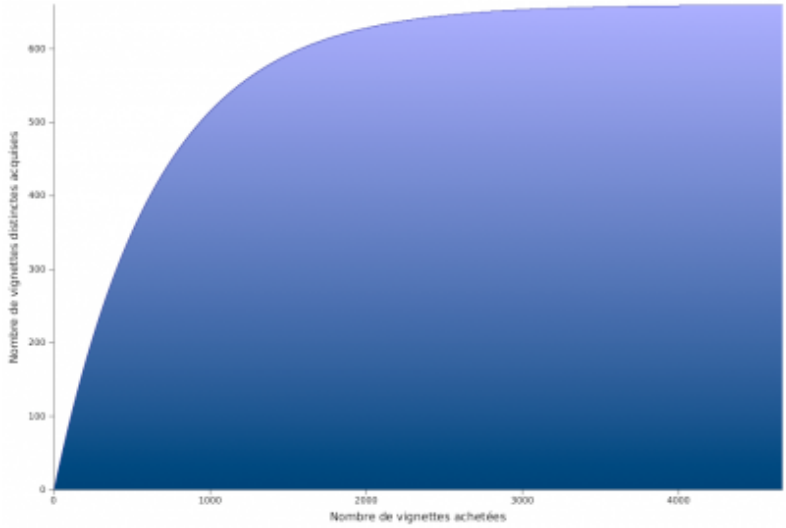
Les gens participant à ce type d'activités ont souvent l'impression que certaines vignettes sont artificiellement rendues plus rares que les autres. Si ceci peut-être vrai dans un certain nombre de cas, il s'agit bien souvent plutôt d'un phénomène psychologique, dû à la très mauvaise intuition des phénomènes aléatoires qu'ont les êtres humains [3]. Nous verrons plus bas quelle procédure peut être utilisée afin de déterminer si cette impression de rareté est justifiée (test du chi-carré pour un échantillon de vignettes collectées).

Quels sont ces effets psychologiques pouvant donner une fausse impression de rareté ? Donnons juste deux exemples, sur lesquels nous reviendrons plus bas :

- Lorsqu'un individu remplit seul son album, les premières pochettes achetées lui donnent très peu de vignettes en double, et son album se remplit donc assez rapidement. À mesure que le collectionneur se rapproche de la complétion de son album, il lui faut acheter de plus en plus de pochettes pour obtenir de nouvelles vignettes. À titre d'exemple, sur les 931 pochettes qu'il lui faudra acheter en moyenne, les premières 232 lui apporteront environ 547 vignettes distinctes, alors que les 232 dernières ne lui fourniront que 3 vignettes ! Il n'est donc pas surprenant que le collectionneur ait l'impression que ces dernières trois vignettes sont quasiment introuvables.
- De même, lorsqu'un groupe de 10 individus pratiquant un échange optimal de leurs vignettes achètent chacun 100 pochettes, soit un total de 5000 vignettes pour le groupe, il y a plus d'une chance sur 4 qu'il manque au moins une même vignette à chaque membre du groupe ! Comme avant, une telle vignette manquant à tout le monde leur donnera l'impression d'être anormalement rare.

hal-00586348, version 1 - 15 Apr 2011

hal-00586348, version 1 - 15 Apr 2011



Combien coûte une collection complète ?

1 *collectionneur,* 1 *vignette par pochette*

Considérons tout d'abord le cas $m = 1$, c'est-à-dire la situation où les vignettes sont acquises une par une (dans une pochette close). Ce problème classique de théorie des probabilités est connu sous le nom de « problème du collecteur de coupons ». Son étude, ainsi que celle de ses diverses extensions, a donné lieu à de nombreux travaux en mathématiques. Une première question très naturelle est de déterminer le nombre moyen de pochettes qu'il faut acheter afin de compléter l'album (ou, ce qui est équivalent, le coût total, en moyenne, de la collection).

Avant de faire cela, faisons une brève digression. Combien de fois *en moyenne* faut-il lancer un dé à 6 faces jusqu'à obtention d'un 6 ? La probabilité d'obtenir un 6 étant de 1/6 à chaque lancer, l'intuition nous dit que le nombre moyen de lancers nécessaires sera de 6, ce qui est correct (voir plus bas). Plus généralement, si l'on cherche à réaliser un événement de probabilité p, le nombre moyen de tentatives nécessaires est de $1/p$.

Revenons à nos vignettes, et déterminons la probabilité d'obtenir une nouvelle vignette lors d'un achat.

- Il est clair que la première vignette achetée peut être immédiatement collée dans l'album.
- La deuxième vignette achetée peut, par contre, coïncider avec la première ; ceci a lieu avec probabilité $1/n$, ou, ce qui est équivalent, la probabilité d'obtenir une nouvelle vignette est égale à $(n-1)/n$.
- Plus généralement, si k vignettes ont déjà été collées dans l'album, la probabilité d'obtenir une nouvelle vignette à l'achat suivant est de $(n-k)/n$.
- Pour finir, lorsque l'album contient toutes les vignettes sauf une, la probabilité d'obtenir cette dernière vignette est égale à $1/n$, ce qui est très faible lorsque n est grand. Nous verrons plus loin qu'il s'agit d'une des causes de l'impression de rareté de certaines vignettes.

hal-00586348, version 1 - 15 Apr 2011

Combien de vignettes faudra-t-il donc acheter *en moyenne* pour que la collection passe de $k-1$ à k vignettes ? On a vu que la probabilité d'obtenir la $k^{\text{ème}}$ vignette est de $(n-k+1)/n$. Le nombre moyen de vignettes à acheter sera donc de $n/(n-k+1)$.

Le nombre total de vignettes à acheter pour compléter l'album est donc donné, en moyenne, par

$$1+\frac{n}{n-1}+\frac{n}{n-2}+\cdots+\frac{n}{2}+\frac{n}{1}\simeq n(\log n+0,577).$$

Plus précisément...

Le nombre N_k de pochettes à acheter pour faire passer la collection de $k-1$ à k vignettes suit une loi géométrique de paramètre $p_{n,k}=(n-k+1)/n$, c'est-à-dire [4]

$$\mathbf{Prob}(N_k=\ell)=\left(1-p_{n,k}\right)^{\ell-1}p_{n,k}\qquad(\ell\in\mathbb{N}^*).$$

*En particulier, l'**espérance** de N_k est donnée par*

$$\mathbf{E}(N_k)=\sum_{\ell=1}^{\infty}\ell\left(1-p_{n,k}\right)^{\ell-1}p_{n,k}=1/p_{n,k}.$$

On peut à présent aisément déterminer le nombre moyen de pochettes qu'il faut acheter afin de compléter l'album : en effet, le nombre N d'achats nécessaires satisfait $N=N_1+N_2+\cdots+N_n$. Par linéarité de l'espérance, on obtient donc

$$\mathbf{E}(N)=\mathbf{E}(N_1)+\cdots+\mathbf{E}(N_n)=\sum_{k=1}^{n}\frac{1}{p_{n,k}}=\sum_{k=1}^{n}\frac{n}{n-k+1}=nH_n,$$

*où H_n est le $n^{\text{ème}}$ **nombre harmonique** : $H_n=1+(1/2)+\cdots+(1/n)=\log n+\gamma+O(\frac{1}{n})$, et $\gamma\simeq 0,577$ est la **constante d'Euler-Mascheroni**.*

Ainsi, il faudra acheter en moyenne environ 4666 pochettes (contenant chacune une vignette) pour compléter un album de 660 vignettes.

1 collectionneur, m vignettes par pochette

Le cas plus général où chaque pochette contient m vignettes distinctes est un peu plus compliqué. Dans ce cas, le nombre moyen de pochettes à acheter est égal à

$$C_n^m\sum_{k=1}^{n}(-1)^{k-1}\frac{C_n^k}{C_n^m-C_{n-k}^m}.$$

Dérivation de la formule.

*Notons, comme précédemment, N le nombre de pochettes à acheter. L'approche utilisée ci-dessus, reposant sur la décomposition $N=N_1+\cdots+N_n$, ne se prête pas bien à cette généralisation. Il est toutefois possible de déterminer la loi de N de manière particulièrement simple [5]. On commence par rappeler la **formule d'inclusion-exclusion** : si $A_1,\ldots,A_r$ sont r ensembles, alors*

$$\mathbf{Prob}(A_1 \cup \cdots \cup A_r) = \sum_{k=1}^{r} (-1)^{k-1} \sum_{1 \le i_1 < i_2 < \cdots < i_k \le r} \mathbf{Prob}(A_{i_1} \cap \cdots \cap A_{i_k}).$$

On numérote les pochettes dans l'ordre d'achat, et on note T_k le numéro de la première pochette contenant la $k^{\text{ème}}$ vignette de l'album. On peut alors exprimer le nombre N de pochettes à acheter comme $N = \max\{T_1, \ldots, T_n\}$.

Commençons par déterminer la probabilité que N soit supérieur à ℓ. Observons tout d'abord que $\max\{T_1, \ldots, T_n\} > \ell$ si et seulement si il existe $k \in \{1, \ldots, n\}$ tel que $T_k > \ell$. De façon similaire, pour $1 \le i_1 < i_2 < \cdots < i_k \le n$, $\min\{T_{i_1}, \ldots, T_{i_k}\} > \ell$ si et seulement si $T_{i_j} > \ell$ pour tout $j \in \{1, \ldots, k\}$.

En appliquant la formule ci-dessus avec $A_k = \{T_k > \ell\}$, on obtient alors

$$\begin{aligned}
\mathbf{Prob}(N > \ell) &= \mathbf{Prob}(A_1 \cup \cdots \cup A_n) \\
&= \textstyle\sum_{k=1}^{n} (-1)^{k-1} \sum_{1 \le i_1 < i_2 < \cdots < i_k \le n} \mathbf{Prob}(A_{i_1} \cap \cdots \cap A_{i_k}) \\
&= \textstyle\sum_{k=1}^{n} (-1)^{k-1} \sum_{1 \le i_1 < i_2 < \cdots < i_k \le n} \mathbf{Prob}(\min\{T_{i_1}, \ldots, T_{i_k}\} > \ell).
\end{aligned}$$

L'intérêt de cette opération est que la probabilité dans le membre de droite se calcule facilement ! En effet, l'événement n'est réalisé que si aucune des vignettes $i_1, \ldots, i_k$ ne se trouve dans une des ℓ premières pochettes achetées. La probabilité qu'une pochette donnée ne contienne aucune des ces k vignettes est C_{n-k}^m / C_n^m. On obtient donc, puisque les contenus des différentes pochettes sont indépendants,

$$\mathbf{Prob}(\min\{T_{i_1}, \ldots, T_{i_k}\} > \ell) = \left(\frac{C_{n-k}^m}{C_n^m} \right)^{\ell},$$

d'où l'on tire aisément la loi de N :

$$\begin{aligned}
\mathbf{Prob}(N > \ell) &= \textstyle\sum_{k=1}^{n} (-1)^{k-1} \sum_{1 \le i_1 < i_2 < \cdots < i_k \le n} \left(\frac{C_{n-k}^m}{C_n^m} \right)^{\ell} \\
&= \textstyle\sum_{k=1}^{n} (-1)^{k-1} C_n^k \left(\frac{C_{n-k}^m}{C_n^m} \right)^{\ell}.
\end{aligned}$$

On peut à présent facilement en extraire l'espérance :

$$\mathbf{E}(N) = \sum_{\ell=0}^{\infty} \mathbf{Prob}(N > \ell) = C_n^m \sum_{k=1}^{n} (-1)^{k-1} \frac{C_n^k}{C_n^m - C_{n-k}^m}.$$

En particulier, il faudra acheter en moyenne environ 931 pochettes contenant chacune 5 vignettes distinctes, soit un total d'environ 4655 vignettes, pour compléter un album de 660 vignettes. Observer que ce résultat est très proche de celui obtenu pour $m = 1$ (4666 vignettes). Ceci s'explique facilement par le fait que la probabilité que 5 vignettes achetées successivement (lorsque $m = 1$) soient toutes différentes est égale à $(660)_5 / 660^5 \simeq 0,985$, où l'on a utilisé la notation $(n)_r = n(n-1) \cdots (n-r+1)$. L'absence de doubles dans une pochette de 5 vignettes n'affecte donc en moyenne qu'environ $1,5\%$ des pochettes.

q collectionneurs, 1 vignette par pochette

Un autre problème, plus délicat, consiste à tenir compte des échanges possibles entre q individus. Nous nous restreindrons au cas le plus simple à traiter mathématiquement : l'échange optimal,

hal-00586348, version 1 - 15 Apr 2011

décrit plus haut. Dans ce cas, le nombre de pochettes à acheter pour que chacun des q individus parvienne à compléter son album est identique au nombre de pochettes qu'un unique individu doit acheter pour compléter q albums. Ce problème a également été étudié (toujours dans le cas $m = 1$). Malheureusement, les formules explicites deviennent vite très complexes. On peut toutefois obtenir des informations sur le comportement asymptotique, c'est-à-dire lorsque $n \longrightarrow \infty$ en maintenant le nombre q de personnes fixé. Par exemple [6], le nombre moyen de pochettes à acheter est approximativement égal à

$$n \log n + (q-1) n \log \log n.$$

Si la qualité de cette approximation est assez mauvaise lorsque n est petit, celle-ci devient très bonne lorsque n est grand.

On voit que lorsque n est (très) grand, le nombre moyen d'achats nécessaires ne dépend que très faiblement du nombre q d'individus ($\log \log n$ étant beaucoup plus petit que $\log n$). Ceci peut se comprendre facilement en réalisant que lorsque n est grand, au moment où le premier album est complet, la plupart des vignettes auront déjà été obtenues en de nombreux exemplaires. En fait, on peut être plus précis [6] : le nombre moyen de coupons dont on ne possède qu'un unique exemplaire au moment où l'album est complété est égal au $n^{\text{ème}}$ nombre harmonique H_n.

L'étude du cas $m > 1$ produit des expressions encore plus compliquées, et nous ne la discuterons pas ici. Comme précédemment, lorsque m est beaucoup plus petit que n, la différence avec le cas $m = 1$ est de toute façon relativement faible. La figure 1 montre la dépendance du nombre de pochettes à acheter, en moyenne, en fonction du nombre de personnes dans le groupe.

Rareté orchestrée ou apparente ?

Parmi les collectionneurs de vignettes, il est souvent supputé que certaines sont intentionnellement rares, afin de rendre plus difficile la complétion d'un album et par conséquent de la rendre plus onéreuse. On se propose ici de tester statistiquement et de donner une explication probabiliste à ce sentiment de rareté.

Test du chi-deux

Comment peut-on vérifier si certaines vignettes sont effectivement plus rares que d'autres ? La procédure est semblable à celle utilisée dans les sondages : comme l'on n'a pas accès à l'ensemble des vignettes émises par le fabriquant, il est nécessaire de baser l'analyse sur un échantillon de vignettes. On achète donc un certain nombre de vignettes, et on comptabilise les fréquences d'apparition des différentes vignettes ; on parle alors de fréquences empiriques. Le problème est de déterminer si ces fréquences sont compatibles avec l'hypothèse d'une équiprobabilité théorique de l'apparition de chaque vignette (c'est-à-dire d'absence de vignettes rares). Afin de réaliser cette tâche, un statisticien aura recours au test du chi-deux (ou chi-carré), que nous allons à présent décrire [7].

Commençons par expliquer en quelques mots la procédure d'un test statistique. Comme toujours en sciences, il est impossible de démontrer qu'une théorie, ou une hypothèse, est correcte, mais seulement de l'invalider par des observations adéquates. Nous partons donc d'une hypothèse, que l'on appelle **hypothèse nulle** et que l'on note H_0, que l'on fait sur une population donnée. On

hal-00586348, version 1 - 15 Apr 2011

hal-00586348, version 1 - 15 Apr 2011

désire, sur la base d'un échantillon collecté dans cette population, déterminer s'il y a des raisons de rejeter notre hypothèse H_0. Supposons qu'il existe une propriété D_α qui, si l'hypothèse H_0 est vraie, ne sera vérifiée par un échantillon qu'avec une très faible probabilité α (disons $\alpha = 5\%$). On effectue alors la procédure suivante :

- on collecte l'échantillon ;
- on détermine si l'échantillon satisfait la condition D_α ;
- si c'est le cas, c'est une forte indication que notre hypothèse H_0 est fausse, et on rejettera donc cette hypothèse.
- sinon, l'observation est compatible avec l'hypothèse H_0, que l'on pourra donc conserver.

Une telle propriété D_α, calculée à partir des données, est appelée une *statistique*.

Retournons à nos vignettes. Dans ce problème, l'hypothèse nulle H_0 est la suivante :

$$H_0 : \text{toutes les vignettes ont la même fréquence } 1/n.$$

Évidemment, l'échantillon étant de taille faible par rapport à l'ensemble des vignettes émises, on ne s'attend pas à ce que les fréquences empiriques d'apparition de chaque vignette coïncident avec les fréquences théoriques, ici toutes égales à $1/n$. Cependant, il suit de la **loi des grands nombres** que l'écart entre les fréquences empiriques et les fréquences théoriques devrait tendre vers 0 lorsque la taille t de l'échantillon croît, si l'hypothèse nulle est vérifiée. Il est donc naturel de construire la propriété D_α en se basant sur une comparaison quantitative entre les fréquences empiriques et les fréquences théoriques. Pour cela, on utilise la statistique de Pearson [8], qui, pour un échantillon donné de numéros de vignettes, se calcule de la façon suivante [9] :

$$P = \sum_{i=1}^{n} \frac{(O_i - e_i)^2}{e_i},$$

où O_i est le nombre d'apparitions de la vignette i dans l'échantillon et $e_i = t/n$ est le nombre d'apparitions théorique si l'hypothèse nulle est vraie. On se fixe ensuite un seuil α comme ci-dessus, et on définit la propriété D_α par

$$D_\alpha : \; P^{\text{calc}} \geq C,$$

où P^{calc} est la valeur que prend P pour notre échantillon, et C est choisie de sorte que la probabilité de D_α, si l'hypothèse nulle est vraie, soit égale à α. Malheureusement, calculer analytiquement la valeur de C pour une valeur de α donnée est en général compliqué. Toutefois, une propriété intéressante de la statistique de Pearson P est que, si l'on suppose que (i) H_0 est vraie, (ii) la taille t de l'échantillon est suffisamment grande, (iii) le nombre moyen e_i de copies de la $i^{\text{ème}}$ vignette prédit sous l'hypothèse H_0 satisfait $e_i \geq 5$ pour $i = 1, \ldots, n$, et (iv) les vignettes sont achetées une à une (c'est-à-dire dans le cas $m = 1$ avec les notations introduites précédemment), alors la statistique de Pearson satisfait

$$P \approx \chi^2_{n-1},$$

c'est-à-dire que P suit approximativement une **distribution chi-deux avec $n-1$ degrés de**

hal-00586348, version 1 - 15 Apr 2011

liberté. On peut donc, lorsque l'approximation est bonne, déterminer C en utilisant la distribution χ^2_{n-1} au lieu de P, ce qui est beaucoup plus simple.

L'approximation (notée par $\approx$ au lieu $\sim$ qui dénote l'égalité en loi) de la distribution de P par χ^2_{n-1} est d'autant meilleure que t est grand et que les e_i sont grands (avec cinq comme borne inférieure pour une bonne approximation).

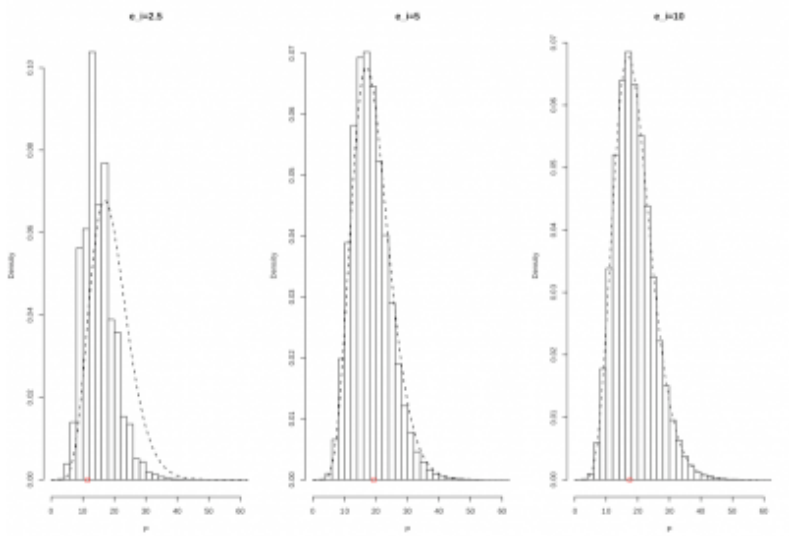
Observez finalement que $P \geq C$ si et seulement si $\mathbf{Prob}(P \geq P^{\text{calc}}) \leq \alpha$. On appelle $\mathbf{Prob}(P \geq P^{\text{calc}})$ la « **valeur-p** » [10]. Ainsi avec par exemple $\alpha = 0,05$, la probabilité de rejeter l'hypothèse nulle alors que celle-ci est vraie (ce que l'on appelle faire une erreur de type I) est de 5%.

Nous illustrons ce test du chi-deux avec un échantillon de taille $t = 200$ d'une collection de $n = 20$ vignettes. Les comptages o_i de cet échantillon sont reportés dans le Tableau 1. Un calcul simple donne $P^{\text{calc}} = 17,4$ pour cet échantillon, et la valeur-p correspondante est $0,563$ en utilisant l'approximation du χ^2_{19} [11]. Par conséquent, on ne peut pas rejeter ici l'hypothèse nulle au niveau de signification $\alpha = 0,05$.

Tableau 1 : Données pour $m = 1, n = 20, t = 200$: $P^{\text{calc}} = 17,4$

i	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
o_i	4	9	11	15	15	6	9	9	7	14	9	7	7	11	11	9	10	11	12	14

Avec $t = 200$, $e_i = 200/20 = 10$ et $m = 1$, l'approximation de la distribution de la statisique de Pearson P sous l'hypothèse nulle par une chi-deux à 19 degrés de liberté est bonne car t est grand et $e_i > 5$. Pour le confirmer, une simulation Monte-Carlo utilisant 10^5 répétitions permet d'estimer la fonction de densité de P (avec un histogramme par exemple) sous l'hypothèse nulle et de la comparer avec la fonction de densité χ^2_{19}. Le graphique de droite de la Figure 3 montre une très bonne concordance entre la densité estimée par Monte-Carlo (l'histogramme) et la densité d'une χ^2_{19} (ligne pointillée) ; en particulier la valeur-p estimée par Monte-Carlo est de $0,555$ (à comparer au $0,563$ obtenu ci-dessus par approximation).



hal-00586348, version 1 - 15 Apr 2011

Que faire si t faible et $e_i \leq 5$?

Dans l'application du test du chi-deux ci-dessus, les conditions du test étaient satisfaites. Dans cette section, on considère la situation où la taille de l'échantillon est faible et les espérances e_i sont inférieures à cinq.

Avec $t = 100$, $e_i = 100/20 = 5$ et $m = 1$, les comptages o_i d'un échantillon simulé sous l'hypothèse nulle sont reportés dans le Tableau 2. Vu que la limite de cinq pour les e_i est atteinte, l'approximation de la distribution de la statistique de Pearson P sous l'hypothèse nulle par une chi-deux à 19 degrés de liberté devrait encore être tolérable. Le résultat de la simulation Monte-Carlo dans cette situation est représenté sur le graphique du milieu de la Figure 3 : on observe un léger décalage à gauche de la vraie distribution de P par rapport à une chi-deux. Pour cet échantillon $P^{\text{calc}} = 14,2$, et la valeur-p approximée par chi deux est légèrement supérieure à celle de référence estimée par Monte-Carlo (voir Tableau 4).

Tableau 2 : Données pour $m = 1$, $n = 20$, $t = 100$: $P^{\text{calc}} = 19,2$

i	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
o_i	3	5	2	4	11	5	6	5	5	5	6	5	6	6	5	2	7	7	5	0

Finalement les comptages o_i d'un échantillon simulé sous l'hypothèse nulle sont reportés dans le Tableau 3 pour la situation plus extrême où $t = 50$, $e_i = 50/20 = 5$ et $m = 1$: d'une part t est faible, et d'autres part $e_i < 5$. Le résultat de la simulation Monte-Carlo dans cette situation est représenté sur le graphique gauche de la Figure 3 : on observe un fort décalage à gauche de la vraie distribution de P par rapport à une chi-deux. De plus $P^{\text{calc}} = 28,3$ pour cet échantillon, et la valeur-p approximée par chi-deux est fortement supérieure à celle de référence estimée par Monte-Carlo (voir Tableau 4).

Tableau 3 : Données pour $m = 1$, $n = 20$, $t = 50$: $P^{\text{calc}} = 11,6$

i	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
o_i	3	1	4	2	2	4	1	1	1	3	3	5	2	2	4	4	2	3	2	1

Tableau 4 : Valeurs-p calculées par approximation du chi-deux comparées à celles estimées par Monte-Carlo sous l'hypothèse nulle pour $m = 1$, $n = 20$ et $t = 50, 100, 200$ (c'est-à-dire $e_i = 2,5; 5; 10$)

e_i	2.5	5	10
Approximation chi-deux	0,902	0,444	0,563
Estimation Monte Carlo	0,733	0,394	0,555

Que faire si $m \geq 1$?

On considère maintenant le cas où le collectionneur achète des pochettes contenant m vignettes toutes différentes avec $m > 1$, contrairement à ce qui se produisait dans les deux sections précédentes (où $m = 1$), la condition $m > 1$ crée de la dépendance entre les vignettes collectées, ce qui perturbe également le test du chi-deux, comme on peut le voir sur la figure 4.

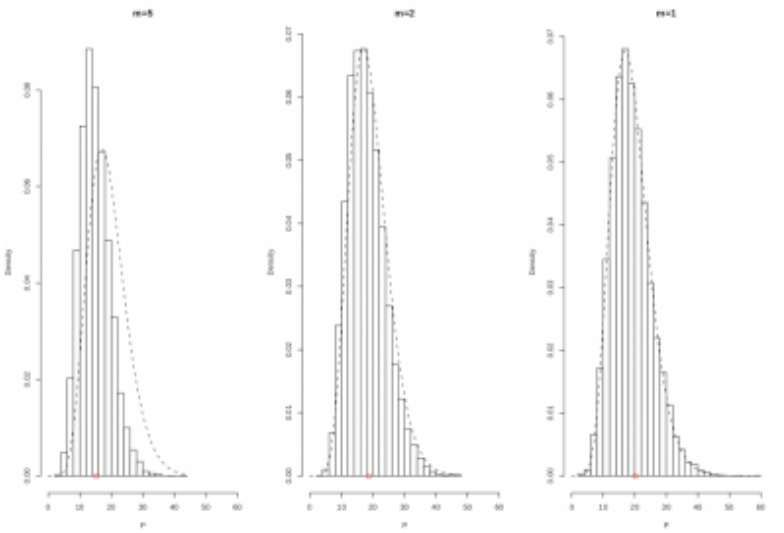
Tableau 5 : Données pour $m = 2$, $n = 20$, $t = 200$: $P^{\text{calc}} = 18,6$

i	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
o_i	10	6	8	8	6	8	14	8	15	6	9	10	12	7	11	11	10	17	14	10

Tableau 6 : Données pour $m = 5$, $n = 20$, $t = 200$: $P^{\text{calc}} = 15,4$

i	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
o_i	8	10	8	10	12	9	6	8	4	11	10	9	14	11	14	14	10	10	7	15

hal-00586348, version 1 - 15 Apr 2011



En conclusion, il est important de vérifier les hypothèses du test du chi-deux. Quand celles-ci sont violées, l'approximation du chi-deux est mauvaise, mais une simulation Monte-Carlo permet d'estimer la valeur-p du test.

Explication « psychologique »

L'impression de rareté de certaines vignettes subsiste souvent même lorsque celles-ci sont distribuées de façon équiprobable. L'explication est alors plutôt de nature psychologique, et suit d'une mauvaise intuition du hasard par l'être humain.

Considérons tout d'abord le cas d'un individu collectant seul des vignettes vendues à l'unité ($m = 1$, $n = 660$). Dans ce cas, comme on l'a vu plus haut, il lui sera nécessaire d'acheter en moyenne 4666 vignettes. Cependant, la vitesse à laquelle les 660 vignettes de l'album sont acquises décroît très rapidement :

- les premières 1166 vignettes achetées fourniront approximativement 547 vignettes distinctes (zone bleue dans la figure 5 ci-dessous) ;
- les 1166 vignettes suivantes fourniront approximativement 94 vignettes nouvelles (zone violette) ;
- les 1166 vignettes suivantes n'en fourniront qu'environ 16 (zone jaune) ;
- il faudra acheter approximativement 1168 vignettes supplémentaires pour obtenir les 3 vignettes manquantes (zone verte).

Sous une forme plus visuelle (les blocs de couleurs représentant les contributions indiquées ci-dessus à l'album final) :

Plus de détails sur la dérivation

Le calcul effectué plus haut montre en effet que, sous l'hypothèse d'équiprobabilité, le nombre de vignettes à acheter pour obtenir r vignettes distinctes est égal en moyenne à

$$\mathbf{E}(N_1) + \cdots + \mathbf{E}(N_r) = \sum_{k=1}^{r} \frac{1}{p_{n,k}} = \sum_{k=1}^{r} \frac{n}{n-k+1}.$$

Cet exemple explique pourquoi l'individu aura l'impression que les dernières vignettes dont il a besoin sont beaucoup plus rares que les autres (il aura en effet énormément de difficulté à les obtenir).

Une seconde situation, plus réaliste, est celle d'un groupe de 10 individus procédant à un échange optimal de leurs vignettes (toujours avec $n = 660$ et $m = 1$). Dans ce cas, même après l'achat de 500 vignettes par chacun d'entre eux, la probabilité qu'une des vignettes au moins leur manque simultanément *à tous* est supérieure à 28%. Ce groupe aura évidemment l'impression que les vignettes manquantes sont très rares.

Détails mathématiques

En appliquant la formule dérivée plus haut avec $n = 660$, $m = 1$ *et* $\ell = 5000$, *on obtient*

$$\mathbf{Prob}(N > 5000) = \sum_{k=1}^{660} (-1)^{k-1} C_{660}^{k} \frac{(660-k)^{\ell}}{660^{\ell}} \simeq 0,28.$$

P.S. :

*Les auteurs remercient **Pierre de la Harpe** pour tous ses commentaires, ainsi que **Vincent Beffara** pour son aide avec la conversion de latex vers SPIP.*

Notes

[▲1] Voir par exemple : A. Boneh et M. Hofri, *The coupon-collector problem revisited---a survey of engineering problems and computational methods*, Comm. Statist. Stochastic Models **13** (1997), no. 1, 39—66.

[▲2] Le coefficient binomial $C_n^m = \frac{n!}{(n-m)!m!}$ est égal au nombre de façons d'extraire m éléments distincts d'un ensemble de n éléments.

[▲3] Voir par exemple : P. Diaconis et F. Mosteller, *Methods for studying coincidences*, J. Amer. Statist. Assoc. **84** (1989), no. 408, 853—861.

[▲4] On utilise la convention usuelle $0^0 = 1$.

[▲5] L'argument présenté est tiré de : I. Adler, S. Oren et S.M. Ross, *The coupon-collector's problem revisited*, J. Appl. Probab. **40** (2003), no. 2, 513—518.

[▲6] A.N. Myers et H.S. Wilf, *Some new aspects of the coupon collector's problem*, SIAM J. Discrete Math. **17** (2003), no. 1, 1—17.

[▲7] Pour une application à des données réelles, voir par exemple : S. Sardy et Y. Velenik, *Paninimania : sticker rarity and cost-effective strategy*, Swiss Statistical Society, Bulletin nr. **6** (2010), 2—6

hal-00586348, version 1 - 15 Apr 2011

[▲8] **Karl Pearson**, 1857—1936 : mathématicien britannique et l'un des fondateurs des statistiques modernes. Il introduisit le test du chi-carré dans un article en 1900.

[▲9] Observez que P est une distance pondérée entre les O_i et les e_i.

[▲10] Le p fait référence à probabilité, pas à Pearson.

[▲11] Vous pouvez refaire par vous-même ce calcul à l'aide de l'applet disponible sur le site **http://www.stat.tamu.edu/ west/appl...**.

Crédits images

Pour citer cet article : **Sylvain Sardy** et **Yvan Velenik**, **Petite collection d'informations utiles pour collectionneur compulsif**. *Images des Mathématiques*, CNRS, 2010. En ligne, URL : `http://images.math.cnrs.fr/Petite-collection-d-informations.html`

hal-00586348, version 1 - 15 Apr 2011