Le problème du "Coupon collector" Rappels de probabilité

Remplissage d'une table de hachage

Jean-Marc.Vincent@imag.fr

Université Grenoble-Alpes

Grenoble 11 février 2018

LE PROBLÈME DU "COUPON COLLECTOR"

Soit une table de hachage de taille M (hachage ouvert) et n objets à hacher dans la table.

Le problème des *Panini*

Votre petit cousin collectionne les vignettes autocollantes des footballeurs célèbres. Il achète les vignettes à l'unité à 0.10 euros dans des pochettes opaques et cherche à compléter son album en collectionnant les $M = 300$ vignettes distinctes de son album. On suppose que l'éditeur des vignettes est honnête et qu'il imprime les vignettes dans les mêmes proportions (il n'y a pas de vignettes plus rares que d'autres et chaque pochette achetée contient une des 300 vignettes au hasard).

Notations

- $U_1, U_2, U_3, \cdots, U_n, \cdots$ la séquence modélisant la suite des objets à insérer dans la table, $U_k \in \{1, \cdots, M\}$, U_k sera l'adresse où sera haché l'objet k.

$U_1, U_2, U_3, \cdots, U_n, \cdots$, sont des variables aléatoires, c'est à dire qu'elles prendront des valeurs lors de l'exécution

TABLE DE HACHAGE : QUESTIONS

Point de vue utilisateur

- Coût de la recherche d'un objet (fructueuse ou infructueuse)
- Coût d'une insertion/suppression

Point de vue système

- Taux de remplissage de la table $\alpha = \frac{n}{M}$
- Taux de cases vides

HYPOTHÈSES STATISTIQUES

Uniformité

On suppose que la fonction de hachage est parfaite, c'est à dire que le calcul de la case d'insertion d'un objet est modélisé par un tirage aléatoire de loi **uniforme** parmi toutes les valeurs possibles, ici dans $\{1, \cdots, M\}$.

$$\text{Pour tout } k \in \{1, \cdots, M\} \text{ on a } \mathbb{P}(U_i = k) = \frac{1}{M}.$$

Indépendance

On suppose que la fonction de hachage "mélange" bien les objets, c'est à dire que les tirages modélisant les choix de cases pour les objets sont **indépendants**. Dans ce cas, la probabilité de collision entre 2 objets sera $\mathbb{P}(U_i = U_j) = \frac{1}{M}$.

La suite $\{U_n\}_{n\in\mathbb{N}}$ est donc une suite de variables aléatoires indépendantes de même loi uniforme sur $\{1, \cdots, M\}$

UFR IM²AG
UNIVERSITÉ Grenoble Alpes

LOI DE BERNOULLI

Considérons un jet de pièce de monnaie dont on observe le résultat "pile ou face". La pièce est dite biaisée si la proportion d'observation de pile sur un grand nombre de lancers n'est pas $\frac{1}{2}$. On modélise ce jet de pièce par une variable aléatoire X par convention "pile" est représenté par 1 et "face" par 0.

$$X = \begin{cases} 1 & \text{si la pièce tombe sur "pile"} \\ 0 & \text{sinon, c'est à dire la pièce tombe sup "face"} \end{cases}$$

On note

$$p = \mathbb{P}(X = 1) = 1 - \mathbb{P}(X = 0).$$

La loi de X est appelée **loi de Bernoulli de paramètre** p et est notée $\mathcal{B}(p)$.

LOI DE BERNOULLI (2)

- La **loi** d'une variable aléatoire X de loi de Bernoulli $\mathcal{B}(p)$ est

  $$\mathbb{P}(X = a) = p^a(1-p)^{1-a} \quad \text{pour } a = 0 \text{ ou } a = 1$$

- Sa **moyenne**

  $$\mathbb{E}X = p.$$

  Ce qui s'interprète en terme de fréquence : dans une suite de lancers de la pièce la fréquence d'apparition de "pile" est p.

- La **variance** d'une variable aléatoire de loi de Bernoulli est

  $$\mathbb{V}ar\, X = \mathbb{E}(X - \mathbb{E}X)^2 = p(1-p).$$

  Refaire le calcul à partir de la loi. On observe que pour $p = 0$ ou $p = 1$ la pièce tombe toujours sur le même coté et la variance est nulle. La variance est maximale pour $p = \frac{1}{2}$, lorsque la pièce n'est pas biaisée.

UFR IM²AG
UNIVERSITÉ Grenoble Alpes

SCHÉMA DE BERNOULLI

Considérons un jeu de "pile/face" biaisé modélisé par une suite de variables aléatoires $X_1, \cdots, X_n, \cdots$ à valeur dans $\{0, 1\}$, indépendantes de même loi de Bernoulli $\mathcal{B}(p)$ avec $0 \leqslant p \leqslant 1$, c'est à dire

$$\mathbb{P}(X_i = 1) = p, \quad \mathbb{P}(X_i = 0) = 1 - p$$

Un tel modèle s'appelle ***schéma de Bernoulli***.
Pour un n donné, calculons la probabilité d'observer un vecteur $[a_1, \cdots, a_n]$ avec $a_i \in \{0, 1\}$

$$\mathbb{P}\left([X_1, X_2, \cdots, X_n] = [a_1, \cdots, a_n]\right)$$

$$\begin{aligned}
&= \mathbb{P}(X_1 = a_1, X_2 = a_2, \cdots, X_n = a_n) \\
&= \mathbb{P}(X_1 = a_1)\mathbb{P}(X_2 = a_2) \cdots \mathbb{P}(X_n = a_n) \\
&\quad \text{car les variables } X_i \text{ sont indépendantes,} \\
&= p^{a_1}(1-p)^{1-a_1} \times p^{a_2}(1-p)^{1-a_2} \times \cdots \times p^{a_n}(1-p)^{1-a_n} \\
&\quad \text{car les } X_i \text{ont la même loi } \mathcal{B}(p) \\
&= p^{\sum_i a_i}(1-p)^{n-\sum_i a_i}
\end{aligned}$$

LA LOI BINOMIALE (1)

Considérons un jeu de n "pile/face" biaisé modélisé par une suite de n variables aléatoires $X_1, \cdots, X_n$ à valeur dans $\{0, 1\}$, indépendantes de même loi de Bernoulli $\mathcal{B}(p)$ avec $0 \leqslant p \leqslant 1$.
On s'intéresse à la variable S_n qui compte le nombre de "pile", c'est à dire

$$S_n = X_1 + X_2 + \cdots + X_n.$$

On en déduit que

$$\mathbb{P}(S_n = k) = \underbrace{\binom{n}{k}}_{(a)} \underbrace{p^k (1-p)^{n-k}}_{(b)}.$$

(a) nombre de vecteurs de n bits ayant exactement k bits à 1

(b) probabilité d'observer un vecteur donné de n bits ayant k bits à 1

La loi de S_n est appelée loi binomiale et notée $\mathcal{B}in(n, p)$

LA LOI BINOMIALE (2)

- La **loi** d'une variable aléatoire S_n de loi binomiale est

$$\mathbb{P}(S_n = k) = \binom{n}{k} p^k (1-p)^{n-k} \quad \text{pour } k = 0, 1, 2, 3, \cdots, n$$

- Sa **moyenne**

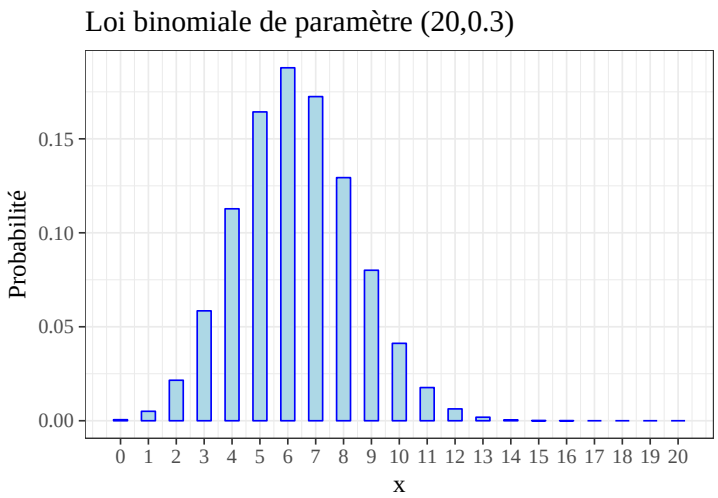
$$\mathbb{E}S_n = np.$$

- Sa **variance**

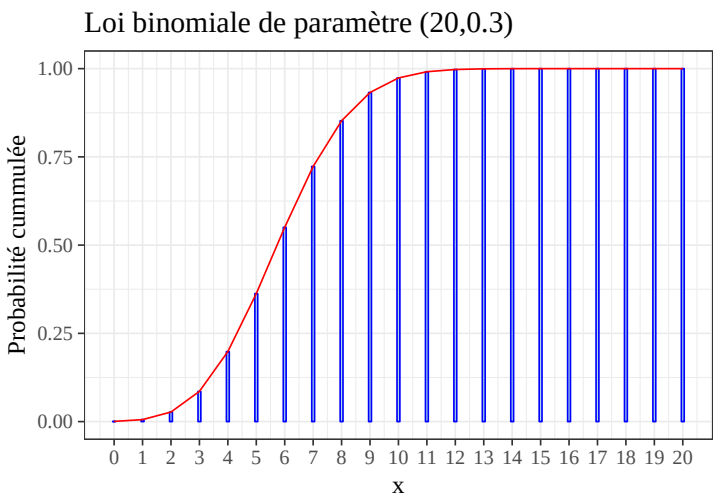
$$\mathbb{V}ar\, S_n = \mathbb{E}(S_n - \mathbb{E}S_n)^2 = np(1-p)$$

LA LOI BINOMIALE (3)

Densité discrète



Fonction de répartition



$$\mathbb{E}S_{20} = 20 \times 0.3 = 6$$

$$\mathbb{V}ar\, S_{20} = 20 \times 0.3 \times (1 - 0.3) = 4.2$$

APPLICATION AUX TABLES DE HACHAGE

Soit X_k le nombre de clés hachées sur la case j. Prenons comme variable auxiliaire $X_{i,k}$ la variable booléenne indiquant si la clé i a été hachée vers la case k.

$$X_j = \sum_{i=1}^{n} X_{i,j}.$$

Or

$$\mathbb{P}(X_{i,j} = 1) = \mathbb{P}(U_i = k) = \frac{1}{M},$$

et pour j fixé les $X_{i,j}$ sont indépendantes (car concernent des clés différentes) de même loi $\mathcal{B}(\frac{1}{M})$.
Donc X_j suit une loi $\mathcal{B}in(n, \frac{1}{M})$.

$$\mathbb{P}(X_j = k) = \binom{n}{k} \frac{1}{M}^k (1 - \frac{1}{M})^{n-k}$$

APPLICATION AUX TABLES DE HACHAGE (2)

On note $\alpha = \frac{n}{M}$ le taux de remplissage de la table.

Recherche infructueuse

Pour rechercher un élément x qui n'est pas dans la table, on calcule son hash et on teste s'il appartient à l'ensemble des éléments hachés sur cette case.
Donc le nombre de tests correspond à X_j si $h(x) = j$
Donc en moyenne $\mathbb{E}X_j = \frac{n}{M} = \alpha = \mathcal{O}(\alpha)$

Recherche d'un élément de la table

On va tester les éléments hachés sur la même case $j = h(x)$, on s'arrête lorsque l'on a trouvé x.
On regarde donc un ensemble d'éléments qui sont associés à j donc en moyenne on a

$$\mathbb{E}(X_j | X_j \geqslant 1) = \frac{\frac{n}{M}}{\mathbb{P}(X_j \geqslant 1)} = \frac{n}{M} \frac{1}{1 - (1 - \frac{1}{M})^M} \simeq \alpha \frac{1}{1 - e^{-1}} \simeq \alpha 1.6 = \mathcal{O}(\alpha)$$

éléments hachés dans la même case j, donc la complexité d'une recherche fructueuse sera de l'ordre de α. [a]

a. en faisant un développement limité on montre facilement que

$$(1 + \frac{a}{x})^x \underset{x \to +\infty}{\longrightarrow} e^a.$$

APPLICATION AUX TABLES DE HACHAGE (3)

Calculons le nombre moyen de cases vides.
Pour cela associons à chaque case j une variable Z_j indiquant si elle est vide ou non

$$Z_j = \begin{cases} 1 & \text{si la case } j \text{ est vide ;} \\ 0 & \text{sinon, c'est à dire qu'au moins un élément à été haché sur } J. \end{cases}$$

Z_j est une variable aléatoire de loi de Bernoulli $\mathcal{B}((1-\frac{1}{M})^n)$, c'est-à-dire

$$\mathbb{P}(Z_j = 1) = (1 - \frac{1}{M})^n$$

Le nombre total de cases vides est $Z = Z_1 + Z_2 + \cdots + Z_M$, d'où

$$\mathbb{E}Z = \mathbb{E}(Z_1 + Z_2 + \cdots + Z_M) = \mathbb{E}Z_1 + \mathbb{E}Z_2 + \cdots + \mathbb{E}Z_M = M(1 - \frac{1}{M})^n.$$

Si on suppose n grand et un taux de remplissage α

$$\mathbb{E}Z = M(1 - \frac{\alpha}{n})^n \simeq Me^{-\alpha}.$$

Donc en moyenne la proportion de cases libres est environ $\frac{\mathbb{E}Z}{M} \simeq e^{-\alpha}$.

APPLICATION AUX TABLES DE HACHAGE (4)

Variabilité du nombre de cases vides.
Il faut calculer $\mathbb{V}ar\, Z = \mathbb{V}ar\,(Z_1 + Z_2 + \cdots + Z_M)$, or les variables Z_i **ne sont pas** indépendantes, donc $\mathbb{V}ar\, Z \neq M\mathbb{V}ar\, Z_1 = Me^{-\alpha}$.

$$
\begin{aligned}
\mathbb{V}ar\, Z &= \mathbb{E}Z^2 - (\mathbb{E}Z)^2 = \mathbb{E}Z^2 - M^2 e^{-2\alpha} \\
&= \mathbb{E}(Z_1 + Z_2 + \cdots + Z_M)^2 - M^2 e^{-2\alpha} \\
&= \sum_{i=1}^{M} Z_i^2 + 2\sum_{i<j}^{M} \mathbb{E}Z_i.Z_j - M^2 e^{-2\alpha} \\
&= Me^{-\alpha} + 2\frac{M(M-1)}{2}\mathbb{E}Z_1.Z_2 - M^2 e^{-2\alpha} \\
&\quad \text{or } \mathbb{E}Z_1.Z_2 = \mathbb{P}(Z_1 Z_2 = 1) = \left(\frac{M-2}{M}\right)^n \simeq e^{-2\alpha} \\
&= Me^{-\alpha} + M(M-1)e^{-2\alpha} - M^2 e^{-2\alpha} \\
&= Me^{-\alpha}(1 - e^{-\alpha})
\end{aligned}
$$

tout se passe comme si les variables Z_i et Z_j étaient indépendantes pour n et M grands

CONVERGENCE

Théorème de convergence vers une loi de Poisson

Soit $S_1, S_2, \cdots, S_n, \cdots$ une suite de variables aléatoires de loi $\mathcal{B}in(n, \frac{\alpha}{n})$, alors

$$\lim_{n\to\infty} \mathbb{P}(S_n = k) = e^{-\alpha}\frac{\alpha^k}{k!}$$

on dit que la suite S_n converge en loi vers une loi de Poisson de paramètre α et on note

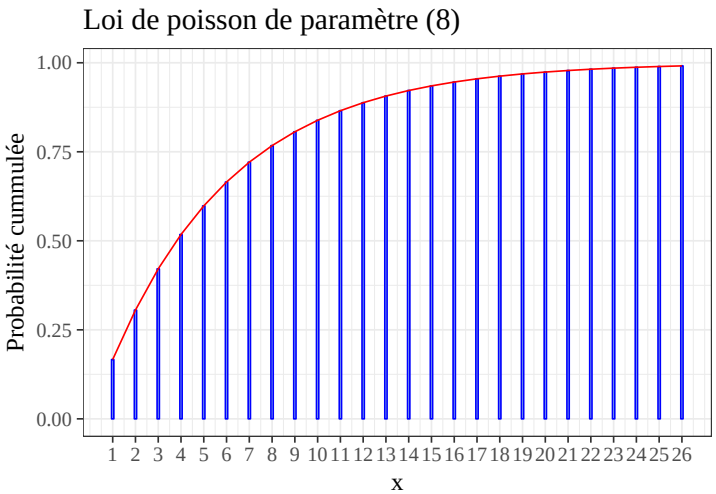
$$\mathcal{B}in(n, \frac{\alpha}{n}) \xrightarrow{\mathcal{L}} \mathcal{P}(\alpha).$$

Preuve

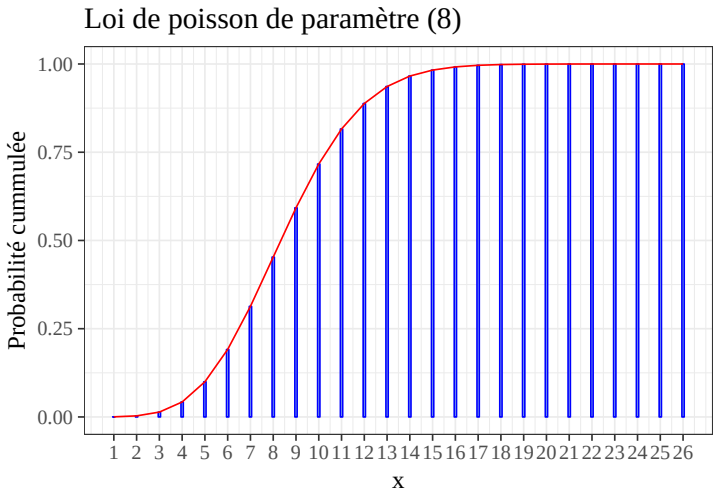
$$\begin{aligned}
\mathbb{P}(S_n = k) &= \binom{n}{k}(\frac{\alpha}{n})^k(1-\frac{\alpha}{n})^{n-k} \\
&= \frac{n(n-1)\cdots(n-k+1)}{k!}\frac{\alpha^k}{n^k}(1-\frac{\alpha}{n})^n(\frac{n-\alpha}{n})^{-k} \\
&= \underbrace{\frac{n(n-1)\cdots(n-k+1)}{(n-\alpha)^k}}_{\longrightarrow 1}\frac{\alpha^k}{k!}\underbrace{(1-\frac{\alpha}{n})^n}_{\longrightarrow e^{-\alpha}} = e^{-\alpha}\frac{\alpha^k}{k!}
\end{aligned}$$

LA LOI DE POISSON

Densité discrète $\mathscr{P}(8)$



Fonction de répartition



Pour $X \sim \mathscr{P}(\alpha)$

$$\mathbb{E}X = \alpha \text{ et } \mathbb{V}ar\, X = \alpha$$

GRANDES VALEURS

Pour $X \sim \mathcal{P}(\alpha)$

$$\mathbb{P}(X \geqslant k) = \sum_{i=k}^{\infty} e^{-\alpha} \frac{\alpha^i}{i!} = e^{-\alpha} \alpha^k \sum_{j=0}^{\infty} \frac{\alpha^j}{(j+k)!}$$

or $(k+j)! \geqslant k!j!$ d'où

$$\mathbb{P}(X \geqslant k) \leqslant \frac{\alpha^k}{k!}$$

La décroissance de la probabilité d'excéder k décroit extrêmement rapidement,

LOI GÉOMÉTRIQUE

Considérons un schéma de Bernoulli, suite infinie de variables aléatoires $X_1, \cdots, X_n, \cdots$ à valeur dans $\{0,1\}$, indépendantes de même loi de Bernoulli $\mathcal{B}(p)$ avec $0 \leqslant p \leqslant 1$, c'est à dire que pour tout n et pour tout vecteur $[a_1, \cdots, a_n]$ avec $a_i \in \{0,1\}$ on a

$$\mathbb{P}([X_1, X_2, \cdots, X_n] = [a_1, \cdots, a_n]) = p^{\sum_i a_i}(1-p)^{n-\sum_i a_i}.$$

Soit Y la variable aléatoire modélisant la première apparition de "pile", la loi de Y se calcule comme suit :

$$\begin{aligned}
\mathbb{P}(Y=k) &= \mathbb{P}(X_1 = 0, X_2 = 0, \cdots, X_{k-1} = 0, X_k = 1) \\
&\quad \text{l'événement correspond à } k-1 \text{ tirages successifs de "face"} \\
&\quad \text{suivi d'un "pile"} \\
&= \mathbb{P}(X_1 = 0)\mathbb{P}(X_2 = 0)\cdots\mathbb{P}(X_{k-1} = 0)\mathbb{P}(X_k = 1) \\
&\quad \text{car les variables } X_i \text{sont indépendantes,} \\
&= (1-p) \times (1-p) \times \cdots \times (1-p) \times p \\
&\quad \text{car les } X_i \text{ont la même loi } \mathcal{B}(p), \\
&= (1-p)^{k-1}p
\end{aligned}$$

La loi de Y est appelée loi géométrique et notée $\mathcal{G}eom(p)$

LOI GÉOMÉTRIQUE (2)

- La **loi** d'une variable aléatoire Y de loi géométrique est

$$\mathbb{P}(Y = k) = (1-p)^{k-1}p \quad \text{pour } k = 1, 2, 3, \cdots$$

- Sa **moyenne**

$$\mathbb{E}Y = \frac{1}{p}.$$

Ce qui s'interprète en terme de fréquence : dans une suite de lancers de la pièce la fréquence d'apparition de "pile" est p, donc en moyenne entre 2 "pile" on fait $\frac{1}{p}$ tirages.

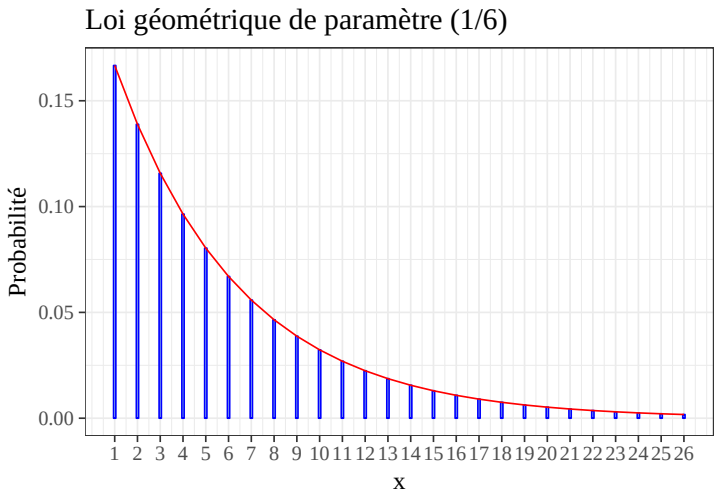
- La **variance** d'une variable aléatoire de loi géométrique est

$$\mathbb{V}ar\, Y = \mathbb{E}(Y - \mathbb{E}Y)^2 = \frac{1-p}{p^2}$$

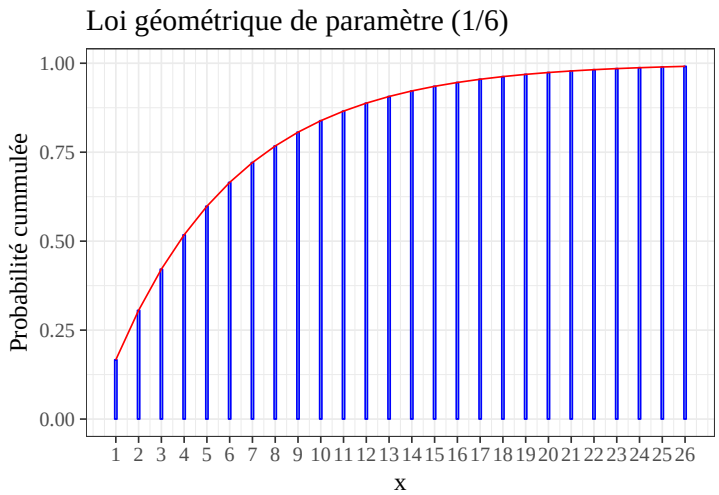
Refaire le calcul à partir de la loi.

LA LOI GÉOMÉTRIQUE (3)

Densité discrète



Fonction de répartition



$$\mathbb{E}Y = \frac{1}{p} = 6$$

$$\mathbb{V}ar\, Y = \frac{1-p}{p^2} = 30$$

TEMPS DE REMPLISSAGE

- On note T_M le nombre d'insertions à réaliser pour que la table soit "remplie", c'est à dire que toutes les cases soient occupées.
- Lorsque la table contient déjà j cases occupées, on note Y_j le nombre d'objets à insérer jusqu'à ce qu'une nouvelle case soit occupée.

Lorsque j est le nombre de cases occupées. Donc on retrouve un schéma de Bernoulli, suite de tirages jusqu'à trouver une case vide. Donc Y_j suit une loi géométrique. Le paramètre de cette loi est la probabilité de tomber sur une case vide, c'est à dire $\frac{M-j}{M}$. On en déduit que pour $0 \leqslant j \leqslant M$

$$\mathbb{P}(Y_j = k) = \left(1 - \frac{M-j}{M}\right)^{k-1} \frac{M-j}{M} = \left(\frac{j}{M}\right)^{k-1} \left(1 - \frac{j}{M}\right)$$

D'où la valeur moyenne

$$\mathbb{E}Y_j = \frac{M}{M-j},$$

et la variance

$$\mathbb{V}ar\, Y_j = \frac{1 - \frac{M-j}{M}}{\left(\frac{M-j}{M}\right)^2} = \frac{M^2}{(M-j)^2} - \frac{M}{M-j}.$$

ESPÉRANCE DE T_M

T_M est le nombre d'insertions pour remplir toutes les cases de la table. Donc

$$T_M = Y_0 + Y_1 + \cdots Y_{M-1} = \sum_{j=0}^{M-1} Y_j.$$

On en déduit

$$\begin{aligned} \mathbb{E}T_M &= \mathbb{E}\left(\sum_{j=0}^{M-1} Y_j\right) = \sum_{j=0}^{M-1} \mathbb{E}Y_j, \\ &\quad \text{car } \mathbb{E} \text{ est un opérateur } \textbf{linéaire } (\mathbb{E}(A+B) = \mathbb{E}A + \mathbb{E}B) \\ &\quad \text{puis en remplaçant } \mathbb{E}Y_j \text{ par sa valeur, et} \\ &\quad \text{en faisant le changement de variable } k = M - j \\ &= \sum_{j=0}^{M-1} \frac{M}{M-j} = M \sum_{j=0}^{M-1} \frac{1}{M-j} = M \sum_{k=1}^{M} \frac{1}{k} = M\left(1 + \frac{1}{2} + \frac{1}{3} + \cdots + \frac{1}{M}\right). \end{aligned}$$

$\mathbb{E}T_M$: INTERPRÉTATION

Lorsque M est grand $\sum_{k=1}^{M} \frac{1}{k}$ est de l'ordre de grandeur de $\log M$, (voir la formule d'Euler-Mascheroni), [1].

$$\mathbb{E}T_M \simeq M \log M$$

Il faudra donc en moyenne $M \log M$ objets pour que chaque case de la table soit touchée, et dans ce cas le nombre moyen d'objets hachés dans une case sera de l'ordre de $\mathcal{O}(\log M)$.
Dans le problème des "Panini" on calcule le nombre moyen de vignettes achetées pour constituer une collection $300 \log 300 \simeq 1700$, à 0.10 euros la vignette, la collection coûtera en moyenne 170 euros ! ! !

1. on peut comprendre cet équivalent en remarquant que la somme des $\frac{1}{k}$ est proche de la surface sous la courbe $y = \frac{1}{x}$ entre 1 et M et que la primitive de $\frac{1}{x}$ est $\log x$

$\mathbb{V}ar\, T_M$

On calcule de la même manière la variance de T_M

$$\mathbb{V}ar\, T_M = \mathbb{V}ar\left(\sum_{j=0}^{M-1} Y_j\right) = \sum_{j=0}^{M-1} \mathbb{V}ar\, Y_j,$$

attention : car les variables sont indépendantes

puis en remplaçant $\mathbb{V}ar\, Y_j$ par sa valeur, et

en faisant le changement de variable $k = M - j$

$$= \sum_{j=0}^{M-1}\left(\frac{M^2}{(M-j)^2} - \frac{M}{M-j}\right) = M^2 \sum_{k=1}^{M} \frac{1}{k^2} - M\sum_{k=1}^{M}\frac{1}{k}$$

Le premier terme est de l'ordre de M^2 (on rappelle que $\sum_{k=1}^{M} \frac{1}{k^2} \leqslant \sum_{k=1}^{\infty} \frac{1}{k^2} = \frac{\pi^2}{6}$) et le deuxième en $M \log M$ est négligeable devant M^2.
Donc la variance croit en M^2 et l'écart-type $\sqrt{\mathbb{V}ar\, T_M}$ en $\mathcal{O}(M)$, la loi de T_M est donc "proche" de sa valeur moyenne.
Pour aller plus loin, on peut utiliser l'inégalité de Bienaymé-Tchebychev qui lie l'écart à la moyenne à la variance

$$\mathbb{P}(|T_M - \mathbb{E}T_M| > a) \leqslant \frac{\mathbb{V}ar\, T_M}{a^2}.$$