Loi des Grands Nombres

Il arrive souvent que l'on répète un grand nombre de fois une épreuve de façon indépendante. Soit A un événement qui peut se réaliser avec une probabilité p. Que peut-on dire du lien entre la fréquence observée de A et sa probabilité p ? Il arrive que p soit inconnue ; peut-on en faire une approximation fondée sur la fréquence de A ?

Avant de répondre à ces questions, considérons le cas d'une v.a. Sa variance étant introduite pour mesurer sa dispersion autour de l'espérance. Peut-on, en termes de ce paramètre, confirmer avec une confiance relativement grande, que l'écart entre l'espérance et la valeur prise par la v.a. ne sera pas trop grand ?

Proposition (Inégalité de Bienaymé-Tchebychev). Soit X un v.a. d'espérance μ et de variance V. Alors, pour tout $\epsilon > 0$:

$$Pr(|X - \mu| \geq \epsilon) \leq \frac{V}{\epsilon^2}.$$

Exemple. Soit X_n une v.a. binomiale de paramètres $p = \frac{1}{10}$ et $n \in \mathbb{N}$.

- Peut-on confirmer, que pour $n = 100$ l'écart entre $\frac{X_n}{n}$ et son espérance $\frac{1}{10}$ ne dépasse pas 0.1 avec une probabilité supérieure à 0.9 ?
- Pour quelles valeurs de n, peut-on confirmer, qu'avec une probabilité ≥ 0.99, cet écart ne dépasse pas 0.05 ?

Solution

- La variance de $\frac{X_n}{n}$ pour $n = 100$ vaut :

$$npq \times \frac{1}{n^2} = \frac{9}{10000}.$$

D'après l'inégalité de Bienaymé-Tchebychev, nous avons :

$$Pr\left(|\frac{X_{100}}{100} - \frac{1}{10}| \geq 0.1\right) \leq \frac{9}{10000} \times 100 = 0.09 < 0.1.$$

La réponse est donc positive.

- Nous avons :

$$Pr\left(|\frac{X_n}{n} - \frac{1}{10}| \geq 0.05\right) \leq \frac{9}{100} \times \frac{1}{n} \times \frac{1}{0.05^2}.$$

D'où, pour que la probabilité recherchée soit au moins égale à 0.99 il *suffit* que n soit assez grand pour qu'on ait $\frac{9}{100} \times \frac{1}{n} \times \frac{1}{0.0025} \leq 0.01$, c'est-à-dire $n \geq 3600$.

Théorème (Loi faible des grands nombres). Soit X une v.a. admettant l'espérance μ et la variance V et soit $X_1, X_2, ..., X_n$ une suite de v.a. indépendantes chacune suivant la même loi que X. Désignons par S_n la valeur moyenne de la suite :

$$S_n = \frac{1}{n} \sum_{i=1}^{n} X_i.$$

Alors S_n *converge en probabilité* vers son espérance μ ; i.e. pour tout $\epsilon > 0$, on a :

$$\lim_{n \to \infty} Pr(|\ S_n - \mu\ | \geq \epsilon) \ = \ 0.$$

Exemple. On lance une paire de dés authentiques n fois et l'on calcule la moyenne arithmétique des produits des deux points. Nous avons vu que l'espérance du produit vaut $\frac{49}{4}$. Puisque cette v.a. admet une variance, la loi des grands nombre permet de confirmer que la moyenne des produits converge en probabilité vers $\frac{49}{4}$.

Une Application : Problème de Hachage*

Une technique populaire utilisée en informatique pour l'organisation de données est celle de hachage. On voudrait gérer un ensemble d'enregistrements, chacun ayant une *clé*. L'accès à un enregistrement se fait via sa clé. A titre d'exemple, on peut considérer les enregistrements contenant des informations sur les étudiants d'une promotion. On peut munir cet ensemble de données de clés d'accès qui seront les noms des étudiants. La méthode de hachage place un enregistrement en fonction de sa clé, la transformant directement en une adresse dans une zone de mémoire contiguë. L'ensemble de ces méthodes permet les opérations de recherche, d'adjonction et de suppression.

———————————————————————————

Types de Données et Algorithmes, C. Froidevaux, M.-C. Gaudel, M.Soria.

Dans la suite, nous proposons le modèle simple suivant. On se donne un tableau de hachage à m places. Nous avons un univers U des clés. U étant de taille très grande, nous ne pouvons pas affecter à tous ses éléments une place attitrée dans le tableau de hachage. On utilise alors une *fonction de hachage* h, qui associe à chaque clé un entier dans $[1, m]$:

$$h \ : \ U \ \rightarrow \ [1, m].$$

Le choix de h est fondamental : il faut appliquer U de manière aussi uniforme que possible sur $[1, m]$. Cela revient à dire que, dans le cas idéal, on doit avoir :

$$\forall x \in U \text{ et } \forall i \in [1, m] \quad Pr(h(x) = i) \ = \ \frac{1}{m}.$$

On dit alors que h est *uniforme*. Il est aussi souhaitable que le calcul de h soit rapide.

Nous disons qu'il y a une *collision* entre deux clés distinctes $x \in U$ et $y \in U$ *sur la case* v, si $h(x)$=$h(y)$=v.

Étant donné un ensemble E de n clés distinctes, la probabilité pour qu'il n'y ait pas de collisions entre ses éléments vaut :
$P \ = \ \frac{1}{m^n}.m(m-1)...(m-n+1)$. Cette probabilité est petite lorsque m n'est pas très grand par rapport à n, comme le montre le calcul suivant.

Paradoxe d'anniversaire

En supposant que l'année comporte 365 jours et que la probabilité d'avoir son anniversaire est la même pour les jours de l'année, calculer la probabilité pour que, dans une classe de n étudiants les anniversaires soient tous différents. Calculer numériquement cette probabilité pour une classe de 23 étudiants.

Solution. Portant la valeur $m = 365$ dans l'équation or P, il vient :

$$P = \frac{1}{365^n}.365.364.363...(363 - n + 1).$$

La courbe de variations de $1 - P$ en fonction de n se trouve sur la page suivante.

On voit que pour un nombre d'étudiants $n \geq 23$, ce qui est relativement petit par rapport à 365, la probabilité d'anniversaires différents tombe au dessous de 0.5.

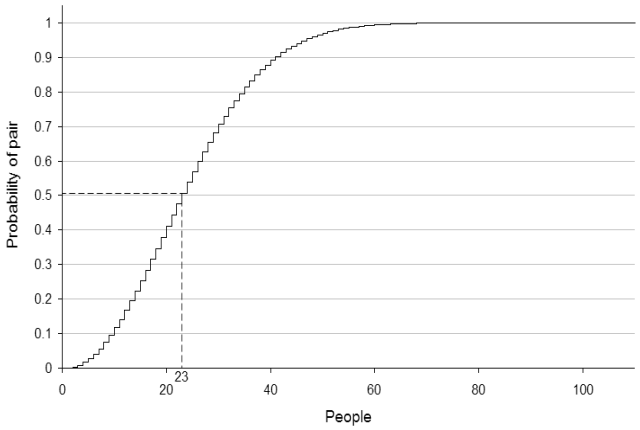
On peut donc raisonnablement s'attendre à des collisions, même lorsque la taille de la table est relativement élevée par rapport au nombre d'éléments qui sont à y placer. Un traitement de collisions s'impose alors.

Soit v un indice quelconque fixé dans l'intervalle réel $[1, m]$. Supposons que n clés soient présentes dans le tableau. La valeur $\alpha = \frac{n}{m}$ est appelée *taux de remplissage du tableau*. Soit $X_{m,n}$ la v.a. désignant le nombre de clés x telles que $h(x) = v$.

Si $m, n \to \infty$, alors $X_{m,n}$ tend vers une v.a. de Poisson de paramètre α ; i.e. :

$$Pr(X_{m,n} = k) = e^{-\alpha}\frac{\alpha^k}{k!}, \quad \forall k \in \mathbb{N}.$$

En particulier, la probabilité pour que la case v soit vide vaut $e^{-\alpha}$.



Considérons maintenant la méthode de chaînage séparé dans la résolution des collisions : on fait une liste chaînée des clés en collision sur la même case dans leur ordre d'arrivée.

Faisons une analyse de la complexité moyenne, en termes de nombre de comparaisons, pour la recherche d'une clé x dans un tableau de hachage de taux de remplissage $\alpha = n/m$. On peut alors considérer deux cas distincts :

- La clé x ne figure pas dans le tableau (recherche négative) ; le nombre de comparaisons nécessaires pour conclure qu'elle n'y est pas est la longueur de la liste chaînée dans la case $h(x)$.
- La clé x figure dans le tableau (recherche positive) ; on pourra la trouver *peut-être avant* de parcourir *toute* la liste chaînée.

Nous sommes donc ramenés à faire une analyse différente pour chacun des cas : la complexité moyenne d'une recherche négative et celle d'une recherche positive.

Commençons par la plus simple.

Complexité Moyenne d'une Recherche Négative

Reprenons les données précédentes pour un tableau de hachage. Soit L_i la v.a. désignant la longueur de la liste située dans dans la $i^{\text{ème}}$ case du tableau. Supposons qu'on cherche une clé x, qui n'existe pas, dans le tableau. D'après l'hypothèse d'uniformité $h(x)$ peut prendre une valeur $i \in [1, m]$ avec la même probabilité $\frac{1}{m}$.

Complexité Moyenne d'une Recherche Positive

Supposons qu'on cherche une clé x qui *figure* dans le tableau de hachage. Nous retenons l'hypothèse d'uniformité qu'elle peut être égale à une des n clés existant dans le tableau avec la *même* probabilité $\frac{1}{n}$. Soient $x_1, ..., x_n$ les clés du tableau dans leur ordre d'insertion. On voit facilement que si $x = x_1$, le nombre de comparaisons pour la trouver est 1 et, de façon générale, si $x = x_i$, le nombre de comparaisons vaut le nombre de comparaisons, effectuées lors de l'insertion la clé x_i, plus 1. Ce dernier nombre en moyenne n'est que le nombre moyen de comparaisons dans une recherche négative dans le tableau lorsqu'il n'a que $i - 1$ clés.

Puisque, pour arriver à la conclusion que x n'est pas dans le tableau, il faut effectuer L_i comparaisons (où $i = h(x)$), l'espérance du nombre de comparaisons vaut :

$$CompRech^{-}(m,n) = \frac{1}{m}\sum_{i=1}^{m}\mathbb{E}(L_i).$$

Par ailleurs ces dernières espérances valent chacune $\alpha = \frac{n}{m}$. En effet chacune des n clés du tableau contribue une augmentation égale à $\frac{1}{m}$ à l'espérance de L_i. Nous avons donc :

$$CompRech^{-}(m,n) = \frac{1}{m} \times m \times \frac{n}{m} = \frac{n}{m} = \alpha,$$

Nous avons donc :

$$\begin{aligned} CompRech^{+}(m,n) &= \frac{1}{n}\sum_{i=0}^{n-1}\left[CompRech^{-}(m,i)+1\right] \\ &= 1 + \frac{n(n-1)}{2nm} \\ &= \frac{\alpha}{2} - \frac{1}{2m} + 1, \end{aligned}$$

ce qui vaut asymptotiquement $\alpha/2$. Ces calculs de complexité moyenne mettent en évidence l'*efficacité* des techniques de chaînage dans le traitement de collisions, en confirmant que dans les deux cas, le nombre moyen de comparaisons pour chercher une clé est proche d'une constante et non pas proportionnel au nombre d'éléments du tableau.