Paradoxe des anniversaires

1 Paradoxe des anniversaires

Le paradoxe des anniversaires, dû à Richard von Mises, estime la probabilité d'avoir deux personnes ayant la même date d'anniversaire dans une assemblée de k personnes. L'intuition courante est que cette probabilité est faible alorsque pour un groupe de 23 personnes par exemple la probabilité d'avoir deux personnes ayant la même date d'anniversaire est d'un peu plus de 50%.

1.1 Principe

Passons à la modélisation du phénomène. On suppose que toutes les années comptent 365 jours et que toutes les dates de naissances sont équiprobables (c'est à dire qu'il n'existe pas de période dans l'année où il y a plus de naissances). On va calculer la probabilité p_k pour que dans un groupe de k personnes toutes aient une date d'anniversaire différente.
Pour un groupe de 2 personnes : La première peut avoir son anniversaire n'importe quel jour. La seconde a le choix entre toutes les dates d'anniversaire sauf celle de la première personne.

$$p_2 = 1 - \frac{1}{365}$$

Pour un groupe de 3 personnes : La première a une date d'anniversaire quelconque, la deuxième doit éviter cette date, la troisième a le choix entre toutes les dates sauf les deux dates d'anniversaire des deux premmières personnes.

$$p_3 = \left(1 - \frac{1}{365}\right)\left(1 - \frac{2}{365}\right)$$

Pour un groupe de k ($k \leq 365$) personnes : en réitérant le même raisonnement on obtient

$$p_k = \left(1 - \frac{1}{365}\right)\left(1 - \frac{2}{365}\right)\ldots\left(1 - \frac{k-1}{365}\right)$$

Si le nombre de personnes est supérieur à 366, on utilise le principe des tiroirs énoncé par Dirichlet au XIX ème siècle. Ce principe s'énonce ainsi : "Si on dispose de k tiroirs et d'au moins $k+1$ paires de chaussettes, il y a au moins un tiroir où sont rangées deux paires de chaussettes." Appliqué à notre problème, ce principe nous dit qu'à partir de 366 personnes, il y a nécessairement deux personnes qui ont la même date de naissance. La probabilité p_k est donc nulle dans ce cas.

1.2 Application numérique

On trace à la fois p_k la probabilité que toutes les personnes dans un groupe de k personnes aient une date d'anniversaire différente et $1 - p_k$ la probabilité qu'il existe au moins un couple de personne ayant le même jour comme anniversaire.

Dès que le groupe atteint la taille de 23 personnes on a 50% de chance qu'il y au moins deux personnes ayant la même date de naissance. Pour 50 personnes, la probabilité atteint 97%.

2 Applications

2.1 Application aux test ADN

Les profils d'ADN représentent un des outils de l'individualisation des échantillons biologiques et peuvent ainsi être utilisé comme preuve pénale.
Le système Quad STR, par exemple, analyse quatre marqueurs génétiques. La probabilité d'un profil génétique décrit par ce système est approximativement égal à 1/10000. La probabilité d'avoir dans un groupe de k deux personnes ayant la même empreinte

génétique par ce test est donnée par :

$$1 - p_k = 1 - \left[\left(1 - \frac{1}{10000}\right)\left(1 - \frac{2}{10000}\right)\dots\left(1 - \frac{k-1}{10000}\right)\right]$$

Sur ce test la probabilité d'avoir deux personnes ayant le même profil génétique est de 50% dès qu'on atteint un groupe de 120 personnes. On trace de nouveau p_k la probabilité que toutes les personnes dans un groupe de k personnes aient des profils génétiques différents et $1 - p_k$ la probabilité qu'il existe au moins un couple de personne ayant le même profil.

Ce test seul ne permet donc pas de construire une base de données d'empreintes génétiques. Ce test est bien entendu couplé à d'autres tests plus discriminants. Mais alors que la probabilité d'un profil génétique n'est que de $0,01\%$, la probabilité d'avoir deux personnes ayant le même profil génétique croît relativement vite avec le nombre de personnes testées. Pour améliorer la fiabilité de ces tests, il faut augmenter le nombre de marqueurs considérés et ainsi réduire la probabilité de chaque profil.

2.2 Application aux fonctions de hachage

Une fonction de hachage est une fonction qui fait subir une succession de traitements à une donnée quelconque fournie en entrée pour en produire une « empreinte » servant

à identifier la donnée initiale (comme précédemment l'empreinte génétique servait à identifier un individu).
Les fonctions de hachage sont utilisées en cryptographie. Dans le cadre de la signature électronique, au lieu de signer l'intégralité d'un message ce qui serait coûteux en temps de calcul on applique au message une fonction de hachage et on signe par l'empreinte du message. La personne qui reçoit le message peut vérifier si l'empreinte et le message sont compatibles ce qui doit l'assurer de l'authencité du message. Pour que ce procédé suffise à identifier le message il faut que la fonction de hachage renvoie le moins souvent possible à deux antécédents (ou messages) différents la même image (ou empreinte). Car si les collisions sont trop nombreuses, le message pourrait être modifié sans que la signature le soit. Le destinataire du message ne pourrait alors pas s'apercevoir de la falsification. Le paradoxe des anniversaires permet d'estimer le taux de collision provoqué par une fonction de hachage en fonction du nombre d'images (ou empreintes) qu'elle peut renvoyer.

Références

[1] http ://fr.wikipedia.org/wiki/Paradoxe_des_anniversaires

[2] http ://www.library.uu.nl/publarchief/jb/congres/01809180/15/b24.pdf