L’INCROYABLE STATISTIQUE DES PREMIERS CHIFFRES

Dès l’école primaire, vous avez appris à utiliser les chiffres de 0 à 9 pour représenter les nombres « en base 10 ». Depuis, cette écriture vous est sans doute devenue familière et vous pensez qu’elle n’a plus de mystère pour vous.

Pourtant, le phénomène dont nous allons parler est si inattendu qu’il vous faudra probablement le constater vous-même pour y croire !

Photographie du bandeau / miniature sur la page d’accueil : “Atelier sur la loi de Benford à la Fête de la Science” – Thierry de la Rue, Gaëlle Chagny

Expérimentez la loi de Benford

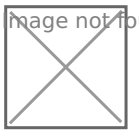
Ouvrez au hasard des pages de journaux, de revues, de sites d’information ou de réseaux sociaux, et relevez tous les nombres que vous y trouvez. Puis intéressez-vous au premier chiffre significatif de chacun de ces nombres : c’est le chiffre le plus à gauche, qui n’est pas zéro. Ne tenez compte ni du signe ni de la place de la virgule : par exemple, le premier chiffre significatif des nombres 0,038 3,14159 et -32 est 3. On peut penser a priori que chacun des chiffres de 1 à 9 sera vu avec la même fréquence comme premier chiffre significatif. Pourtant, si vous relevez beaucoup de nombres d’origines variées, vous constaterez que le chiffre 1 apparaît au début de près d’un tiers des nombres, le chiffre 2 environ une fois sur 6, et que les fréquences diminuent jusqu’au chiffre 9 (moins d’une fois sur 20).

Image not found or type unknown

Fréquences théoriques des premiers chiffres significatifs selon la loi de Benford.

Cette distribution du premier chiffre significatif est aujourd'hui connue sous le nom de « Loi de Benford », d'après l'ingénieur américain qui l'a vérifiée en 1938, en répertoriant plus de 20 000 nombres provenant de multiples sources (longueurs de fleuves, cours de la bourse, résultats de base-ball, poids des éléments chimiques, etc.).

Frank Benford propose même une formule précise pour décrire avec quelle distribution apparaissent les chiffres de 1 à 9 comme premier chiffre significatif : la fréquence du chiffre i (i variant entre 1 et 9) est donnée par le logarithme à base 10 de (1+i)/i. Par exemple lorsque i est le chiffre 2, vous pouvez vérifier sur une calculatrice que la fréquence donnée par la formule de Benford vaut :



Formule donnant la fréquence du chiffre 2 en tant que premier chiffre significatif.

Une découverte grâce à l'absence de calculatrice

La fonction logarithme qui apparaît dans la formule ci-dessus a joué un grand rôle dans la découverte de cette étrange loi. Le logarithme était très utilisé avant l'avènement de l'ordinateur pour sa faculté à transformer les multiplications et divisions, opérations très compliquées à effectuer à la main, en additions et soustractions (un peu plus simples !). Pour effectuer des calculs, on avait donc couramment recours à des tables de logarithmes, petits livres qui donnaient les logarithmes des nombres que l'on voulait multiplier.

Ainsi, pour calculer rapidement le quotient 12 345 ÷ 6 789, on commençait par consulter la table pour obtenir les logarithmes de 12 345 et 6 789, qui valent respectivement 4,0915 et 3,8318. On calculait à la main la différence entre ces deux nombres, qui donne 0,2597, puis en utilisant dans l'autre sens la table de logarithmes, on trouvait le quotient, qui est le nombre dont le logarithme est égal à cette différence, soit environ 1,818.

Calcul d’un quotient en utilisant les tables de logarithmes.

C’est l’astronome américain Simon Newcomb qui a remarqué que les premières pages de ces tables de logarithmes étaient plus rapidement usées que les dernières, pour la raison que l’on utilisait plus souvent des nombres commençant par un 1 que par un 9. Newcomb publia le premier article sur cette surprenante distribution des premiers chiffres dès 1881, mais son travail est à l’époque passé inaperçu.

Près de 50 ans plus tard, en observant à nouveau l’usure irrégulière des tables de logarithmes, Benford refit la même découverte.

Plus les données sont variées, mieux ça marche

La distribution prédite par la loi de Benford se vérifie expérimentalement sur toute série de données issues du monde réel, pourvu que cette série soit assez « riche » (nombres d’origines variées et/ou réparties sur plusieurs ordres de grandeur).

En effet on comprend bien que, si par exemple on ne considère que des tailles d’individus exprimées en centimètres, le premier chiffre significatif sera presque tout le temps le 1 et donc la loi de Benford ne sera pas satisfaite. En revanche, la série constituée des nombres d’habitants par commune sur un territoire assez grand se conforme plutôt bien à la loi de Benford, car la taille des villes peut varier de quelques centaines à plusieurs millions d’habitants.

Ainsi, le graphique ci-dessous illustre les résultats obtenus en étudiant le premier chiffre significatif des nombres d’habitants des communes de la région Normandie. Globalement, on retrouve bien l’allure du diagramme en barres prévu par la loi de Benford.

Diagramme en barres du premier chiffre significatif des nombres d’habitants des communes normandes (Insee 2019).

Pourquoi plus de 1 ?

On peut se demander pourquoi le 1 et le 2 sont plus souvent utilisés comme premier chiffre significatif que le 8 ou le 9. Après tout, il y a autant de nombres dans l'intervalle [9 000, 10 000) (donnant un 9 comme premier chiffre) que dans l'intervalle [1 000, 2 000), qui vont donner un 1.

Mais l'erreur bien naturelle que l'on commet en comparant ainsi les tailles de ces deux intervalles consiste à les mesurer de manière additive : dans les deux cas, il faut ajouter 1 000 à la borne inférieure pour obtenir la borne supérieure. Or, comme le montre très bien Mickaël Launay dans son livre ce raisonnement « additif » n'est pas pertinent : quand on compare des nombres de la vie réelle, on le fait plutôt multiplicativement. La taille « multiplicative » du premier intervalle vaut 10 000 ÷ 9 000 soit environ 1,11, elle est beaucoup plus petite que celle du second, qui vaut 2 000 ÷ 1 000, soit 2.

Voici une situation très concrète pour montrer en quoi ce point de vue multiplicatif est beaucoup mieux adapté. Intéressons-nous aux prix de biens de consommation courante, et disons pour simplifier que ces prix suivent tous une même inflation lente et régulière. Prenons un prix dont le premier chiffre significatif est 1, disons la baguette de pain à 1 euro. Son premier chiffre significatif va rester 1 tant que le prix de la baguette n'aura pas atteint 2 euros, soit pendant tout le temps nécessaire pour obtenir une augmentation des prix de 100 %. Considérons en parallèle le prix d'un litre d'huile d'olive à 9 euros : son premier chiffre significatif restera 9 seulement le temps que l'inflation le fasse monter à 10 euros (augmentation de seulement 11 %).

En pensant ainsi multiplicativement, la distribution prédite par la loi de Benford devient beaucoup plus naturelle. Les intervalles [1,2), [2,4), [3,6), [4,8) et [5,10) ont la même taille multiplicative 2. Les sommes des fréquences des premiers chiffres significatifs vus dans chacun de ces intervalles sont alors égales :

Sommes des fréquences des premiers chiffres significatifs prévues par la loi de Benford sur des intervalles de taille multiplicative 2.

Cette vision multiplicative se retrouve dans un autre argument couramment avancé pour expliquer la loi de Benford : la distribution du premier chiffre significatif doit être la même en France, où l'on mesure les distances en kilomètres et les prix en euros, qu'aux États-Unis où l'on utilise les miles et les dollars. Autrement dit elle ne doit pas dépendre du

choix des unités utilisées pour mesurer les grandeurs. Ainsi les fréquences des premiers chiffres significatifs ne doivent pas changer si l'on multiplie toutes les données par un même nombre (ce qui correspond à un changement d'unité). Or la loi de Benford est la seule distribution qui satisfait cette invariance.

Un détecteur de fraude

La loi de Benford peut sembler n'être qu'une curiosité anecdotique. Cependant, au début des années 90, l'économiste Mark Nigrini lui trouva une application très concrète : il eut l'idée de l'utiliser pour la détection de fraudes dans des données, et y a même consacré un ouvrage en 2012.

En effet, si une série de nombres variés provenant de données réelles suit théoriquement la distribution prédite par Benford, Nigrini montre que dans des données comptables falsifiées, la fréquence de nombres commençant par 5 ou 6 est largement plus élevée : la plupart des faussaires ignorent la loi de Benford ! Des experts-comptables peuvent ainsi mettre en évidence les fraudes des sociétés. Il semble courant aujourd'hui de se baser sur la loi de Benford (incluant des tests plus approfondis considérant également le second chiffre significatif des nombres) pour suspecter une fraude dans des données, qu'elles soient fiscales, comptables, électorales ou même scientifiques. Bien qu'un écart à la loi de Benford ne constitue pas une preuve de fraude, il peut orienter les experts vers des vérifications plus poussées.

Sans même être expert, ni avoir envie de débusquer des fraudes, la simple curiosité vous poussera peut-être à constater par vous même la loi de Benford quelles que soient vos prochaines lectures.

Auteurs

Thierry de la Rue, Chargé de recherche CNRS en mathématiques, *Université de Rouen Normandie* et **Gaëlle Chagny**, chargée de recherche CNRS en mathématiques (statistique), *Université de Rouen Normandie*

Publié le : 2022-06-29 11:59:51