INSA INSTITUT NATIONAL DES SCIENCES APPLIQUÉES ROUEN

Projet de Physique P6
STPI/P6/2024-018

Loi de Newcomb-Benford

Étudiants :

KERHOMEN Charlie
DECHAMPS Gatien
LE BORGNE Armand
QUIDÉ Alexandre

Enseignant-responsable du projet :
ZIDANI Hasnaa

Date de remise du rapport : 14/06/2024

Référence du projet : STPI/P6/2024 – 018

Intitulé du projet : Dévoilez l'authenticité des données expérimentales avec la loi de Newcomb-Benford

Type de projet : *Model/simulation numérique*

Objectifs du projet :
Partie bibliographique :
— Comprendre les fondements historiques et mathématiques de la loi de Newcomb-Benford
— Étudier son applicabilité et ses limites dans le cadre de la détection de fraudes
Modélisation et Application :
— Mettre en œuvre un programme pour évaluer la conformité de données avec la loi de Newcomb-Benford
— Analyse d'un cas issu du monde réel : des données fiscales italiennes (reproduction d'un résultat de la littérature)
— Vérification de résultats attendus dans la littérature

Mots-clefs du projet :
— Probabilités
— Statistiques
— Modélisation
— Programmation

Remerciements

Nous adressons nos premiers remerciements à Hasnaa Zidani, qui fut notre enseignant encadrant dans le cadre de ce projet. Elle nous a accompagnés avec sérieux, implication et bienveillance au cours du semestre, et sa (juste) exigence nous a poussés à pleinement développer nos travaux dans une atmosphère studieuse et épanouissante.

Merci également aux responsables d'EC, David Honoré et Jérôme Yon, de permettre à ces projets scientifiques ambitieux d'exister. Ce format pédagogique unique fut pour nous une opportunité d'explorer de façon approfondie un phénomène mathématique passionnant, tout en développant nos compétences de gestion de projet.

Merci enfin à nos camarades, Nour Chaabane, Kangping Huang, Naël Mathlouthi et Laurène Lecomte pour leurs précieux retours et leur aide dans la préparation de notre soutenance orale.

Institut National des Sciences Appliquées de Rouen
Département Sciences et Techniques Pour l'Ingénieur
685 Avenue de l'Université BP 08- 76801 Saint-Etienne-du-Rouvray
Tél : 33 2 32 95 66 21 - Fax : 33 2 32 95 66 31

Table des matières

Chapitre 1: Introduction

Si l'on s'intéresse à de nombreuses séries statistiques issues de mesures du monde réel, comme par exemple la longueur des fleuves du monde, on peut remarquer que le chiffre 1 apparaît plus souvent comme le premier chiffre significatif des valeurs prises par la série. Et il en va de même pour le chiffre 2, qui est lui-même plus fréquent que le 3, et ainsi de suite. Contrairement à ce que nous pourrions considérer intuitif, les premiers chiffres significatifs ne suivent pas une distribution uniforme la plupart du temps.

Fort de ce constat empirique, l'astronome Simon Newcomb - et après lui l'ingénieur Frank Benford-, formulèrent la loi aujourd'hui connue sous le nom de Newcomb-Benford, ou loi des nombres anormaux.

1.1 Organisation

Notre projet consiste en la compréhension de la loi de Newcomb-Benford, ainsi que de ses fondements mathématiques, autant par sa découverte que par la preuve de son existence. Mais aussi par la modélisation de cette loi, et comparer des données statistiques avec les valeurs théoriques de cette loi.

Ainsi, une bonne partie de notre travail fut centrée autour de la recherche et la mise en commun de documentation liée à chacune de ces parties ; l'histoire de la découverte de la loi, la recherche de preuve appuyant sa véracité, la recherche de domaines dans lesquels cette loi peut s'appliquer, et enfin des méthodes de comparaison entre la loi et des données statistiques.

C'est ainsi que notre groupe fonctionna durant l'intégralité de ce projet (voir Figure 1 page 24) :

Tout d'abord, chacun était responsable de sa propre partie (citées plus haut), mais chacun avait aussi le droit et était encouragé à ajouter son aide dans d'autres parties si nécessaire ou si sa propre partie était bien entamée par rapport aux autres, et chaque responsable de partie se chargeait de la rédaction de sa partie pour le rapport.

Cependant, nous nous voyions chaque semaine le mardi matin pour mettre en commun nos avancées dans le projet, ainsi que pour aborder les questions de chacun avant de faire un compte rendu à notre professeure responsable de projet plus tard dans la journée. Suite aux retours de notre professeure, nous corrigions notre cap pour les semaines à venir dans la gestion du projet.

Chapitre 2: Histoire de la loi de Newcomb-Benford

La première description connue de ce phénomène est attribuée à l'astronome, mathématicien et économiste américain Simon Newcomb (1835 - 1909).

En 1881, il publie un article intitulé *Note on the frequency of use of the different digits in natural numbers*, dans lequel il observe que les premières pages des tables de logarithme couramment utilisés (qui contenaient les logarithmes de nombres à chiffre significatifs étant 1) alors montraient des signes d'usures plus marqués que les pages suivantes. Il s'est donc ainsi posé une question qui amena bien plus tard à la loi que nous connaissons :

> *"La question que nous devons considérer est la suivante : quelle est la probabilité que, si un nombre entier était tiré aléatoirement, son premier chiffre significatif soit n, son second n', ect ... "* [1]

Aussi, il suppose que, comme les nombres entiers "se produisent" dans la nature, il est légitime de les considérer comme des rapports de quantités et que de fait, il ne faudrait pas simplement déterminer la probabilité d'apparition du chiffre significatif n sur un simple nombre aléatoire, mais plutôt sur le résultat du rapport de deux nombres entiers, pris aléatoirement.

Ainsi pour obtenir une réponse à la question posée ultérieurement, nous devriions former un nombre indéfini de tels rapports, pris indépendamment, puis réaliser cette expérience une telle quantité de fois que cela nous amènerait à tendre vers la limite à laquelle la probabilité n s'approcherait.

Après avoir définit la mantisse (que nous définiront en début de partie 3.1) , il énonce la loi ainsi :

> *"La loi de probabilité d'apparition des nombres est telle que toutes les mantisses de leurs logarithmes sont également probables."* [1]

Finalement, c'est dans cet article que l'on peut pour la première fois un tableau avec lesdites probabilités d'apparition (Voir annexe 1).

Cette première publication passe relativement inaperçue, de telle sorte que la même observation est mentionnée par plusieurs auteurs au cours des décennies suivante. En particulier, en 1938, l'ingénieur américain Frank Benford publie des observations similaires à celles de Newcomb dans un article intitulé *The Law of anomalous Numbers* [2].

Cette publication, bien plus remarquée, est la raison pour laquelle le phénomène est encore souvent connu sous le nom de "Loi de Benford", même si la découverte par ce dernier est postérieure à celle de Newcomb. Dans cette publication-ci, on peut y lire une formalisation plus poussée par rapport à celle posée par Simon Newcomb, complètant ainsi une partie jusquelà peu abordée : la proposition concrète d'une formule mathématique permettant de généraliser la solution au problème :

> *"La fréquence des premiers chiffres [significatifs] suit de près la relation logarithmique* $F_a = log[_{10}](\frac{a+1}{a})$" [2]

Dans la suite de cet article, Benford donne aussi plusieurs tableaux dans lesquels il montre les tests qu'il a pu réaliser sur des données dans 20 domaines différents : tailles de populations aux

USA, fleuves et rivières, constantes physiques, masses moléculaires, taux de mortalité, addresses postales des 300 premières personnes listées dans *"American men of science"*, pour un total de plus de 20.000 données.

Cependant, on remarque qu'entre les valeurs théoriques avancées par Benford varient extrêmement peu de celles de l'article de Newcomb (Voir Tableau 1 page 24).

Dans cet article, Benford y mentionne bon nombre d'informations afin de compléter la proposition qu'il apporte avec sa loi : il y apporte déjà une amélioration par rapport à ce qui avait été avancé par Newcomb en proposant aussi que cette loi n'est pas seulement applicable à des chiffres significatifs à l'ordre 1, mais aussi à 2, 3, 4 etc ... En effet il donne cette justification :

"L'intervalle logarithmique entre ab et ab + 1[ab étant un nombre composé de a et b des chiffres. ex : a=5 et b =3 alors ab=53]*est log(ab+1)-log(ab), tandis que l'intervalle couverte par les dix chiffres de deuxième place possibles est log (a + 1) - log a. Alors, la fréquence Fb d'un deuxième chiffre b suivant un chiffre de premiere place a est*

$$F_b = \frac{log(\frac{ab+1}{ab})}{log\frac{a+1}{a}}\text{"}[2] \tag{2.1}$$

On obtiens alors la généralisation suivante : *"Il s'ensuit que la probabilité pour un chiffre à la q-ième position est*

$$F_b = \frac{log\frac{abc...p(q+1)}{abc...pq}}{log\frac{abc...o(p+1)}{abc...op}}\text{"}[2] \tag{2.2}$$

Bon nombre d'années après, un financier sud africain a eu l'idée d'employer la loi de Newcomb-Benford afin de donner un premier avis sur la véracité de données statistiques dans le domaine financier. En effet, Mark Nigrini, depuis les années 2000, propose d'employer la loi de Newcomb-Benford pour réaliser une préanalyse de détection des fraudes afin de limiter les coûts d'analyses de fraude.[4][5]

Chapitre 3: Partie Mathématique de la loi

L'un de nos objectifs dans le cadre de notre projet scientifique est de comprendre les fondements mathématiques de la loi de Newcomb-Benford. Nous avons donc réalisé un travail de revue bibliographique avec l'aide de notre enseignant encadrant, Mme ZIDANI, afin de mieux comprendre ce phénomène intriguant. Nous avons notamment étudié en détails un article publié par Vincent et Christian Genest en 2011[8] qui propose plusieurs éléments intéressants :

— Une définition mathématiquement rigoureuse de la loi de Newcomb-Benford comme loi de probabilités

— Une proposition établissant qu'une variable aléatoire dont la fonction de densité vérifie certaines hypothèses tend à suivre cette loi. Dans nos travaux, nous désignons cette proposition "Théoème de Genest".

Cette partie du rapport vise à rendre compte de nos travaux dans ce sens, mais avant toute chose, il faut définir précisément les outils mathématiques dont nous avons besoin :

3.1 Formalisme

Tout d'abord, une variable aléatoire, en probabilités, est une fonction associant une valeur x à tout événement possible ω dans un univers Ω. Concrètement parlant, une variable aléatoire permet de définir les probabilités de chaque événement dans un ensemble d'événements fini ou infini, celui-ci dépendant de ce pour quoi l'on utilise la variable aléatoire.

On considère une variable aléatoire parcourant l'ensemble des réels positifs. Cette variable S peut s'écrire de la forme.

$$S = s \times 10^k \text{ avec } k \in Z$$

s est couramment appelé le significande, et son logarithme en base 10, compris entre 0 et 1 la mantisse de S.

$$\text{Soit } c \in \{1; ...; 10\}.$$

Si l'on considère l'évènement A(c) : le premier chiffre significatif de S est strictement inférieur à c, alors cela revient à savoir si le premier chiffre significatif de s est strictement inférieur à c. Si cet évènement est vérifié, alors $1 \leq s < c$, et en passant au logarithme : $0 \leq log_{10}(s) < log_{10}(c)$ De plus, comme $log(S) = log(s) + k$, on a l'inéquation finale :

$$k \leq log_{10}(S) < log_{10}(c) + k$$

avec k l'ordre de grandeur du nombre qui n'a donc ici que peu d'importance

En prenant cette fois $\epsilon \in [0; 1]$, on pose $P(\epsilon)$ la probabilité que le logarithme décimal de S appartienne à un intervalle $[k; k + \epsilon]$ avec k un entier relatif :

$$P(\epsilon) = P(log_{10}(S) \in \bigcup_{k \in Z} [k; k + \epsilon[)$$

3.2 Définitions mathématiques de la loi

On peut alors introduire la loi de Newcomb-Benford sous sa forme faible, puis forte :

Definition 1

1. Une variable aléatoire S suit la loi faible de Newcomb-Benford ssi,
$$\forall \epsilon \in \{log_{10}(1); ...; log_{10}(10)\}, P(\epsilon) = \epsilon$$
2. Une variable aléatoire S suit la loi forte de Newcomb-Benford ssi,
$$\forall a \in [0; 1], P(\epsilon) = \epsilon$$

où $P(\epsilon)$ représente, d'après le raisonnement précédent, la probabilité que le premier chiffre significatif de X soit inférieur ou égal à 10^{ϵ}. Ainsi, cette définition nous indique - en particulier - que si X suit la LNB, alors la probabilité que le premier chiffre significatif de X soit inférieur ou égal à 3 est égal à log(3).

On voit donc que la version forte de la loi est une généralisation de la version faible, qui est la plus immédiate. Celle-ci permet de ne pas se limiter aux raisonnements en LNB1 (c'est à dire fondés sur la première décimale uniquement) par la suite.

Notons que cette définition est cohérente avec l'équation (1) présentée en introduction et posée par Newcomb déjà. On peut en effet retrouver cette dernière en quelques lignes en remarquant simplement que la probabilité F(c) que le premier chiffre significatif de X soit c s'écrit F(c)= P(log(c+1))-P(log(c))=log(c+1)-log(c)=log(1+1c). Pour tout c entier entre 1 et 9.

3.3 Théorèmes de Genest

On s'intéresse maintenant à mettre en évidence l'apparente vaste répartition de cette loi dans le monde qui nous entoure. En effet, on peut s'apercevoir que les conditions d'apparition de cette loi laissent assez de liberté quant au choix de la variable aléatoire.

Si l'on s'appuie sur la proposition suivante :

Proposition 1 *Soit S une variable aléatoire strictement positive de densité continue. Si la densité f de $log_{10}(S)$ est majorée par une constante M positive et s'il existe K un entier naturel et L un entier relatif tel que f soit d'abord croissante jusqu'à L, puis décroissante à partir de L+2K, alors $|P(a) - a| \leq 2(K + 1)M$*

La démonstration de cette proposition est disponible (Référence 7 page 22) Remarquons que les variables aléatoires qui satisfont de telles hypothèses sont diverses, d'autant plus que le passage au logarithme décimal, fonction croissante et bijective, n'a pas d'influence sur la forme globale de la fonction de densité de S. Ainsi, de nombreuses variables aléatoires valident ces conditions, notamment celles suivant une loi normales ou plus simplement toute variable aléatoire dont les valeurs probables se situent dans un intervalle particulier (la consommation électrique d'un logement comme la taille d'un arbre varieront principalement dans des ordres de grandeur du kWh ou de la dizaine de mètres)

Pour obtenir une variable suivant une loi de Newcomb-Benford, on cherche à faire tendre le produit $2(K + 1)M \rightarrow 0$. Pour cela, on peut imaginer une suite de variables aléatoires strictement positives (X_n) vérifiant toutes les conditions de la proposition précédente. Pour chaque variable X_n, M devient M_n et K devient K_n . Si maintenant la suite K_n est bornée et $M_n \rightarrow 0$ à l'infini, alors pour un N assez grand, $P_n(a) = a$, c'est à dire la variable X_n suit la loi de Newcomb-Benford.

Il est intéressant de remarquer que la variable M est un marqueur de l'étalement de la fonction de densité. En effet, l'aire sous sa courbe devant rester constante et égale à 1, plus M est petit, plus la fonction est étalée et inversement si M est grand. On en déduit alors que plus les valeurs possibles ("probables") de cette variable sont étalées, plus elle peut suivre de près la loi de Newcomb-Benford.
De manière plus visuelle, en se concentrant sur le graphique de la fonction de densité de $log_{10}(S)$, il est fort probable qu'il présente une forme similaire à celle d'un "Chapeau", comme voici :

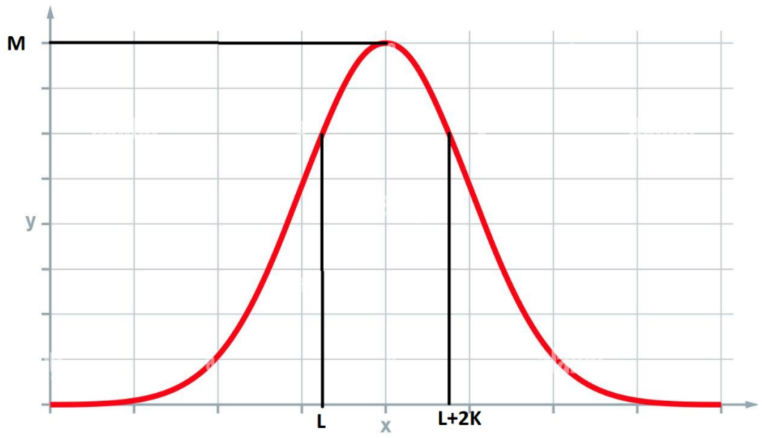


FIGURE 3.1 – Allure d'une fonction de densité de la loi normale, vérifiant les hypothèses du théorème de Genest

Si l'on surligne en noir les bandes sous la courbe concernant les valeurs de S ayant 1 pour premier chiffre significatif, on remarque que l'aire coloriée représente une fraction bien plus grande que seulement 11% de l'aire totale. Cette aire, que l'on peut assimiler à la fréquence d'apparition du 1 en premier chiffre significatif, tend en réalité à atteindre le tiers de la surface totale, (en effet, on considère l'aire entre k et $k + log_{10}(1) = k + 0.301$) notamment lorsque l'on parcourt un grand nombre d'ordres de grandeur.
En effet, le nombre d'ordres de grandeur parcouru représente le nombre de bandes noires sur la figure (x*10 correspond à $log_{10}(S)$+1). Plus on a de bandes noires, plus on se rapproche statistiquement de la loi en réduisant l'impact des variations locales de densité.

3.4 Caractérisation de la Loi par invariance

Dans une publication datée de 1961, Roger Pinkham[3] pose comme prémisse que s'il existe une loi de distribution numérique, alors celle-ci doit être invariante par changement d'échelle.

Il semble en effet trivial, à première vue, que si la longueur des fleuve suit une loi, alors l'unité dans laquelle ces grandeurs sont exprimées ne doit pas avoir d'impact sur l'adéquation de la série à cette loi. Des grandeurs suivant la loi de Newcomb-Benford, après avoir été multipliées par une constante non nulle, continuent de suivre la loi de Newcomb-Benford.

Pour poursuivre notre raisonnement, on introduit la notion de transformation linéaire d'échelle :
Une transformation linéaire d'échelle est une application de la forme $S \to cS$ avec c>0.

Une propriété importante de la loi de Newcomb-Benford est qu'elle est invariante par transformation linéaire d'échelle. C'est notamment ce qui explique que l'unité utilisée n'a pas d'incidence sur le fait qu'une série statistique suit ou non la LNB. De plus, la loi est caractérisée par cette propriété.

Par exemple, si S est une variable aléatoire positive, on a $log_{10}(cS) = log_{10}(c) + log_{10}(S)$ Donc si la fonction de densité de $log_{10}(x)$ satisfait les hypothèses de la Proposition 1 pour un certain couple d'entiers K naturel et L relatif, la densité de $log_{10}(cS)$ y répond également en prenant $K + 1$ et $L + log_{10}(c)$

Proposition 2 *Une v.a. S strictement positive obéit à la version forte de la LNB ssi, pour tout c>0 et tout* $\epsilon \in [0; 1]$*, on a :*

$$P_c(\epsilon) \equiv Pr\{log_{10}(cS) \in M(\epsilon)\} = P(\epsilon)$$

Cette proposition peut être prouvée en montrant d'abord que

$$\forall \delta \in [0; 1], \bigcup_{k \in Z} [k - \delta, k[= \mathbb{R} \diagdown M(1 - \delta)$$

Ce qui entraîne

$$Pr\left\{log_{10}(cS) \in \bigcup_{k \in Z} [k - \delta, k[\right\} = 1 - P(1 - \delta)$$

. Les détails de cette preuve sont présentés dans la publication de Vincent Genest. On établit ainsi que pour tout c>0, le fait que S et cS aient la même probabilité d'avoir des premiers chiffres significatifs inférieurs à 10^{ϵ} entraîne nécessairement que S suit la loi forte de Benford.

Remark 1 Il existe d'autres caractérisations de la LNB. En particulier, Hill a montré que c'était la seule loi invariante par changement de base. Janvresse et de La Rue ont montré que cette loi émerge naturellement dans le cas où les observations sont issues d'un mélange de lois uniformes.

3.5 Loi du i-ième chiffre significatif

Il est également possible, grâce à sa définition forte, de généraliser la loi de Newcomb Benford pour les i premiers chiffres significatifs. Soit l'évènement

"les i premiers chiffres significatifs de X sont $c_1, c_2, c_3, ..., c_i$ *dans l'ordre"*

On peut montrer que la probabilité de cet événement vaut

$$F(c_1, c_2, c_3, ..., c_i) = log_{10}(1 + \frac{1}{c_1 c_2 c_3 ... c_i})$$

où $c_1 c_2 c_3 ... c_i$ représente la concaténation des chiffres $c_1, c_2, c_3, ..., c_i$.

Par exemple, la probabilité que X commence par les chiffres 1, 7, 8 et 2 vaut

$$F(1, 7, 8, 2) = og_{10}(1 + \frac{1}{1782}) \approx 2,436 \times 10^{-4}$$

Une preuve de cette propriété est présentée dans la publication de Vincent Genest[8]. Celle-ci nous permet de développer des analyses plus fines de données, ce qui complique la tâche des fraudeurs désireux de conformer artificiellement leurs données manipulées avec la LNB.

Chapitre 4: Programme de simulation et analyse de données

Dans cette partie, nous étudierons les méthodes informatiques permettant de mesurer la conformité d'une série de données avec la loi de Newcomb-Benford. En particulier, nous nous intéresserons aux tests d'adéquation, qui sont des moyens de quantifier cette conformité par une unique statistique. On peut ainsi dépasser l'aspect arbitraire des comparaisons visuelles basées sur des représentations graphiques.

Dans un second temps, nous commenterons le programme informatique que nous avons mis en place afin de représenter visuellement la conformité de données avec la LNB et de mettre en œuvre ces tests.

4.1 Présentation de quelques tests d'adéquation

4.1.1 Test de la déviation moyenne absolue (MAD)

Le test de la déviation moyenne absolue, ou MAD pour *Mean Absolute Value Deviation*, est couramment utilisé comme indicateur d'adéquation en raison de sa simplicité de calcul. Il s'agit de la moyenne des écarts entre la fréquence attendue et la fréquence observée de chaque chiffre. Formellement, la MAD est obtenue par l'équation suivante :

$$MAD = \frac{1}{K}\sum_{i=1}^{K} |f_{obs,i} - f_{att,i}|$$

où $f_{obs,i}$ représente la fréquence observée du chiffre i, $f_{att,i}$ sa fréquence attendue, et K le nombre de "premiers chiffres possibles". Selon le nombre de chiffres auxquels on s'intéresse, celui-ci peut valoir 9 ou 90 par exemple.

Plus le MAD est faible, plus les données étudiées sont conformes à la loi de fréquences attendues. Cet indicateur permet donc de comparer l'adéquation de deux jeux de données de tailles comparables.

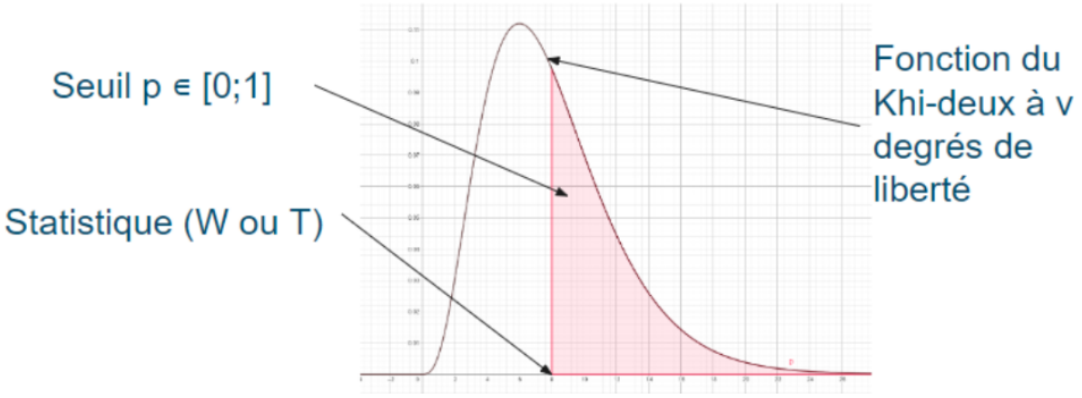
4.1.2 Test du Khi-deux

Le test khi-deux (ou χ^2) consiste à mesurer la distance entre les fréquences observées d'apparition des chiffres significatifs et celles théoriques au moyen de la statistique :

$$W_m = m\sum_{i=1}^{K} \frac{(f_{obs,i} - f_{att,i})^2}{f_{att,i}}$$

Avec m la taille de l'échantillon, $f_{obs,i}$ les fréquences observées et $f_{att,i}$ les fréquences théoriques.

Si l'échantillon étudié s'avère suivre la loi de Newcomb-Benford, alors il est prouvé que la statistique W se comporte comme une variable du khi-deux à q-1 degrés de libertés. Q représente le plus grand nombre pouvant être formé par la suite de PCS sélectionnée (9 si l'on regarde le premier chiffre, 99 si on regarde les deux premiers, etc...). Autrement dit, la fonction khi-deux tend à être la fonction de densité de W lorsque la taille de l'échantillon est suffisamment grande.



La dernière étape est de calculer le seuil p, compris entre 0 et 1, traduisant à partir de ce test la fiabilité du jeu de données. Ce seuil correspond à l'aire de la fonction khi-deux précédente à partir de notre statistique : il correspond à la probabilité d'observer une valeur plus grande que notre statistique. Si celle-ci est faible, nous pouvons discréditer l'hypothèse de la LNB sur le jeu de données.

4.1.3 Smooth Test

Le smooth test quant à lui constitue une alternative au test du khi-deux, plus spécifique à l'hypothèse nulle formulée : l'échantillon suit la loi de Newcomb-Benford. Ce test se fonde sur le théorème suivant :

Théorème 2.1. *Soit $X_1, ..., X_n$ des copies indépendantes d'une variable aléatoire X de densité $f(\cdot)$ par rapport à une mesure ν. Soit $\{h_0(\cdot) \equiv 1, h_k(\cdot), k = 1, 2, ...\}$ une suite de fonctions orthonormales par rapport à $f(\cdot)$; plus précisément, $\int h_k(x)h_{k'}(x)f(x)d\nu(x) = \delta_{kk'}$, la fonction delta de Kronecker. Soit $U_k = n^{-1/2}\sum_{i=1}^{n} h_k(X_i)$ et pour un entier $K \geq 1$, soit $T_K = \sum_{k=1}^{K} U_k^2$. Alors sous H_0, $T_K \xrightarrow{L} \chi^2_K$, la loi khi-deux à K degrés de liberté, et un test de niveau asymptotique α rejette H_0 si la valeur observée de T_K dépasse $x^2_{K,1-\alpha}$, le quantile d'ordre $1-\alpha$ de cette loi χ^2_K.*

Ce test nécessite l'utilisation de fonctions orthonormales à la densité de probabilisé de notre hypothèse. Après quelques recherches, nous avons obtenu une suite de polynômes H_k de degré k vérifiant cette condition, ces fonctions sont bien-sûr des approximations numériques. Il suffit alors de calculer les statistiques U_k présentées ci-dessus en prenant les premiers chiffres significatifs des valeurs de l'échantillon à la place de X_i. Nous nous contenterons ici de k = 2, déjà bien suffisant. Nous réalisons enfin le même protocole d'intégration que le test précédent afin de déterminer un seuil p, en prenant cette fois la fonction du khi-deux à deux degrés de liberté

4.1.4 Conclusion

Il est ainsi intéressant de pouvoir mettre en compétition ces tests. En effet, bien que le khi-deux soit le test de conformité le plus utilisé actuellement pour vérifier la loi de Newcomb-Benford, certaines études tendent à souligner l'efficacité du smooth test. (Cf *Tests d'adéquation de la loi de Newcomb-Benford comme outil de détection des fraudes* [9] pages 39 à 42), ce que nous verrons dans une prochaine partie.

4.2 Commentaires sur la partie informatique

Nous nous penchons dans cette partie à l'étude et l'explication du programme informatique (python) permettant de représenter les données et effectuer les tests.

Différents modes de représentation

Si l'on se concentre d'abord sur le cas n=1, l'histogramme en barre s'avère être un mode de représentation clair et facilement compréhensible. On trouve alors la fréquence d'apparition en ordonnée, en fonction du chiffre significatif en abscisses. Cette fréquence peut alors aisément être comparée à la fréquence théorique tracée juste à côté.

Pour ce qui est des cas où n >1, l'histogramme devient rapidement moins pratique d'utilisation, saturant alors le graphique. Pour y remédier, nous avons opté pour une représentation en nuage de points, sans changer les axes, mais en y ajoutant une courbe continue de tendance. La fréquence théorique est aussi tracée sous forme d'une courbe continue, par soucis de lisibilité.

Explication du programme informatique pour l'affichage graphique

En premier lieu, nous extrayons les données initialement présentes sur un fichier csv, téléchargé dans l'ordinateur, à l'aide de la fonction

```
Extract_Data(filename, columnName)
```

et les *stockons* sous forme d'une liste de valeurs.

Il s'agit alors de transformer cette liste de nombres en une liste " table " de fréquences d'apparition. Ainsi, chaque élément de cette nouvelle liste doit contenir la fréquence observée d'apparition dans les données de son indice en tant que chiffre significatif.

Par exemple : si l'on s'intéresse aux deux premiers chiffres significatifs, le nombre 1578 augmentera la fréquence d'apparition de nombres commençant par 15, renseignée à la 15e position de table.

Pour le cas n = 1 : table comporte 10 éléments de 0 à 9, (0 étant l'emplacement réservé aux nombres que le programme n'a pas pu classer).

Lorsque n = 2, table comporte 100 éléments, de 0 à 99, mais dont seuls les emplacements 10 à 99 servent à calculer les fréquences (0 en 1^{ere} position n'est pas un chiffre significatif.

Plus généralement, table sera une liste à éléments, dont seuls les éléments compris entre et nous intéressent. Pour passer de la liste de nombre à note liste de fréquences, nous utilisons la fonction

```
Create_Table(data,n)
```

qui utilise elle-même la fonction

```
Count_Digits(number, n)
```

chargée d'extraire les n premiers chiffres significatifs d'un nombre.

Pour chaque nombre présent dans la liste de données, on extrait ses n premiers chiffres significatifs et l'on incrémente de 1 la position correspondante dans Table. A la fin, il suffit de diviser chaque élément de Table par le nombre total d'éléments étudiés.

Enfin, il ne reste plus qu'à tracer le graphique correspondant (voir ci-dessus) en utilisant les données de table, et la bibliothèque graphique mathplotlib. Afin de pouvoir comparer les données à la loi de Newcomb-Benford, nous calculons les valeurs théoriques sur le nombre de chiffres significatifs souhaité dans la liste benford, que nous affichons en parallèle sur le graphique. Dans le cas n > 1, nous ajoutons une courbe de tendance dans le but de mieux visualiser le comportement de nos données. Cette courbe est une régression logarithmique base 10 calculée, dans

```
Regression_Log(x,y)
```

à l'aide d'une régressions polynomiales obtenue en utilisant

```
numpy.polyfit()
```

(régression polynomiale à l'ordre 3 de y en fonction de log(x), puis affichage de l'équation y = somme de (coef calculés$\times log(x)^k$))

Génération des tests

La seconde utilité de cette partie correspond à la réalisation de tests. Nous réaliserons ici deux tests présentés ci-dessus :

```
le smooth test, le test du khi deux(smooth_test(data, total)

et (khi_deux(table, total, n)).
```

Dans un premier temps, les statistiques de chacun des tests sont calculées : Pour le khi deux, les fréquences de la liste "tables" et les fréquences théoriques de la loi de Newcomb Benford sont comparées, tandis que le smooth test procède d'abord à une phase d'extraction des premiers chiffres significatifs avant de calculer la statistique. À noter que le smooth test n'est programmé que pour une étude du premier chiffre signigicatif.

```
Les fonctions Khi-Deux(v , w) et h(k : indice du polynôme, x : variable)
```

correspondent respectivement à la fonction du khi-carré(w) à v degrés de liberté et aux fonctions orthonormales nécessaires pour le smooth test.

Le calcul du seuil est effectué en intégrant la fonction du khi deux (au degré de liberté correspondant) à l'aide de la méthode des trapèzes (1000 points) entre 0 et la statistique, puis en prenant le complément à 1 de cette valeur.

Comparaison des tests

Au-delà de la variété de tests que l'on peut développer, il est important de pouvoir mesurer leur efficacité lorsqu'ils sont mis en application. En particulier, ces tests sont sujets à deux grands types d'erreurs : les erreurs de type 1, correspondant à une fraude détectée injustement (faux positif) et les erreurs de type 2, correspondant à des fraudes non détectées (faux négatif). Pour évaluer la puissance de chacun des tests, une méthode consiste à placer tous ces tests au même niveau d'erreur de type 1, puis à générer plusieurs échantillons de données suivant des lois "modifiées". Ce sont des alternatives de la loi de Newcomb-Benford, avec une certaine marge

d'erreur, comme on pourrait retrouver cela dans les données de la vie courante. On mesure alors le taux de rejet du test afin de déterminer cette puissance.

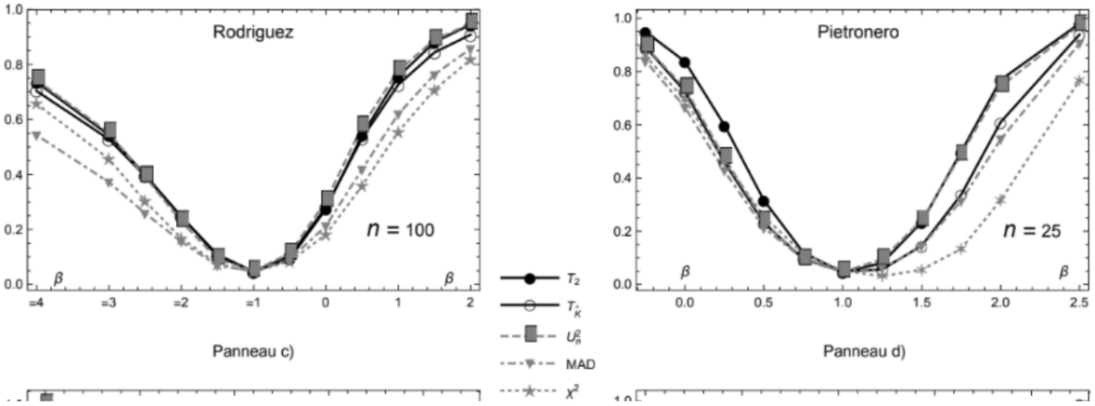
On retrouve de nombreux exemples de ces lois dans le document *Tests d'adéquation de la loi de Newcomb-Benford comme outil de détection des fraudes* [9] (p.39), où différents tests sont notamment comparés, dont les tests du khi deux, smooth test (T2) et MAD (Voir annexe).

Voici pour exemple la famille d'alternatives de Rodriguez :

$$\mathbb{P}[D=d] = p_d^{(Rod)}(\beta) = \begin{cases} \frac{1}{9}(1+\frac{10}{9}\ln(10) + x\ln(x) - (x+1)\ln(x+1)) & \text{si } \beta = 0 \\ \log_{10}[1+1/x] & \text{si } \beta = -1\,, \\ \frac{\beta+1}{9\beta} - ((x+1)^{(\beta+1)} - x^{(\beta+1)})/(\beta(10^{(\beta+1)}-1)) & \text{sinon} \end{cases}$$

Ainsi, lorsque bêta varie, on s'éloigne plus ou moins de la LNB, vérifiée lorsque bêta = 1.

En traçant maintenant les courbes de puissance des différents tests, visualisables dans le même document que précédement[9] (p.40), et dont un extrait est affiché ci-dessous, nous observons une large domination des tests lisses, notés T(k) et T(2), notamment par rapport à des tests comme le khi-deux, ce qui justifie leur utilisation.



Chapitre 5: Applications de la loi

Premiers essais

Dans un premier temps, nous avons tâché de mettre en oeuvre notre programme sur des cas relativement simples pour plusieurs raisons :

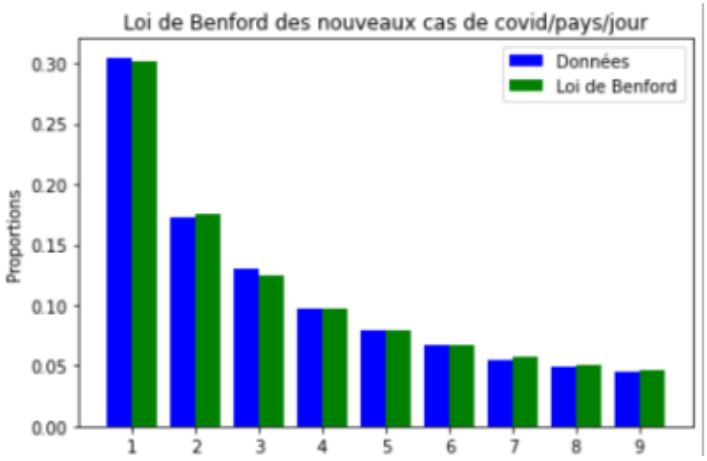
1. Développer notre programme progressivement dans un approche ascendante et descendante
2. Constater par nos propres moyens la validité de la loi de Newcomb Benford

Il s'agissait également de prendre nos marques avec les modalités pratiques des applications, préalablement à des analyses plus fines.

L'un des premiers jeux de données sur lesquels nous nous sommes penchés porte sur le nombre de cas de covid déclarés par pays et par jour depuis le 10 janvier 2020 [1]

Ce premier test permet de mettre en évidence une conformité assez claire de ces données avec la loi de Newcomb-Benford.

Dans la suite de nos travaux, nous avons tenté d'explorer les possibilités d'application de la loi de Newcomb-Benford dans le cadre de la détection de fraudes.



5.1 De l'utilité de détecter les fraudes

Les bases mathématiques de la loi de Newcomb-Benford étant posées, on souhaite mettre en application les possibilités offertes par le phénomène. En particulier, Mark J. Nigrini [4][5], a travaillé en détails sur les applications de la LNB en matière de détection de fraudes. Une idée largement popularisée par ses nombreux travaux veut que les données authentiques tendent à respecter la LNB tandis que les données artificiellement altérées, y compris par des manipulations frauduleuses, n'ont pas de raison particulière de s'y conformer. Ainsi, mesurer l'adéquation d'un jeu de données suspectes avec la LNB est une méthode aisément automatisable et permettant - sans être un élément réellement probant - d'orienter plus efficacement des moyens de recherche humains, plus coûteux à mettre en place.

1. Source des données : ourworldindata.org, sur la base de données de l'OMS.

Par fraude, on entend ici le fait pour un acteur de modifier délibérément une grandeur dans le but d'en tirer un intérêt. Par exemple, le cas d'un.e contribuable ou d'une entreprise qui réduirait artificiellement ses revenus sur sa déclaration fiscale afin d'être moins imposée. Les enjeux économiques de la détection de fraude sont évidemment de première importance. En France, le seul organisme se risquant à donner une estimation du coût total de la fraude fiscale est le syndicat Solidaires Finances Publiques, qui avance le chiffre de 60 à 80 milliards d'euros annuels[10]. Pour information, cela représente entre 11 et 15% de l'ensemble des recettes fiscales perçues par l'état en 2022[11]. Les entreprises peuvent également avoir intérêt à mentir dans leur déclarations financières en majorant leurs actifs et minorant leurs passifs dans le but de séduire des investisseurs. Elles peuvent également falsifier des données techniques pour faire croire qu'elles respectent les normes sociales, environnementales ou de sécurité en vigueur. Il existe donc un enjeu politique à la détection de fraudes, qui peut aussi prendre la forme de la détection de fraudes électorales.

5.2 Démarche et méthodologie

Forts des notions mises en place et des résultats acquis dans les parties précédentes, nous nous sommes proposés d'explorer les possibilités d'application de la loi de Newcomb-Benford dans un cas pratique : la détection de fraudes fiscales.

A l'image de travaux publiés par Marcel Ausloos en 2017[12], nous avons décidé d'analyser des données disponibles sur le site du ministère italien de l'économie et des finances, correspondant au revenu déclaré dans chacune des près de 8 000 communes italiennes entre 2007 et 2011 dans le cadre de l'imposition sur le revenu des particuliers. Comme dans cette publication, nous avons mené une étude comparative entre les 20 régions dans l'espoir de mettre en lumière des différences de conformité d'une région à l'autre, qui seraient susceptibles de révéler des tendances structurelles à la fraude.

Étant donnée du faible nombre de données pour certaines régions, il nous semble plus pertinent d'en rester à une analyse de type LNB1, c'est-à-dire portant sur les premiers chiffres significatifs seulement.

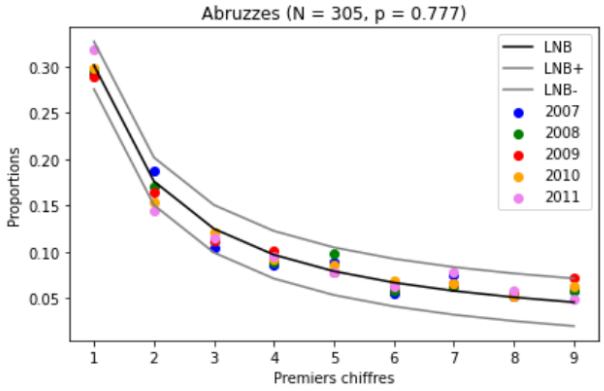
Pour une région donnée, nous représentons visuellement les fréquences d'apparition de chaque chiffre significatif dans les graphes usuels, au moyen de points dont la couleur est associée à une année. On représente aussi par une ligne noire les fréquences théoriques attendues par la loi de Newcomb-Benford. Cette dernière étant une loi tendancielle, on s'attend à ce que les plus grandes régions (sur lesquelles on a le plus de données) soient naturellement plus resserrées autour des fréquences attendues.

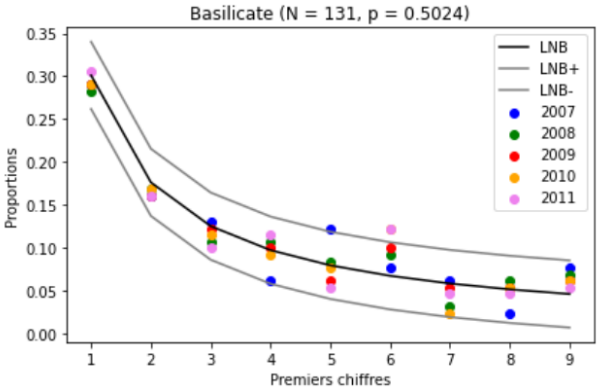
Pour pouvoir comparer visuellement la conformité des données indépendamment de la taille de la région, on définit une barre d'erreur standard de l'échantillon dépendant du nombre de communes N de la région comme suit : $\sigma = \frac{1}{\sqrt{5N-1}}$. On trace les courbes BL1+ et BL1- délimitant cet intervalle de confiance heuristique. Le 5 représente les cinq années qui sont étudiées (entre 2007 et 2011 inclus).

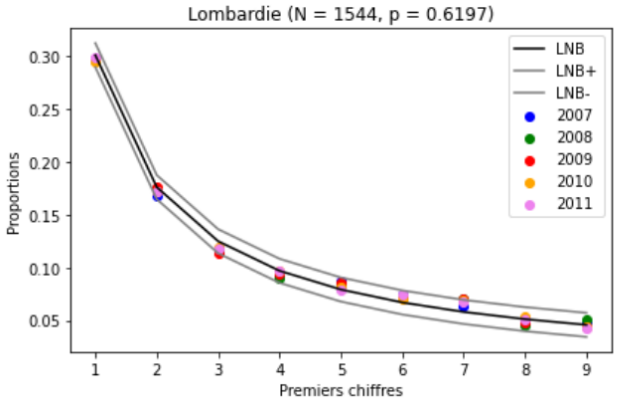
Pour quantifier la conformité de ces données avec la loi de Benford, nous choisissons d'utiliser le test du khi-deux, qui est la méthode de référence dans la littérature et également celle utilisée par Marcel Ausloos. A la différence de ce dernier, nous calculons cependant le seuil de tolérance p du test du khi-deux qui est un indicateur de conformité plus discriminant que la simple statistique W^2. En pratique, puisque les données sont séparées par année, on calcule le seuil p pour chaque année et on en fait la moyenne temporelle.

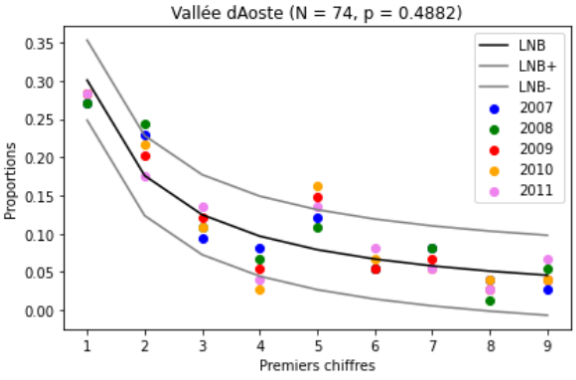
2. Cf. partie 4.1.2

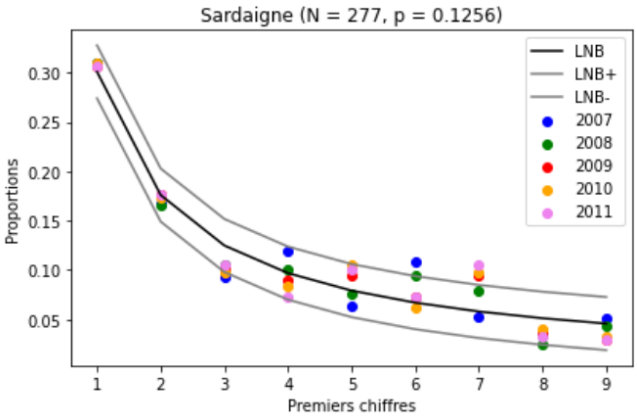
5.3 Résultats

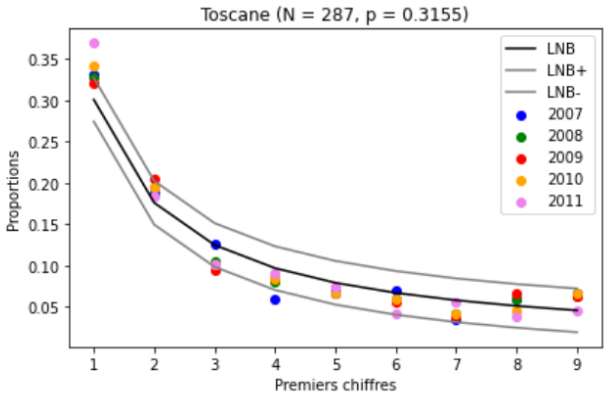


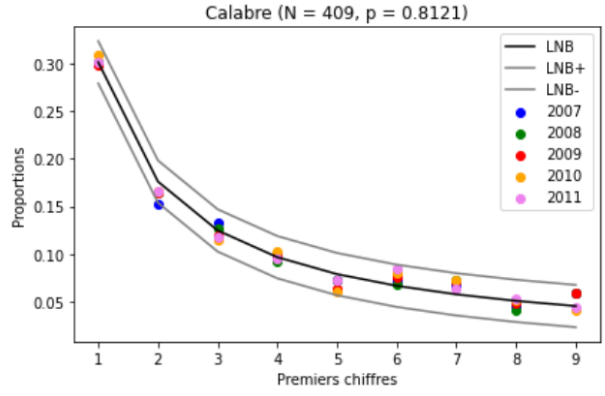


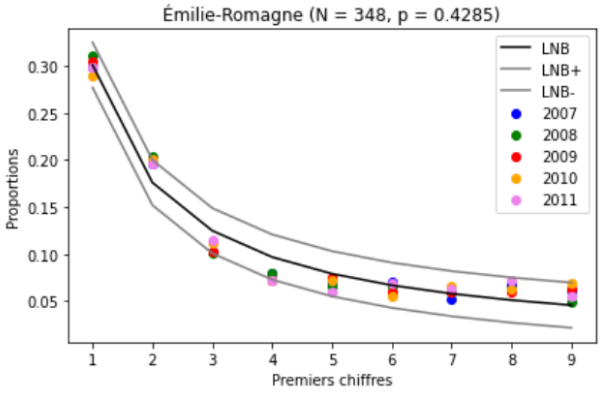












Le but de cette étude était de nous intéresser à quelque chose de plus usuel. En effet, la loi de Newcomb-Benford a pu être utilisée par le passé afin de limiter les enquêtes fiscales qui sont très coûteuses. Ici le but était de découvrir quelles régions étaient les plus éloignées du modèle théorique afin d'orienter au mieux une possible enquête comme le ferait un contrôleur fiscale ou un économiste.

Ce que l'on constate graphiquement c'est que les graphes d'une année à l'autre sont très différents. Certains respectent très bien la LNB avec des seuils avoisinant 0,8. D'autres ont des seuils très faibles, moins de 0,1. Cependant il reste difficile d'affirmer si une de ces régions fraude. Comme on peut le voir certaines régions n'ont que peu de communes, moins de 300 pour certaines. Ainsi même si les tests ne semblent pas cohérent, nous émettons des réserves quant aux supposées fraudes.

En clair, la LNB est un outil puissant permettant de nous indiquer si la théorie est suivie en pratique. Elle ne reste cependant qu'un indice mathématique, ce n'est pas une science exacte.

On peut synthétiser les résultats obtenus pour l'ensemble des vingt régions italiennes ainsi :

TABLE 5.1 – **Résultats de nos analyses par région italienne**

Région	**N**	**p**
Campanie	551	0,0543
Ligurie	235	0,0837
Sardaigne	277	0,1256
Molise	136	0,186
Ombrie	92	0,1874
Toscane	287	0,3155
Marches	239	0,4047
Emilie-Romagne	348	0,4285
Frioul-Vénétie julienne	218	0,4597
Vallée d'Aoste	74	0,4882
Basilicate	131	0,5024
Vénétie	581	0,5095
Sicile	390	0,6065
Lombardie	1544	0,6197
Pouilles	258	0,6585
Piémont	1206	0,7251
Trentin-Haut-Adige	333	0,7495
Latium	378	0,7522
Abruzzes	305	0,777
Calabre	409	0,8121

Chapitre 6: Conclusions et perspectives

Pendant un semestre, nous avons pu étudié de façon approfondie la loi de Newcomb-Benford. Notre compréhension de ce phénomène a évolué, aussi bien par l'étude des étapes historiques de sa formalisation qu'à travers nos recherches bibliographiques sur sa justification mathématique.

Sur la base de ces travaux, nous avons décidé de reproduire une analyse issue de la littérature sur un cas d'application en matière de détection des fraudes.

Tout comme ceux de la publication dont nous nous sommes inspirés, nos résultats mettent en lumière des véritables différences entre les régions. Ils semblent indiquer que la loi de Newcomb-Benford, sans être un indicateur réellement probant, permet d'orienter efficacement des moyens d'investigation plus sophistiqués et plus coûteux.

Ainsi, ce projet nous a non seulement permis de mettre en pratique des connaissances acquises au cours de notre formation (en cours de probabilités, ou en programmation par exemple) et d'exercer notre esprit critique, mais aussi de développer de nouvelles compétences. Nous avons, entre autres, appris à rechercher et lire des publications académiques mathématiques, à organiser notre travail en équipe, à se répartir les tâches, etc.

Pour approfondir ces analyses, on peut envisager d'utiliser d'autres types de tests plus puissants comme par exemple le smooth test (Cf Partie 4.1.3). On peut également mettre en place des études plus fines se basant non seulement sur le premier chiffre significatif, mais aussi sur les deux ou trois premiers. Enfin, on peut remarquer que les champs d'application potentiels de la loi de Newcomb-Benford en détection des fraudes sont nombreux. Nous nous sommes ici intéressés à des données financières, mais on pourrait également se pencher sur des données électorales, techniques ou scientifiques.

Bibliographie

[1] Simon Newcomb, *Note on the frequency of use of the different digits in natural numbers*, *American Journal of Mathematics*, Vol. 4, No. 1 (1881).

[2] Frank Benford, *The Law of Anomalous Numbers*, American Philosophical Society, *Proceedings of the American Philosophical Society*, Vol. 78, No. 4, 1938.

[3] Roger S. Pinkham, *On the Distribution of First Significant Digits.*, The Annals of Mathematical Statistics, vol. 32, no. 4, 1961, pp. 1223–30.

[4] Mark J. Nigrini, *Benford's Law - Applications for Forensic Accounting, Auditing and Fraud Detection*, 9 mars 2012.

[5] P. Drake, Mark J. Nigrini, *Computer assisted analytical procedures using Benford's Law*, Journal of Accounting Education, Volume 18, Issue 2, 2000

[6] Theodore P. Hill, Arno Berger, *An Introduction to Benford's Law*, Princeton University Press, 2015.

[7] Theodore P. Hill, Arno Berger, *The mathematics of Benford's law : a primer*, Statistical Methods & Applications, 2020, URL.

[8] Vincent et Christian Genest, *La loi de Newcomb-Benford ou la loi du premier chiffre significatif*, Association mathématique du Québec, *Bulletin AMQ*, Vol. LI, No 2, mai 2011.

[9] Vovor-Dassu Komlavi, *Tests d'adéquation à la loi de Newcomb-Benford comme outils de détection de fraudes.*, Thèse de doctorat, [Montpellier, France] Université Montpellier 2, décembre 2021, URL : https ://theses.hal.science/tel-03595714v1.

[10] Caroline Félix, *Le Vrai du Faux. Quels sont les chiffres des fraudes fiscales et sociales en France ?*, francetvinfo.fr, 20 avril 2023.

[11] DGFip, *Les recettes fiscales budgétaires collectées par la DGFiP en 2022*, Lien, décembre 2023.

[12] Marcel Ausloos, Roy Cerqueti et Tariq A. Mir , *Data science for assessing possible tax income manipulation : The case of Italy* , Chaos, Solitons & Fractals, Volume 104, 2017,Pages 238-256

Chapitre 7: Annexes

Annexe : Démonstration de la proposition dans l'article de Genest [8]

Soit X un aléa strictement positif de densité continue. Supposons que la densité f de $Y = log_{10}(X)$ soit majorée par une constante M > 0 et admettons qu'il existe des entiers $K \in \mathbb{N}$ et $L \in \mathbb{Z}$ tels que f soit croissante sur $(\inf, L]$ et décroissante sur $[L + 2K, +\inf)$. L'objectif de cette annexe est de démontrer que pour tout $\epsilon \in [0, 1]$, on a $|P(\epsilon) - \epsilon| \leq 2(K + 1)M$.

L'énoncé étant trivial lorsque $\epsilon = 0$, fixons $\epsilon \in [0, 1]$. Pour tout entier k < L, la croissance de f sur l'intervalle $(k, k + 1)$ entraîne que

$$\frac{1-\epsilon}{\epsilon}\int_k^{k+\epsilon} f(y)\,dy \leq \int_k^{k+\epsilon} f(k+\epsilon)\,dy$$

$$= (1-\epsilon)(k+\epsilon) = \textstyle\int_{k+\epsilon}^{k+1} f(k+\epsilon)\,dy \leq \int_{k+\epsilon}^{k+1} f(y)\,dy$$

En remplaçant le terme d'extrême gauche de l'inéquation ci-dessus par

$$\frac{1}{\epsilon}\int_k^{k+\epsilon} f(k+\epsilon)\,dy - \int_k^{k+\epsilon} f(k+\epsilon)\,dy$$

on déduit que

$$\frac{1}{\epsilon}\int_k^{k+\epsilon} f(y)\,dy \leq \int_k^{k+\epsilon} f(y)\,dy + \int_{k+\epsilon}^{k+1} f(y)\,dy = \int_k^{k+1} f(y) \tag{7.1}$$

De même, si k > L + 2K, la décroissance de f sur l'intervalle $(k+\epsilon-1, k+\epsilon)$ permet d'écrire

$$\frac{1-\epsilon}{\epsilon}\int_k^{k+\epsilon} f(y)\,dy \leq \frac{1-\epsilon}{\epsilon}\int_k^{k+\epsilon} f(k)\,dy$$

$$(1-\epsilon)f(k) = \int_{k-1+\epsilon}^{k} f(k)\,dy \leq \int_{k-1+\epsilon}^{k} f(y)\,dy$$

d'où l'on tire

$$\frac{1}{\epsilon}\int_k^{k+\epsilon} f(y)\,dy \leq \int_{k-1+\epsilon}^{k+\epsilon} f(y)\,dy \leq \int_{k-1}^{k} f(y)\,dy \tag{7.2}$$

où la dernière inégalité est justifiée par le fait que f est décroissante sur l'intervalle $(k-1, k)$ Par ailleurs, si $k \in L, ..., L + 2K$, on a :

$$\frac{1}{\epsilon}\int_k^{k+\epsilon} f(y)\,dy \leq M. \tag{7.3}$$

En combinant les inégalités (7.1), (7.2) et (7.3) on trouve

$$\frac{P(\epsilon)}{\epsilon} = \frac{1}{\epsilon}\sum_{k\in\mathbb{Z}}\int_k^{k+\epsilon} f(y)\,dy \leq (2K+1)M + \int_{\mathbb{R}\setminus[L,L+2K]} f(y)\,dy \leq (2K+1)M + 1,$$

puisque f est une densité. Il s'ensuit que

$$P(\epsilon) \leq \epsilon + (2K+1)M\epsilon \leq \epsilon + (2K+1)M$$

et donc que $P(\epsilon) - \epsilon \leq (2K+1)M \leq 2(K+1)M$.

La démonstration de l'inégalité $\epsilon - P(\epsilon) \leq 2(K+1)M$ est analogue. Pour tout entier $k < L$, la croissance de f sur l'intervalle $(k-1+\epsilon, k+\epsilon)$ fait en sorte que

$$\frac{1-\epsilon}{\epsilon}\int_k^{k+\epsilon} f(y)dy\,dy \geq (1-\epsilon)f(k) \geq \int_{k-1+\epsilon}^{k} f(y)\,dy$$

Par conséquent,

$$\frac{1}{\epsilon}\int_k^{k+\epsilon} f(y)\,dy \geq \int_{k-1+\epsilon}^{k+\epsilon} f(y)\,dy \geq \int_{k-1}^{k} f(y)\,dy. \tag{7.4}$$

De façon semblable, si $k > L+2K$, la décroissance de f sur l'intervalle $(k-1,k)$ conduit à

$$\frac{1-\epsilon}{\epsilon}\int_k^{k+\epsilon} f(y)\,dy \geq (1-\epsilon)f(k+\epsilon) \geq \int_{k+\epsilon}^{k+1} f(y)\,dy,$$

d'où il découle que

$$\frac{1}{\epsilon}\int_k^{k+\epsilon} f(y)\,dy \geq \int_k^{k+1} f(y)\,dy. \tag{7.5}$$

En se servant des inégalités (7.4) et (7.5), on conclut que

$$\begin{aligned}\frac{P(\epsilon)}{\epsilon} &\geq \sum_{k=-\inf}^{L-1}\int_{k-1}^{k} f(y)\,dy + \sum_{k=L+2K+1}^{\inf}\int_k^{k+1} f(y)\,dy \\ &= 1 - \int_{L-1}^{L+2K+1} f(y)\,dy \geq 1 - 2(K+1)M\end{aligned}$$

pour toute valeur possible de $\epsilon \in (0,1]$. Ceci achève la démonstration.

Nous tenons à noter que l'entièreté de cette démonstration est tirée de l'article écrit par les frères Genest [8] que nous avons retranscrite par soucis de facilité de lecture.

FIGURE 1 – Organisation du groupe dans le projet

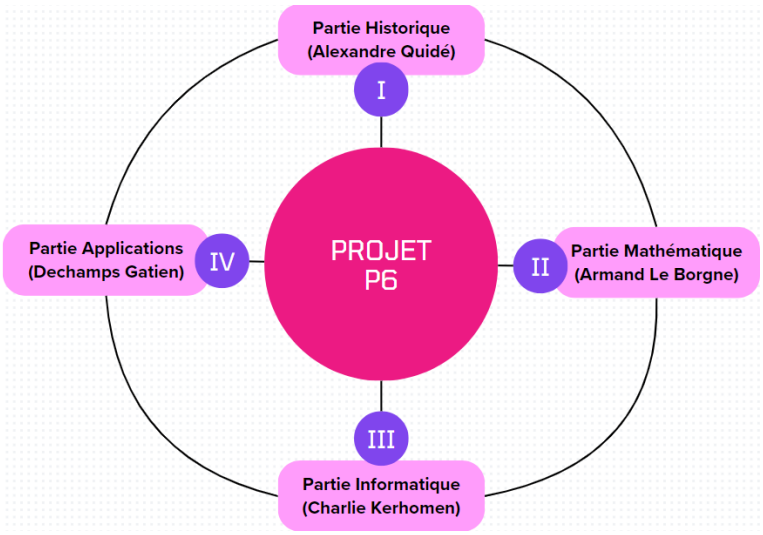


TABLE 1 – **Tableau des probabilités d'apparition de chaque chiffre significatif, jusqu'au deuxième chiffre | Newcomb (à gauche) et Benford (à droite)**

Chiffre	*1er chiffre*	*2e chiffre*
0		0.1197
1	0.3010	0.1179
2	0.1761	0.1088
3	0.1249	0.1043
4	0.0969	0.1003
5	0.0792	0.0967
6	0.0669	0.0934
7	0.0580	0.0904
8	0.0512	0.0876
9	0.0458	0.0850

Chiffre	*1er chiffre*	*2e chiffre*
0	0.000	0.120
1	0.301	0.114
2	0.176	0.108
3	0.125	0.104
4	0.097	0.100
5	0.079	0.097
6	0.067	0.093
7	0.058	0.090
8	0.051	0.088
9	0.046	0.085