section

plain definitionremarkplain

LU3MA201 Projet

Rapport final:
La loi de Benford

Avril 2022

Auteurs: Ahmed Mansour Bourassine (28610115)
ahmed.bourassine@etu.sorbonne-universite.fr

Thomas Dittmar (21102938)
thomas.dittmar@etu.sorbonne-universite.fr

Encadrante: Anna Bonnet

Contents

1 Introduction

Notre projet concerne la loi de Benford, qui décrit une distribution inattendue des premiers chiffres pour certaines sources de nombres dans la vie réelle. Cette distribution se produit notamment lorsque les chiffres représentent plusieurs ordres de grandeur et, dans notre cas, dans les bilans et les données d'entreprise. Notre objectif est de détecter les fraudes financières avec la loi de Benford qui a la formule suivante:

$$\forall k \in \{1, 2, ..., 9\} : p(k) = \log_{10}(1 + \frac{1}{k})$$

où $p(k)$ est la probabilité que le premier chiffre significatif est k.

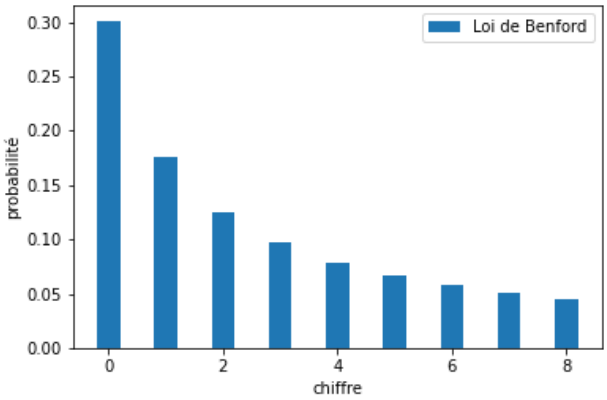


Figure 1: Distribution de la loi de Benford

1.1 Contexte historique

La découverte a eu lieu deux fois. La première fois par l'astronome canadien Simon Newcomb en 1881qui a remarqué que les premières pages des livres de logarithmes étaient beaucoup plus usées que les dernières. Il a proposé la formule ci-dessus dans l'American Journal of Mathematics, mais n'a pas reçu beaucoup d'intérêt. Près de cinquante années plus tard Frank Benford, un physicien américain, a redécouvert cette distribution et elle a été nommée en son honneur. [hist-1]
La détection de fraude est l'un des applications les plus connus de la loi de Benford. La loi est utilisée par exemple par l'IRS (l'agence du gouvernement fédéral des États-Unis qui collecte l'impôt) pour détecter des invraisemblances dans des bilans comptables, et détecter des chiffres manipulés. La loi est considéré par quelques experts comme un outil de comptabilité forensique (juricomptabilité), notamment Mark Nigrini. [appl-1]

1.2 Problématique pour les mathématiques

La problématique qu'on s'est posée dans ce projet est assez claire: Pour des données de déclaration de revenue fiscal, est-il possible de détecter d'éventuelles fraudes et avec quelle précision?
Le problème que nous essayons principalement de résoudre est de savoir si les données doivent suivre la distribution de Benford ou non. Si c'est le premier cas, notre approche est très efficace, mais si les données ne convergent pas naturellement en loi vers la distribution théorique, nous ne pouvons rien faire.
Pour ce faire, nous déterminons la distribution des chiffres dans nos données et nous la comparons à la distribution attendue pour la loi de Benford. Nous pouvons ainsi savoir quelle est la probabilité que nos données suivent la loi.

2 Test statistique

2.1 Hypothèse

Le but d'un test statistique est de décider entre deux hypothèses, l'hypothèse nulle H_0 et l'hypothèse alternative H_1. Pendant le test, on cherche à savoir quelle est la probabilité qu'une distribution sous l'hypothèse H_0 a une distance plus grande par rapport à l'espérance sous H_0, que la distance entre les données et l'espérance pour H_0. Nous appelons cette probabilité la valeur p. Si cette valeur p est inférieure à une valeur *alpha*, appelée risque de première espèce, nous rejetons l'hypothèse.
Notre hypothèse H_0 est que les données suivent la loi de Benford et notre H_1 est que les données ne suivent pas la loi.

2.2 Modèle

Pour un vecteur aléatoire X (représentant le premier chiffre significatif) qu'on a supposé qu'il suit une loi inconnu L, (plus de détaille sur les hypothèses dans la partie 2.2) la question qu'on se pose est "Est-ce que L est la loi de Benford ?".

Formalisme

Soit $X \sim L$ un vecteur aléatoire qui prend le premier chiffre significatif i d'un nombre et envoie un vecteur de N^9 avec 1 dans la i ème composante du vecteur.

$$\begin{array}{rccc} X & : & \Omega & \to & \mathbb{N}^9 \\ & & PCS* & \mapsto & (0,0,..0,1,0,..,0) \end{array}$$

$PCS*$: premier chiffre significatif

Notre hypothése est que $L =$ Loi de Benford.
Par loi forte de grand nombre si on pose $S_N = \sum_{i=1}^{N} X_i$ on aura :

$$\lim_{N\to\infty} \frac{S_N}{N} \xrightarrow{presque-s\hat{u}r} (\log(2), \log(\frac{3}{2}), \log(\frac{4}{3}), ..., \log(\frac{10}{9}))$$

Le but est de calculer, avec N le cardinal de l'échantillon, cette probabilité noté $\mathscr{A}(x)$:

$$\mathscr{A}(x) = \mathbb{P}(d(S_N, \mathscr{C}_{th}^N) \geq x) = ?$$

$\mathscr{C}_{th}^N$: courbe (ou graphe) théorique pour un échantillon de taille N (vecteur de $\mathbb{R}^9$).
$d()$: une distance qu'on va définir plus en détaille dans la suite (partie 2.3).

On peut même normaliser vu que la distance utilisé (qu'on discutera dans une partie à part dans le rapport) est obtenu par proportionnalité de chaque coordonné des deux graphes. on obtient donc:

$$\mathbb{P}(d(\frac{S_N}{N}, V_{LB}) \geq x) =?$$

avec V_{LB} le vecteur de loi de Benford normalisé $(\log(2), \log(\frac{3}{2}), \log(\frac{4}{3}), ..., \log(\frac{10}{9}))$.

2.3 Statistique de test

On avait besoin de définir une distance pour mesurer combien un graphe empirique est "éloigne" de la distribution théorique. Ainsi on propose de travailler avec la distance suivante :

$$\begin{array}{rccc} d & : & \mathbb{R}^9 \times \mathbb{R}^9 & \to & \mathbb{R}_+ \\ & & (v_1, v_2) & \mapsto & d(v_1, v_2) \end{array}$$

avec :

$$d(\mathscr{G}_{emp}, \mathscr{G}_{th}) = \sum_{i=1}^{9} max(|\frac{\mathscr{G}_{emp}^{(i)}}{\mathscr{G}_{th}^{(i)} + 1} - 1|, |\frac{\mathscr{G}_{th}^{(i)}}{\mathscr{G}_{emp}^{(i)} + 1} - 1|)$$

$\mathscr{G}_{th}^{(i)}$: occurence théorique de i
$\mathscr{G}_{emp}^{(i)}$: occurence empérique de i

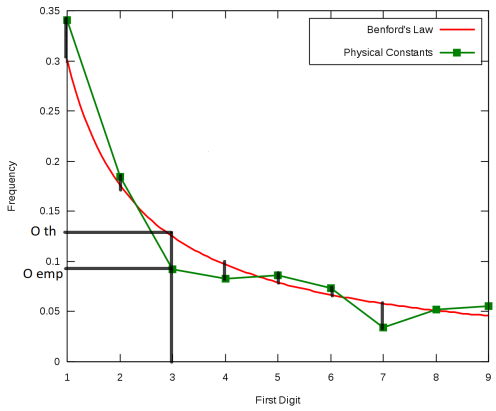


Figure 2: Exemple de calcul de la distance.

pour $i = 3$:

$$|\frac{\mathscr{G}_{th}^{(3)}}{\mathscr{G}_{emp}^{(3)}+1} - 1| \sim |\frac{0.125}{0.095} - 1| = |1.32 - 1| = 0.32$$
$$|\frac{\mathscr{G}_{emp}^{(3)}}{\mathscr{G}_{th}^{(3)}+1} - 1| \sim |\frac{0.095}{0.125} - 1| = |0.76 - 1| = 0.24$$

Donc la distance est de 0.32 en $i = 3$; il faut calculer aussi pour le restant des chiffres.

Comparaison des tests

Afin d'améliorer la statistique de test, on pensait à appliquer une autre façon de calculer la distance entre $\mathscr{C}_{emp}$ et $\mathscr{C}_{th}$ (les courbe).

La nouvelle distance qu'on va tester avec est la distance euclidienne alors qu'avant on testait avec la distance de Manhatan:

$$\begin{array}{rccc} d.2 & : & \mathbb{R}^9 \times \mathbb{R}^9 & \to & \mathbb{R}_+ \\ & & (v_1, v_2) & \mapsto & d.2(v_1, v_2) \end{array}$$

avec :

$$d.2(\mathscr{G}_{emp}, \mathscr{G}_{th}) = \sqrt{\sum_{i=1}^{9} max(|\frac{\mathscr{G}_{emp}^{(i)}}{\mathscr{G}_{th}^{(i)} + 1} - 1|, |\frac{\mathscr{G}_{th}^{(i)}}{\mathscr{G}_{emp}^{(i)} + 1} - 1|)^2}$$

On peut remarquer après l'implémentions que la distance $d.2$ est un peu moins pénalisante pour les déviations que d mais pas de façon significatif. On va mesurer le rapport entre la distance maximale générée et la distance de notre graphe empirique. Avec d on obtient 1.15 alors qu'avec $d.2$ on obtient 1.18 (une différence de 3%).
Nous avons décidé qu'une modification n'était pas nécessaire.

2.4 Région de rejet

Les régions de rejet sont les valeurs de la distance pour lesquelles nous rejetons l'hypothèse H_0. Cette région peut être définie par la valeur α, dans ce cas nous comparons notre α avec la valeur p comme décrit au début et rejetons notre hypothèse si la valeur p est inférieure à α. Le choix de α conditionne la probabilité de deux erreurs considérées dans la discussion et doit donc être réfléchir en conséquence, $\alpha = 0.05$ est un choix fréquent. Pour déterminer ce p, nous devons par la suite déterminer la distribution de la probabilité pour les différentes distances.

Approche théorique

(On reprend les mêmes notation de la partie 2.1)
On peut remarquer que pour un certain N, $S_N = \sum_{i=1}^{N} X_i$ suit la loi multinomiale $(N, 9)$ car en partie:

$$\mathbb{P}(S_N = (n_1, n_2, ..., n_9)) = \mathbb{P}(N_1 = n_1, N_2 = n_2, ..., N_9 = n_9)$$

On a $\sum_{i=1}^{9} n_i = N$ et $p_k = \log(1 + \frac{1}{k})$ avec les N_i , $i \in (1, 2, ..., 9)$ qui correspondent aux probabilités p_i . alors d'après la loi multinomiale :

$$\mathbb{P}(N_1 = n_1, N_2 = n_2, ..., N_9 = n_9) = \frac{N!}{\prod_{i=1}^{9} n_i!} \cdot \prod_{i=1}^{9} p_i^{n_i}.$$

On déduit donc à l'aide de cette formule que :

$$\mathbb{P}(S_N = (n_1, n_2, ..., n_9)) = N! \cdot \prod_{i=1}^{9} \frac{p_i^{n_i}}{n_i!}$$

En bref, c'est la probabilité d'obtenir un certain vecteur (ou graphe) pour un échantillon de cardinal N d'un jeu de donné qui suit la loi de Benford. On peut donc maintenant déduire une formule générale pour $\mathscr{A}(x)$ avec $x \geq 0$:

$$\mathscr{A}(x) = \sum_{S_N \in \mathscr{D}_x} \mathbb{P}(S_N)$$

avec $\mathscr{D}_x = \{\ S_N \in [0, .., N]^9 \mid d(S_N, \mathscr{C}_{th}^N) \geq x\ \}$

Naturellement on se pose la question: Peut-on faire ce calcul explicitement?
La formule direct de $\mathbb{P}(S_N)$ est plutôt possible et ne coûte pas trop cher de point de vue de calcul pour l'ordinateur (calcul simple et explicite). Le problème est l'ensemble $\mathscr{D}_x$: Il faut faire des comparaisons pour tous les S_N possibles ce qui peut s'avérer vraiment coûteux pour N assez grand. Dans la suite on va faire un petit détour sur le combinatoire pour essayer de déterminer explicitement le cardinal maximum de $\mathscr{D}_x$.

Pour $\mathbb{P}_0$ (le cas ou $x = 0$) $\mathscr{D}_x$ atteint son cardinal maximal :
Toutes les distribution possible de S_N sont dedans. On doit calculer toutes les distributions possible de N objet identique sur 9 places distinctes. En fixant chaque fois le nombre de "objet" dans les cases avant, on calcule les possibilité de combinaison pour le reste. la somme se fait que sur 8 place et pas 9 car le choix de la dernière case est déterministe. En écrivant explicitement en forme de somme on obtient :

$$|\mathscr{D}_0| = \sum_{a=0}^{N} \sum_{b=0}^{N-a} \sum_{c=0}^{N-(a+b)} ..... \sum_{h=0}^{N-(a+...+g)} N - (a + b + ... + h)$$

On peut trouver une formule beaucoup plus pratique avec un dénombrement assez astucieux du l'espace fonctionnel discret ($\mathscr{F} : \{1, ..., N\} \to \{1, ..., 9\}$).

Effectivement, on trouve cette formule : $|\mathscr{D}_0| = \binom{N+8}{8}$. En regardant la formule qu'on a trouvé on peut se rendre compte que $|\mathscr{D}_0| = \mathcal{O}(N^8)$.
Donc l'algorithme va avoir une complexité de $\mathcal{O}(N^8)$ vu que les comparaison et le calcul de $\mathbb{P}(S_N)$ sont tous de l'ordre de $\mathcal{O}(1)$. Au début on se rend pas compte du problème mais en prenant un exemple on se rend compte de l'ordre du grandeur du calcul à faire.
Prenons $N = 10000$ (cardinal d'un jeu de données qu'on travaille avec pour notre code), alors on aura:
$N_{calcul} = C_0 \cdot (10000)^8 = C_0 \cdot 10^{36}$ en fixant $C_0 = 10^{-5} \approx \frac{1}{8!}$ on obtient $N_{calcul} = 10^{31}$.

Le problème c'est que même si on effectue le calcul avec TaihuLight Sunway (le plus fort ordinateur en 2017) avec une puissance de calcul de 10^{17} cela nous prendra 3000000 ans !

Clairement cette méthode d'approche théorique est assez problématique et trop inefficace. Ce qui est plus pratique, pour N assez grand, est d'approximer la distribution par une approche expérimentale.

Approche expérimentale

Pour l'approche expérimentale, nous avons essayé d'approximer la distribution des distances sous l'hypothèse H_0. Pour ce faire, nous avons utilisé une méthode similaire à la méthode de Monte-Carlo, dans laquelle un problème analytique est approché à l'aide d'échantillons. [MC-1] La loi des grands nombres est invoquée pour justifier une convergence vers la solution exacte. Nous approchons ainsi $\mathscr{A}(x)$ en regardant combien de nos échantillons tombent d'un côté ou de l'autre de x.
Nous avons généré pour plusieurs N différents, compris entre 0 et 100.000, un millier de jeux de données qui suivent la loi de Benford et calculé leurs distances. Si nous voulons analyser des données, nous comparons la distance des données avec les distances du N le plus proche. Nous obtenons ainsi une approximation de $\mathscr{A}(x)$ qui s'améliore à mesure que nous générons initialement plus d'ensembles de données.
L'avantage des distributions comparatives calculées au préalable par rapport à un calcul pour chaque test individuel est que le résultat reste constant et reproductible. Si les distributions de comparaison étaient recréées pour chaque test, il est possible que des tests répétés changent le résultat.

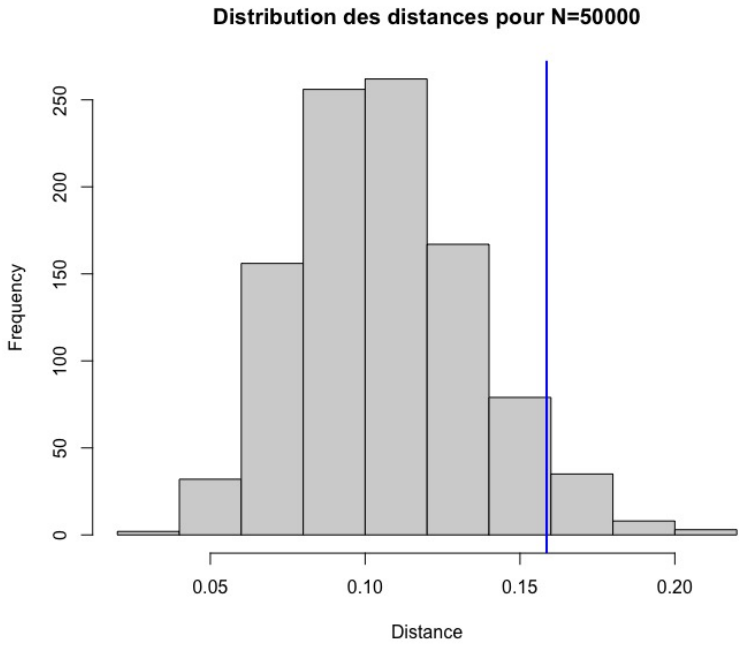


Figure 3: Distribution exemplaire des mille distances pour $N = 5000$, la ligne bleue marque les 5%.

3 Discussion et autres réflexions

3.1 Utilisation

Maintenant, nous pouvons enfin évaluer nos données. Nous prenons comme exemples la suite des puissances de 2 et le PIB[pib-1] des pays du monde. Les résultats sont les suivants :

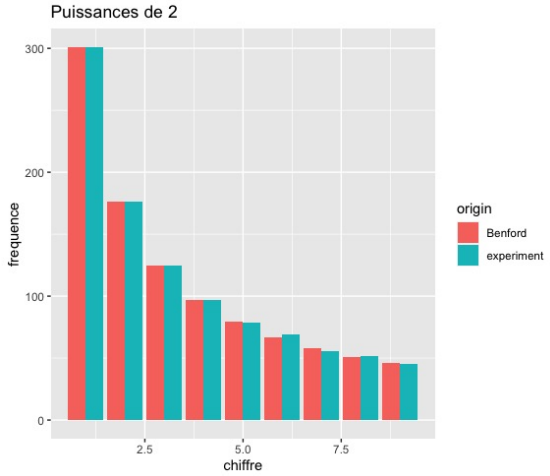


Figure 4: Histogramme des les puissances de 2 pour $N = 1000$

p valeur $= 0.999 > \alpha$

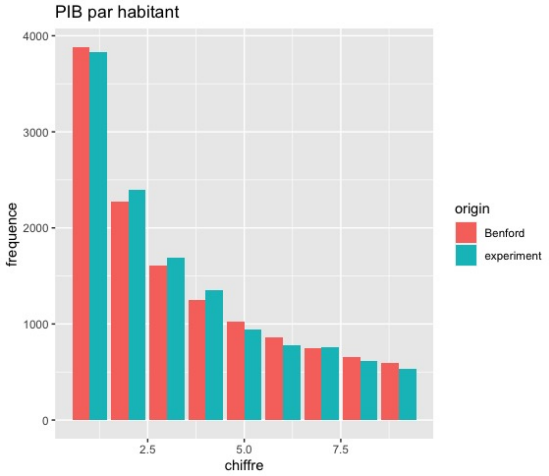


Figure 5: Histogramm de PIB par habitant, $N = 12897$

p valeur $= 0 < \alpha$

Visuellement, dans les deux cas, les fréquences des données (en bleu) sont proches des valeurs attendues (en orange). Par contre, si on regarde la valeur p, il est clair que les puissances de 2 correspondent presque parfaitement à la loi de Benford, alors que pour le PIB, nous devons rejeter notre hypothèse nulle.
Au début, le résultat semble un peu étrange, mais la raison est vite trouvée. Le N est treize fois plus grand dans le deuxième cas et la tolérance converge vers zéro pour N vers l'infini. Cela devient évident lorsque nous regardons l'évolution de la distance associée à $\alpha = 0.05$.

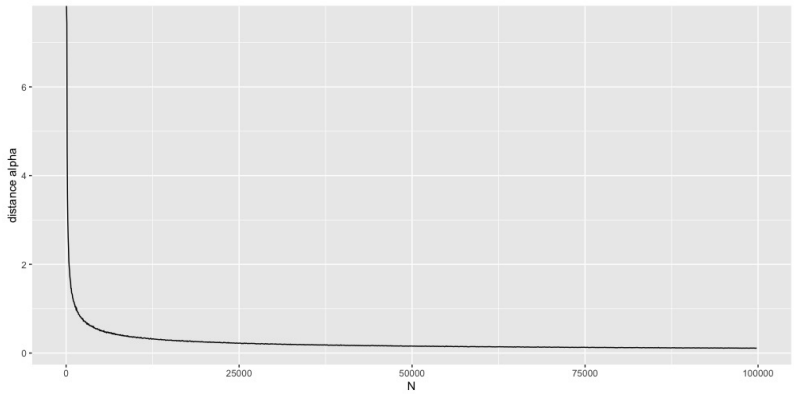


Figure 6: distance assoziée à α

3.2 Choix de α

En choisissant α, il est le défi de contrebalancer les deux erreurs suivantes.

Erreur de type I

Si nous rejetons l'hypothèse H_0 alors qu'elle est correcte, on parle d'une erreur de premier type. Dans notre cas, par exemple, cela signifie que nous accusons quelqu'un de fraude qui n'en a pas vraiment commis. Nous rejetons donc notre hypothèse nulle, même si les données sont en fait correctes. La probabilité pour cette erreur est précisément notre α.

Erreur de type II

La deuxième erreur consiste à confirmer à tort l'hypothèse H_0. La probabilité de cette erreur est plus difficile à trouver et dépend de l'hypothèse H_1.

En général, il faut supposer qu'il y a des erreurs dans les bilans. De plus, nous voulons éviter les fausses accusations. C'est pourquoi nous voulons minimiser la première erreur, même si cela implique de négliger d'autres cas suspects.
Les 0.05 déjà mentionnés sont une valeur populaire pour α, mais dans notre cas, on pourrait même utiliser 0.01 ou 0.005. Il faudrait alors calculer plus de distances de comparaison pour déterminer la valeur p avec précision, en particulier pour des zones aussi petites.

3.3 Comparaison de X^2 et de notre distance

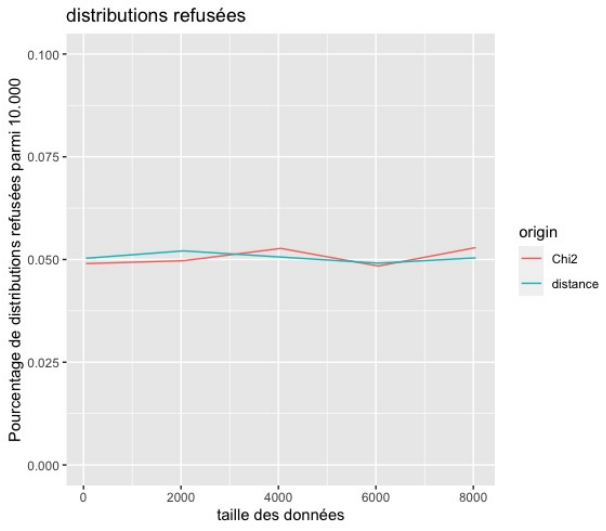


Figure 7: Pourcentage de distributions rejetées pour différents ensembles de données.

Pour comparer les deux tests, nous avons simulé dix mille expériences et analysé leurs résultats avec les deux. Comme on pouvait l'attendre, les deux tests rejettent environ 5% des expériences pour $\alpha = 0.05$. Il est intéressant de noter que les distributions rejetées ne se recoupent que dans la moitié des cas. Ces différences d'évaluation sont toutefois minimes, si l'on augmente α pour le test rejeté à 0.07, les deux tests s'accordent alors dans 80% des cas.

3.4 Loi de Benford pour le deuxième chiffre

La loi de Benford peut également être généralisée avec la même formule pour les chiffres arrière, dans ce qui suit pour le deuxième chiffre. Pour cela, nous additionnons les probabilités des différentes possibilités qui ont le nombre désiré comme deuxième chiffre.

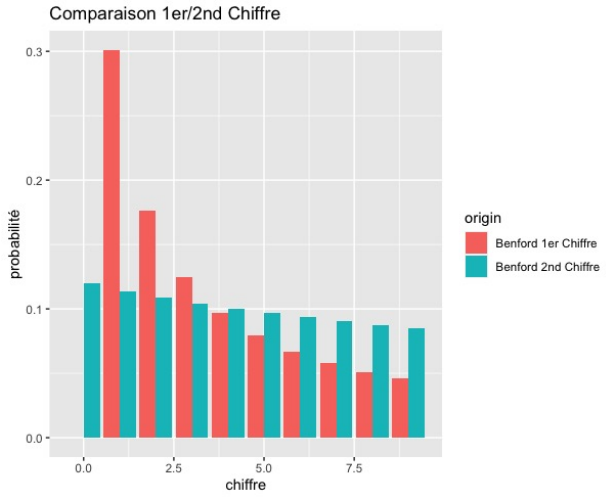
Soit $S_i = \{n \in \{10, 11, ..., 99\} |$ le deuxième chiffre est i $\}$, alors la probabilité d'avoir i comme second chiffre est la suivante:

$$p_i = \sum_{n \in S_i} \log_{10} \left(1 + \frac{1}{n}\right).$$

Par exemple, pour le nombre deux, on obtient:

$$S_2 = \{12, 22, 32, 42, 52, 62, 72, 82, 92\}$$
$$p_2 = \sum_{n \in S_2} \log_{10} \left(1 + \frac{1}{n}\right) = 0,1088.$$

Au total, nous obtenons cette répartition:



Chiffre	Prob. 1er pos	Prob. 2nd pos.
0	0	0.1197
1	0.3010	0.1139
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

Figure 8: Comparaison de la probabilité d'occurrence pour différents nombres.

Comme nous pouvons le constater, la différence des probabilités pour les différents i diminue. La répartition des chiffres arrière est intéressante, car souvent les petits changements ne modifient pas le premier chiffre, mais les chiffres arrière oui. Par conséquent, les chiffres arrière sont plus sensibles à la manipulation.

3.5 Estimer la vrai distribution

Approche bayésien

L'idée est de chercher pour un certain jeu de donné, la possibilité qu'elle suit une loi polynomiale de paramètre (θ_1, θ_2, ...θ_9). c'est une probabilité de probabilité. On peut choisir après un intervalle de confiance pour les paramètres.

Pour mieux formaliser; soit un modèle paramétrique qui représente un ensemble de mesures indexé par des paramètres $\theta \in \mathbb{R}^9$.

$$\mathscr{P}(\mathbb{P}_\theta, \theta \in \Theta) \; avec \; \Theta = [0,1]^9.$$

Sur cette ensemble on peut définir une fonction μ qui associe à chaque mesure de probabilité d'avoir cette distribution.

$$\begin{array}{rccc} \mu : & \mathscr{P} & \to & \mathbb{R}_+ \\ & \mathbb{P}_\theta & \mapsto & \textit{probabilité de générer le jeu de donné} \end{array}$$

Exemple :

Soit un jeu de donné pile-face ou on a eu 1000 pile et 0 face.

Si on pose $\mathbb{P}$(pile) = θ et $\mathbb{P}$(face) = 1-θ alors :

pour $\theta = \frac{1}{2}$ on a : $\mu(\mathbb{P}_{\frac{1}{2}}) = (\frac{1}{2})^{1000}$ donc très improbable.

pour $\theta = 1$ on a : $\mu(\mathbb{P}_1) = 1$.

Le plus probable est que θ se trouve à proximité de 1 avec une très grande certitude.
En faite on a $\mu(x) = x^{1000}$. si on avait 100 face et 900 pile alors dans ce cas $\mu(x) = x^{900} \cdot x^{100} \cdot \binom{1000}{900}$.

Cas de 2 variable :

Si on pose par exemple une expérience avec 3 boule; rouge, bleu et vert. Dans cette expérience on a eu: 5 rouge, 10 bleu et 2 vert . alors $\mathbb{P}$(R)= θ_1, $\mathbb{P}$(B) = θ_2 et $\mathbb{P}(V) = 1 - \theta_1 - \theta_2$ donc :

$$\mu(\theta_1, \theta_2) = \theta_1^5 \cdot \theta_2^{10} \cdot (1 - \theta_1 - \theta_2)^2 \cdot \frac{(5+10+2)!}{5!.10!.2!}$$

.

La figure ci dessous présente le cas où ; 1 rouge, 4 bleu et 6 vert. Les deux axes présentent θ_1 et θ_2, la troisième probabilité est directement déterminé par les deux premières.

Le troisième axe (oz) indique la probabilité d'obtenir les résultats déjà observés pour les deux paramètres θ_1 et θ_2 choisis comme cordonnés.

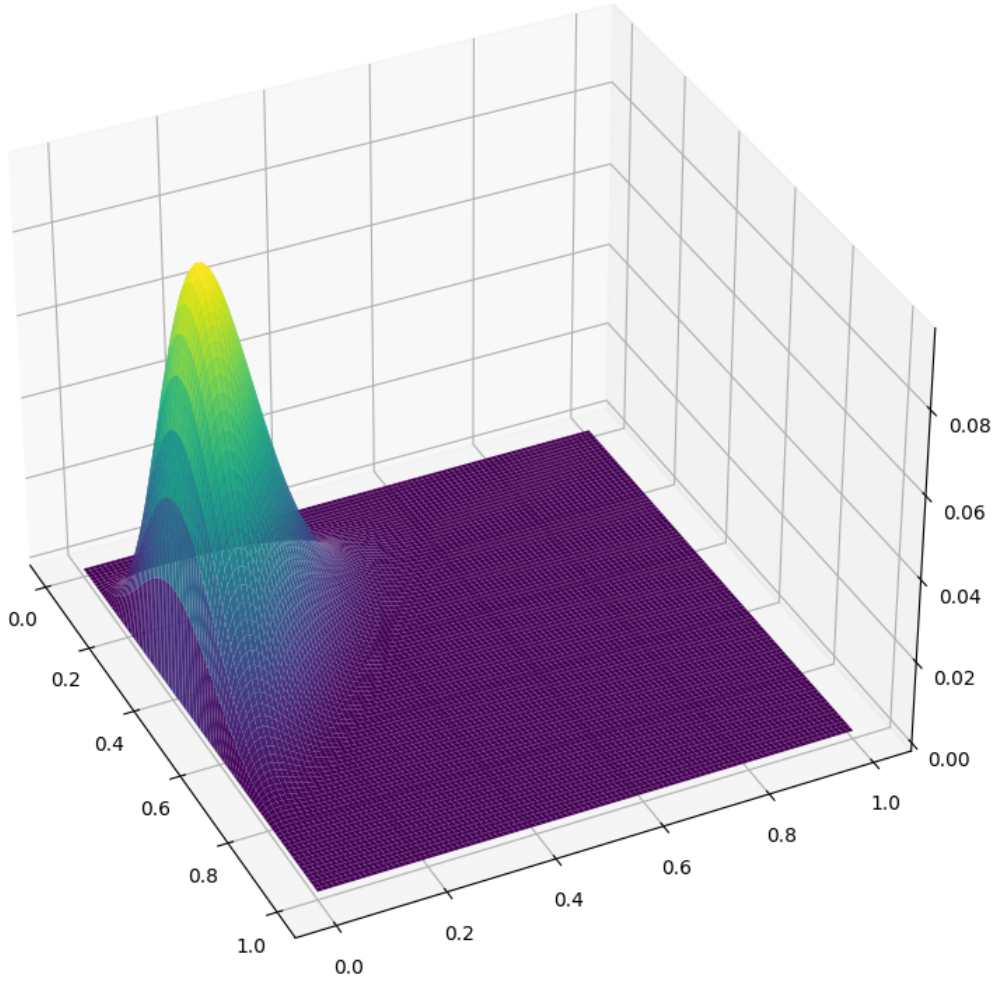


Figure 9: Cas de 2 variables possibles.

4 Application avec python

Github : https://github.com/scihelios/Benford-law
Afin de concrétiser notre approche dans une façon plus réel, on a décidé de créer une application qui est facile à utiliser. On voulait créer un programme qui analyse des données pré-traité et mis sous forme CSV. L'application comporte un GUI (Graphic-User-Interface) où on peut mettre les informations nécessaires par rapport au test qu'on va effectuer. En plus, le programme va générer automatiquement un rapport qui donne un description de l'analyse effectuée et comment l'interpréter.

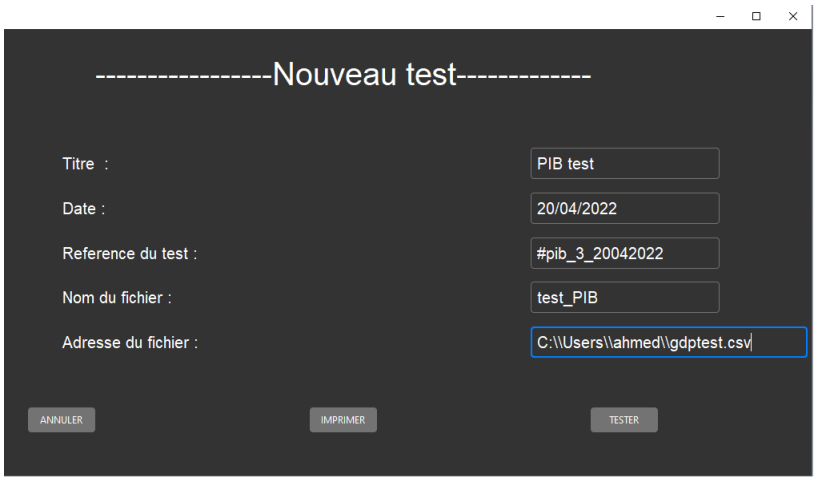


Figure 10: Screenshot de l'App

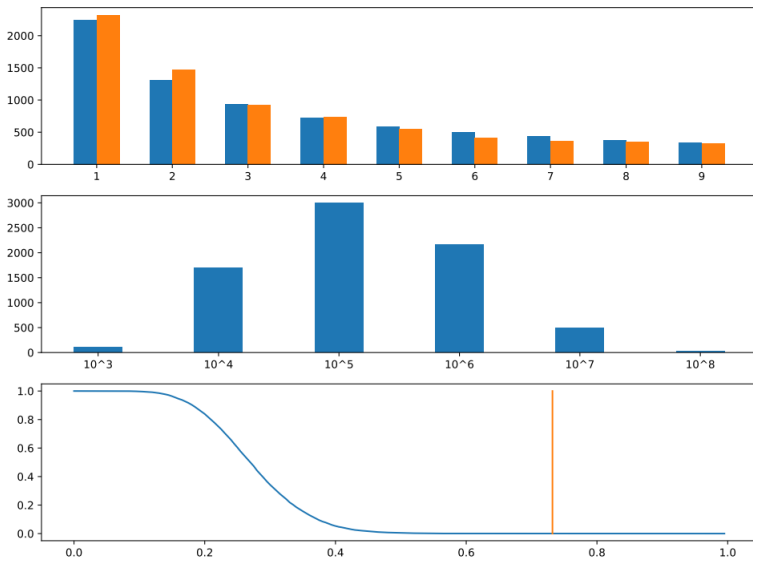
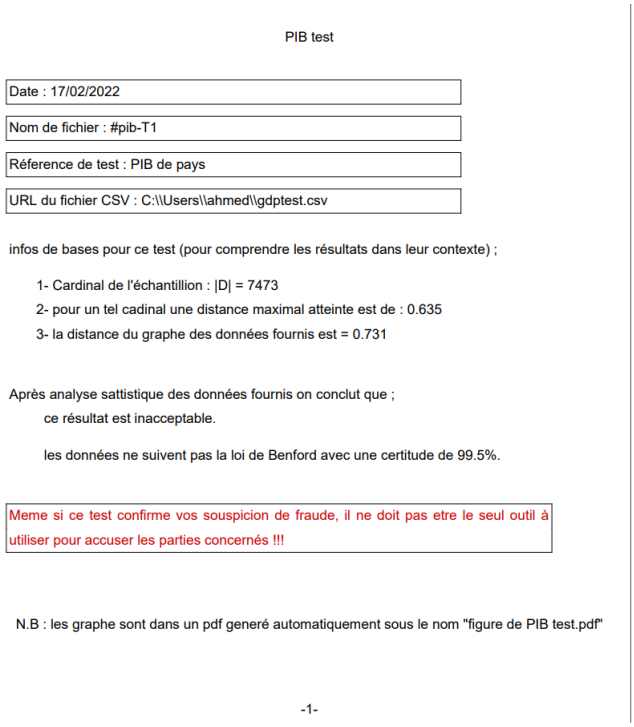


Figure 11: Graphes du test: distribution du premier chiffre, distribution des puissances, position dans la répartition des distances


PIB test

Date : 17/02/2022

Nom de fichier : #pib-T1

Réference de test : PIB de pays

URL du fichier CSV : C:\\Users\\ahmed\\gdptest.csv

infos de bases pour ce test (pour comprendre les résultats dans leur contexte) ;

1- Cardinal de l'échantillion : |D| = 7473
2- pour un tel cadinal une distance maximal atteinte est de : 0.635
3- la distance du graphe des données fournis est = 0.731

Après analyse sattistique des données fournis on conclut que ;
ce résultat est inacceptable.

les données ne suivent pas la loi de Benford avec une certitude de 99.5%.

Meme si ce test confirme vos souspicion de fraude, il ne doit pas etre le seul outil à utiliser pour accuser les parties concernés !!!

N.B : les graphe sont dans un pdf generé automatiquement sous le nom "figure de PIB test.pdf"

-1-

Figure 12: Résultat du test

5 Code dans R

Fonction distance

```
#N est le nombre de donnees et E le vecteur qui, a la
    position i, a le frequence de premiers chiffres du nombre
     i
distance <- function(N,E) {
     if(N==0) {return (0)}
     theo <-  Benford()*N
     Mat <- matrix(c(abs(theo/(E+1) -1),abs(E/(theo+1) -1)),
         ncol=2)
     D= sum(apply(Mat,1,max), na.rm = TRUE)
     return(D)
}
```

Fonction distribution (p-valeur)

```
#La fonction compare la distance des donnees d'etendue N
    avec les distances generees pour la classer
P_value <- function(N, Dist){
     df_Distances <- import("~Distances.csv")
     #trouver la position du N le plus proche
     posi <- which(abs(df_Distances$Amounts-N)==min(abs(df_
         Distances$Amounts-N)))
     #selectionner les distances pour ce N
     comp <- as.numeric(df_Distances[posi,2:1001 ])
     #trouver la position de la distance la plus proche et
         divise la par 1000
     proba <- (which(abs(comp-Dist)==min(abs(comp-Dist)))/
         1000)
     return(proba)
}
```

size	V1	V2	...	V955	**V956**	V957	...	V1000
0	0	0	...	0	0	0	...	0
...	...	...	...	...	...	...	...	...
23900	0.3128715	0.3033510	...	0.08728532	0.08718752	0.08680329	...	0.05016364
24000	0.3197177	0.3041845	...	0.08535157	**0.08506384**	0.08425276	...	0.03500114
24100	0.3321123	0.2956557	...	0.08637042	0.08619278	0.08595219	...	0.04448654
...	...	...	...	...	...	...	...	...

Table 1: df_Distances pour $N = 2403$ et $d_{donnees} = 0.0851534$, alors $p - valeur = 0.956$

Créateur d'expérience statistique

```
#une fonction qui genere un vecteur de longueur 9 avec N
    nombres evalues qui suivent la loi de Benford
Gen_Exp <- function(N) {
    return(as.vector(rmultinom(1, N, log10(1 + 1/(1:9)))))
}
```

6 References

histTed Hill, *Le premier chiffre significatif fait sa loi*, La Recherche, no 316, janvier 1999, p. 73. applMark J. Nigrini, *Benford's Law: Applications for Forensic Accounting, Auditing, and Fraud Detection*, Hoboken, NJ: Wiley, 2012. MCJ.E.Gentle, *International Encyclopedia of Education (Third Edition)*, Elsevier, 2010. pibhttps://www.humanprogress.org/ . guihttps://likegeeks.com/python-gui-examples-tkinter-tutorial/