Agrégation externe de mathématiques, session 2022
Épreuve de modélisation, option A : Probabilités et Statistiques

(*A01-public*)

Résumé : On s'intéresse à la loi des premiers chiffres significatifs des nombres qui nous entourent. Une loi continue et sa contrepartie discrète sont au cœur de l'étude. On justifie leur caractère universel (stabilité par changement d'échelle et tirage aléatoire par étapes). On en déduit un test d'authenticité de déclarations comptables à partir d'un jeu de données fourni dans le fichier `Ventes.txt`.

Mots clefs : Fonctions de répartition, tests, lois des grands nombres.

► *Il est rappelé que le jury n'exige pas une compréhension exhaustive du texte. La présentation, bien que totalement libre, doit être organisée et le jury apprécie qu'un plan soit annoncé en préliminaire. L'exposé doit être construit en évitant la paraphrase et mettant en lumière les connaissances, à partir des éléments du texte. Il doit contenir des illustrations informatiques réalisées sur ordinateur, ou, à défaut, des propositions de telles illustrations. Des pistes de réflexion, indicatives et largement indépendantes les unes des autres, vous sont proposées en fin de texte.*

Ce texte vous est fourni avec un fichier `Ventes.txt` *permettant d'illustrer sur des données réelles certains des résultats présentés. Des instructions pour lire ce fichier à l'aide des logiciels disponibles sont données en dernière page du texte.*

Simon Newcomb, un astronome et mathématicien du dix-neuvième siècle, avait remarqué que les tables de logarithmes étaient plus usées au niveau des premières pages ; il en a déduit que les nombres qui nous entourent commencent plus souvent par un premier chiffre petit que grand. Il a postulé que, pour tout $i \in \{1,\dots,9\}$, la fréquence d'apparition de i en premier chiffre significatif est $\log_{10}\big((i+1)/i\big)$, où $\log_{10}$ désigne le logarithme en base 10. Ce travail a été publié en 1881 et indépendamment redécouvert et approfondi par le physicien Franck Benford en 1938. Benford a mené un travail minutieux de collecte de données de diverses sources (numéros de rue de personnes tirées au hasard dans l'annuaire, hauteurs d'immeubles, nombres lus dans le *Reader's Digest*) ; beaucoup de ces jeux de données semblaient distribués selon la loi citée, mais pas tous. Et ce qui était le plus frappant, c'était que l'union de ces jeux de données collait quasi-parfaitement, elle, à cette loi.

Dans ce texte, nous étudierons deux lois de Benford, l'une continue $\mathscr{B}_c$, l'autre discrète $\mathscr{B}_d$ se déduisant de $\mathscr{B}_c$ en regardant le premier chiffre. Nous montrerons que $\mathscr{B}_c$ est stable par changement d'échelle puis nous obtiendrons une justification partielle du fait que des procédures de tirages aléatoires par étapes améliorent l'adéquation à $\mathscr{B}_c$. Ensuite, nous utiliserons ces lois pour tester l'authenticité de données. Enfin, nous montrerons que des suites sont distribuées, en un sens à préciser, selon la loi $\mathscr{B}_c$.

Dans la suite, X et U seront des variables aléatoires à valeurs respectives dans $\mathbb{R}_+^*$ et $[0,1[$ (U ne suivra pas nécessairement la loi uniforme), $\lfloor x \rfloor$ sera la partie entière de x et $\oplus$ sera le

résultat dans $[0,1[$ de l'addition modulo 1 (par exemple : $0,4 \oplus 0,7 = 0,1$). On notera $X \sim \mu$ lorsque X suit la loi μ.

1. Définition des lois de Benford et premières propriétés

Définition 1. *Pour tout $x \in \mathbb{R}_+^*$, on note $\mathscr{M}(x)$ l'unique réel de $[1,10[$, $k(x)$ l'unique entier relatif et $\mathscr{S}(x)$ l'unique entier de $\{1,\dots,9\}$ tels que*

$$x = \mathscr{M}(x)\,10^{k(x)}, \quad \mathscr{S}(x) = \lfloor \mathscr{M}(x) \rfloor . \tag{1}$$

On appelle $\mathscr{M}(x)$ la mantisse de x et $\mathscr{S}(x)$ le premier chiffre significatif de x. Par exemple, $\mathscr{M}(0,0761) = 7,61$, $\mathscr{S}(0,0761) = 7$, $\mathscr{M}(247,3) = 2,473$ et $\mathscr{S}(247,3) = 2$.

Définition 2. *La loi de Benford continue, notée $\mathscr{B}_c$, est la loi de 10^U où U suit la loi uniforme sur $[0,1[$. La loi de Benford discrète, notée $\mathscr{B}_d$, est la loi sur $\{1,\dots,9\}$ définie par $\mathscr{B}_d(\{i\}) = \log_{10}\big((i+1)/i\big)$ pour tout $i \in \{1,\dots,9\}$.*

Proposition 3. *$\mathscr{B}_c$ est absolument continue par rapport à la mesure de Lebesgue, de densité $x \mapsto \frac{\mathbf{1}_{[1,10[}(x)}{x \ln 10}$. Par ailleurs, si $X \sim \mathscr{B}_c$ alors $\mathscr{S}(X) \sim \mathscr{B}_d$.*

La propriété suivante caractérise $\mathscr{B}_c$ par son invariance par changement d'unité.

Proposition 4. *$\mathscr{M}(X) \sim \mathscr{B}_c$ si et seulement si $\mathscr{M}(\alpha X)$ et $\mathscr{M}(X)$ ont même loi pour tout $\alpha > 0$.*

Démonstration du sens direct. Posons $\beta = \log_{10}\alpha$ et $U = \log_{10}\mathscr{M}(X)$. On a $\mathscr{M}(\alpha X) = 10^{U \oplus \beta}$. Si $\mathscr{M}(X) \sim \mathscr{B}_c$ alors U suit la loi uniforme sur $[0,1[$. Ainsi, $U \oplus \beta$ suit aussi la loi uniforme sur $[0,1[$ donc $\mathscr{M}(\alpha X) \sim \mathscr{B}_c$. □

Introduisons un outil utile pour la suite. Pour T une variable aléatoire réelle, nous noterons $c_n(T) = \mathbb{E}(e^{2i\pi nT})$ pour tout $n \in \mathbb{Z}$. La suite $(c_n(T))_{n\in\mathbb{Z}}$ détermine $\mathbb{E}(f(T))$ pour tout polynôme trigonométrique f puis, par densité, pour toute fonction f continue 1-périodique. Donc, si T est à valeurs dans $[0,1[$, la suite $(c_n(T))_{n\in\mathbb{Z}}$ détermine la loi de T.

Démonstration de la réciproque. D'après le calcul fait pour le sens direct, il s'agit de voir que si $U \oplus \beta$ a même loi que U pour tout $\beta \in \mathbb{R}$ alors U suit la loi uniforme sur $[0,1[$. Pour tout $n \in \mathbb{Z}$, on a donc $c_n(U) = c_n(U \oplus \beta) = e^{2i\pi n\beta} c_n(U)$. En prenant β irrationnel, il vient $c_n(U) = 0$ pour tout $n \in \mathbb{Z}^*$. Donc U suit la loi uniforme sur $[0,1[$. □

2. Obtention de la loi de Benford par tirage en deux temps

Les numéros de rue recueillis par Benford semblaient correspondre à la loi $\mathscr{B}_d$. On va voir que cela vient de la situation suivante : on choisit une rue au hasard, elle a un nombre N aléatoire de numéros. On choisit ensuite un numéro au hasard dans cette rue, entre 1 et N. Voici une version continue de la situation discrète précédente : on prend X une variable aléatoire à valeurs dans $\mathbb{R}_+^*$ et, conditionnellement à X, on tire une variable aléatoire X' de loi uniforme sur $]0,X]$.

Proposition 5. *Avec les notations précédentes, $\mathscr{M}(X')$ et $\mathscr{M}(X)$ ont même loi si et seulement si $\mathscr{M}(X) \sim \mathscr{B}_c$.*

Démonstration. Soit A suivant une loi uniforme sur $]0,1]$ et indépendante de X. Alors X' a même loi que AX. Si l'on suppose $\mathscr{M}(X) \sim \mathscr{B}_c$ alors, d'après la proposition 4,

$$\mathbb{E}(f(\mathscr{M}(X'))) = \mathbb{E}(f(\mathscr{M}(AX))) = \int_0^1 \mathbb{E}(f(\mathscr{M}(\alpha X)))\, d\alpha = \int_0^1 \mathbb{E}(f(\mathscr{M}(X)))\, d\alpha = \mathbb{E}(f(\mathscr{M}(X))), \tag{2}$$

ce qui prouve que $\mathscr{M}(X')$ et $\mathscr{M}(X)$ ont même loi. Réciproquement, supposons que $\mathscr{M}(X')$ et $\mathscr{M}(X)$ ont même loi. Posons $U = \log_{10} \mathscr{M}(X)$ et $W = \log_{10} A$ si bien que $\mathscr{M}(AX) = 10^{U \oplus W}$. Ainsi, $U \overset{\text{loi}}{=} U \oplus W$. Par indépendance, $c_n(U) = c_n(U \oplus W) = c_n(U)c_n(W)$. Pour tout $n \in \mathbb{Z}^*$, l'inégalité triangulaire donne $|c_n(W)| < 1$ donc $c_n(U) = 0$. Ainsi, U suit la loi uniforme sur $[0,1[$ donc $\mathscr{M}(X) \sim \mathscr{B}_c$. □

Dans les notations précédentes, la loi μ' de $\mathscr{M}(X')$ dépend uniquement de la loi μ de $\mathscr{M}(X)$: notons $\mu' = F(\mu)$. Si $\mathscr{P}$ est l'ensemble des mesures de probabilité sur $[1,10[$, F est une application de $\mathscr{P}$ dans $\mathscr{P}$ dont l'unique point fixe est la loi de Benford. Si Y, Z sont deux variables aléatoires à valeurs dans $[1,10[$ de lois respectives μ, ν, la formule

$$d(\mu,\nu) = \sup_{n \in \mathbb{Z}} \left| c_n(\log_{10} Y) - c_n(\log_{10} Z) \right| \tag{3}$$

définit une distance sur $\mathscr{P}$. Un calcul de la démonstration précédente assure que

$$d(F(\mu),F(\nu)) = \sup_{n \in \mathbb{Z}^*} \left| c_n(W)c_n(\log_{10} Y) - c_n(W)c_n(\log_{10} Z) \right| \leq a d(\mu,\nu), \tag{4}$$

où $a = \sup_{n \in \mathbb{Z}^*} |c_n(W)| < 1$. Ainsi, $d(F(\mu),\mathscr{B}_c) \leq a d(\mu,\mathscr{B}_c)$, ce qui justifie que la procédure F renforce l'adéquation avec la loi de Benford. Précisément, une application répétée de cette procédure fait converger à vitesse géométrique vers la loi de Benford.

3. Application : test de sincérité de déclarations financières

Principe. Supposons que l'on dispose d'observations $X_1, X_2, \ldots$ indépendantes et identiquement distribuées selon une certaine loi sur $\mathbb{R}_+^*$ inconnue : comment tester que $\mathscr{M}(X_1)$ suit la loi $\mathscr{B}_c$? En théorie, on peut utiliser un test de Kolmogorov-Smirnov ; mais les données fournies sont généralement arrondies : on ne procure que quelques premiers chiffres significatifs. Ainsi, en pratique, on regroupe plutôt les données en classes en fonction de leur premier (ou de leurs deux premiers) chiffre(s) significatif(s), puis on applique un test du χ^2 d'ajustement. On fera simplement attention à ce qu'il y ait suffisamment d'observations n : une règle empirique est que la probabilité de chaque classe soit plus grande que $5/n$, ce qui conduit à la contrainte

$$n \geq 110 \tag{5}$$

dans le cas où l'on regroupe selon le premier chiffre significatif.

Utilisation sur un jeu de données. On pourra s'exercer sur le jeu de données fourni dans `Ventes.txt`, qui retrace des ventes effectuées lors de l'année 1996. La première colonne reporte l'index du vendeur et la seconde les montants des ventes de ce vendeur (négatifs pour des remboursements). Il s'agit de répondre à la question suivante : quels sont, parmi les huit vendeurs, ceux qui reportent de manière honnête leurs transactions ? Si on suppose que ces ventes résultent de procédures décrites à la partie 2 (choix d'un catalogue au hasard, puis choix d'une marque dans ce catalogue puis choix d'une rubrique dans cette marque, etc), il devient raisonnable de penser que les montants sont des réalisations de variables indépendantes, identiquement distribuées dont les mantisses suivent la loi $\mathscr{B}_c$. Si des données sont déclarées non conformes à cette modélisation par les tests précédemment décrits, c'est peut-être qu'elles ont été manipulées. On n'oubliera pas de préciser pourquoi le résultat du test de conformité n'inculpe pas directement un vendeur lorsque ses données sont jugées non conformes.

4. Suites distribuées selon la loi de Benford $\mathscr{B}_c$

On va s'intéresser ici à la convergence, en un sens à préciser, de suites déterministes (y_k) ou aléatoires (Y_k) de réels strictement positifs vers la loi $\mathscr{B}_c$. Ce que l'on veut modéliser, c'est que, par exemple, la proportion de termes y_k admettant i pour premier chiffre significatif parmi les premiers termes de la suite tend, lorsque n croît, vers $\log_{10}\big((i+1)/i\big)$.

4.1. *Cas de suites déterministes*

Dans la définition suivante, δ_t désignera la masse de Dirac en un point réel t et on rappelle que la convergence étroite de lois de probabilité signifie la convergence en loi d'une suite de variables aléatoires distribuées selon ces probabilités.

Définition 6. *On dit qu'une suite (y_k) de réels strictement positifs induit la loi de Benford $\mathscr{B}_c$ si on a la convergence étroite*

$$\frac{1}{n}\sum_{k=1}^{n}\delta_{\mathscr{M}(y_k)} \underset{n\to+\infty}{\rightsquigarrow} \mathscr{B}_c\,. \tag{6}$$

Par ailleurs, on dit qu'une suite (u_k) à valeurs dans $[0,1[$ est équirépartie si on a la convergence étroite

$$\frac{1}{n}\sum_{k=1}^{n}\delta_{u_k} \underset{n\to+\infty}{\rightsquigarrow} \mathscr{U}\,, \tag{7}$$

où $\mathscr{U}$ est la loi uniforme sur $[0,1[$.

En utilisant que $\log_{10} y \pmod 1 = \log_{10}\mathscr{M}(y)$ et en faisant appel à la définition 2, on voit qu'une suite (y_k) induit la loi de Benford si et seulement si la suite $\big(\log_{10} y_k \pmod 1\big)$ est équirépartie.

Or, il existe un critère d' équirépartition d'une suite déterministe, le critère de Weyl.

Lemme 7 (Critère de Weyl)**.** *Une suite* (u_k) *à valeurs dans* $[0,1[$ *est équirépartie si et seulement si pour tout entier relatif* $j \in \mathbb{Z} \setminus \{0\}$*,*

$$\frac{1}{n}\sum_{k=1}^{n} e^{i2\pi j u_k} \underset{n\to+\infty}{\longrightarrow} 0. \tag{8}$$

Démonstration. Le sens direct est par définition de la convergence en loi ; le sens réciproque utilise le fait (théorème de Stone-Weierstrass) que les polynômes trigonométriques forment un ensemble dense dans l'espace vectoriel des fonctions continues 1–périodiques sur $\mathbb{R}$ muni de la norme de la convergence uniforme. □

4.2. *Cas de suites de variables aléatoires*

Il est utile d'avoir un critère de Weyl pour des suites de variables aléatoires.

Proposition 8. *Soit* (U_k) *est une suite de variables aléatoires indépendantes à valeurs dans* $[0,1[$*. On rappelle la notation* $c_j(U_k) = \mathbb{E}(e^{2i\pi jU_k})$*. Alors* (U_k) *est presque-sûrement équirépartie si et seulement si pour tout entier relatif* $j \in \mathbb{Z} \setminus \{0\}$*,*

$$\frac{1}{n}\sum_{k=1}^{n} c_j(U_k) \longrightarrow 0. \tag{9}$$

Démonstration. Le sens direct est par intégration du critère donné par le Lemme 7. Le sens réciproque s'appuie sur une généralisation de la loi forte des grands nombres (dite de Kolmogorov) : si une suite de variables aléatoires indépendantes (Z_k) est telle que

$$\sum_{k\geq 1} \frac{\mathrm{Var}|Z_k|}{k^2} < \infty \qquad \text{et} \qquad \frac{1}{n}\sum_{k=1}^{n} \mathbb{E}\big[Z_k\big] \longrightarrow 0, \tag{10}$$

alors

$$\frac{1}{n}\sum_{k=1}^{n} Z_k \longrightarrow 0 \qquad \text{p.s.} \tag{11}$$

On pourra en admettre la preuve, qui est la suivante. En notant $Z'_k = Z_k - \mathbb{E}\big[Z_k\big]$ les versions recentrées des variables aléatoires, on a par indépendance que la suite $\sum_{k\leq n} Z'_k/k$ (lorsque n varie) forme une martingale convergente dans $\mathbb{L}^2$ donc convergente presque-sûrement ; on conclut ensuite par le lemme de Kronecker, un résultat classique d'analyse sur les suites, que

$$\frac{1}{n}\sum_{k=1}^{n} Z'_k \longrightarrow 0 \qquad \text{p.s.} \tag{12}$$

□

Suggestions et pistes de réflexion

► *Les pistes de réflexion suivantes ne sont qu'indicatives et il n'est pas obligatoire de les suivre. Vous pouvez choisir d'étudier, ou non, certains des points proposés, de façon plus ou moins approfondie, mais aussi toute autre question à votre initiative. Vos investigations comporteront une partie traitée sur ordinateur et, si possible, des représentations graphiques de vos résultats. À défaut, si vos illustrations informatiques n'ont pas abouti, il est conseillé d'expliquer ce que vous auriez souhaité mettre en œuvre.*

— *Suggestions mathématiques*
  — On pourra détailler certaines preuves du texte.
  — On pourra préciser le lien entre convergence pour la distance d introduite dans l'équation (3) et la convergence en loi.
  — Montrer que la loi de Benford permet même d'obtenir des lois sur les $p \geq 2$ premiers chiffres significatifs.
  — Rappeler l'énoncé et les propriétés d'au moins un test d'adéquation à une loi donnée afin de construire explicitement un test d'authenticité des données.
  — Montrer que si $\alpha > 0$ est tel que $\log_{10} \alpha$ est irrationnel, alors la suite (α^k) induit la loi de Benford.
— *Suggestions d'études empiriques*
  — On pourra expliquer le calcul de la borne inférieure dans l'inégalité (5) et mener le test d'ajustement du χ^2 proposé sur le jeu de données dans sa globalité (on ne sera pas tenu de vérifier que la contrainte (5) est respectée pour chaque classe).
  — Étudier le jeu de données vendeur par vendeur et dans sa globalité ; comment quantifier pour chacun l'accord des données correspondantes à la loi de Benford ?
  — Donner des exemples de processus stochastiques menant à la loi de Benford ; on pourra par exemple considérer $Y_k = \prod_{j=1}^{k} E_{j,k}$ pour un tableau $(E_{j,k})$ de variables aléatoires réelles i.i.d.

Lecture des fichiers de données

Indications pour la session 2022

Ce document est agrafé au présent texte car ce dernier est associé à un jeu de données réelles `Ventes.txt`, qui vous est fourni sous format texte.

La première ligne de ce fichier de données est une ligne de titre, contenant du texte indiquant ce qu'on trouve dans chacune des colonnes ; les lignes suivantes contiennent des données numériques.

Les instructions suivantes expliquent comment charger ce fichier sous différents logiciels de sorte que les données se retrouvent dans une matrice `A`.

Commencez par déplacer `Ventes.txt` dans votre répertoire personnel :

— ouvrir le répertoire `Données` se trouvant sur le bureau en double-cliquant sur la dernière icône (bleue) de la colonne de gauche

— choisir le fichier `Ventes.txt` et le déplacer à la souris (ou le copier) dans le `Répertoire personnel` (première icône du bureau de la colonne de gauche).

Lecture sous Scilab.

```
[A,text] = fscanfMat("Ventes.txt")
```

La matrice `A` contient les donnés numériques, `text` contient la ligne de texte.

Lecture sous R. La commande

```
B<-read.table("Ventes.txt", header = TRUE)
```

crée une variable `B` qui n'est pas exactement une matrice numérique, mais plutôt une structure. Les lignes suivantes transforment cette structure en matrice numérique si besoin est :

```
D<-dim(B); n<-D[1]; p<-D[2];
A<-matrix(0,n,p);
for (i in 1:p) A[,i]<-B[,i];
```

Lecture sous Octave. Commencer par détruire la première ligne du fichier `Ventes.txt`, puis exécuter :

```
A = load("Ventes.txt")
```

Lecture sous Python. On utilise la fonction `loadtxt` de la bibliothèque `numpy` :

```
from numpy import loadtxt
A = loadtxt("Ventes.txt", skiprows=1)
```

Le résultat obtenu est de type `array`. L'option `skiprows=1` sert à ignorer la ligne de commentaire du fichier.

Description du fichier `Ventes.txt` : Ce fichier a 2773 lignes (ligne de titre non comptée) et 2 colonnes. La première colonne contient les numéros des vendeurs (de 1 à 8), la seconde les montants des ventes.