CentraleSupélec

1A - Cursus ingénieur

Année 2020

Statistique des records

Étudiants :
Champagne de Labriolle Guilhaume
Porcher François
Jeannin Hugo
Hainsellin Paul
Laulhère Hugo

Encadrant :
Challet Damien

Remerciements : Nous souhaitons remercier Damien Challet pour nous avoir encadré dans ce projet et orienté dans nos recherches. Nous tenons également à remercier le corps enseignant pour leur aide et l'encadrement du pôle mathématiques : Véronique Letort, Paul-Henry Cournede, Gurvan Hermange et Ioane Muni-Toke.

Table des matières

1 Le sujet

1.1 Contexte et enjeux

L'étude des statistiques a une importance capitale dans notre monde moderne où l'on analyse des grands nombres de données. Elle touche à des domaines divers tels que le sport, le climat, l'économie, la physique... L'étude des statistiques à notamment un but prédictif. Il s'agit, en effet, d'anticiper les événements futurs pour aider à la prise de décision. Par exemple, les statistiques météorologiques (1) (2) permettent aisément de mettre en avant les effets de l'activité humaine sur la composition de l'atmosphère, la température moyenne à la surface du globe ou le nombre d'événements climatiques exceptionnels. Une autre motivation est l'étude de données financières où les statistiques sont autant d'indicateurs qui aide les investisseurs dans leurs transactions. Des statistiques particulièrement cohérentes sont les records. Pour une série temporelle donnée un record apparait lorsque la valeur de la série temporelle est supérieure à toutes ses valeurs précédentes.

Bien souvent, nous n'avons pas accés à la loi d'une série temporelle mais simplement à ses valeurs. La statistique paramétrique permet d'obtenir à partir d'échantillons, des informations sur la loi sous-jacente. Cependant, la beauté de la théorie des records et qu'elle ne requiert par l'existence d'un kurtosis ou d'une variance. Un record existe toujours tant que la loi est finie pp. Nous allons dans ce document étudier quelque aspects de cette théorie, ainsi que le cadre théorique dans lequel nous nous sommes plongés pour tenter de mieux comprendre l'influence des permutations sur cette théorie.

1.2 État de l'art

En 1952, K.N. Chandler introduit la statistique des records (3) . Depuis le sujet a été étudié essentiellement dans le cadre de variable indépendantes et identiquement distribuées (4) (5). Récement Krug s'est intéressé au cas des distributions non identiques. Toutefois dans la majorité des cas, les séries temporelles sont corrélées, c'est dans ce cadre que s'inscrit l'article de Satya N. Majumdar1and et Robert M. Ziff (6) proposé par notre encadrant.

2 Notre travail

2.1 Compréhension de l'article

Dans un premier temps nous nous sommes approprié les notions mathématiques abordées dans l'article de départ dont l'objectif est de proposer des formules probabilistes concernant les records d'une marche aléatoire dont le pas suit une loi de distribution symétrique, centrée et continue (propriété P). Puisque ces formules ne dépendent pas de la loi considérée pour le pas mais sont vérifiées dès que (P) est vraie, notre étude s'inscrit dans le cadre des statistiques non paramétriques et nous appellerons donc « universalité » toute propriété valable pour la famille de loi vérifiant (P). Un exemple d'universalité est la loi de probabilité (et donc également l'espérance) du nombre de records en N étapes, ou encore la loi de probabilité de la durée moyenne d'un record en N étapes.

Définition 1 - Record d'une série temporelle.

Soit $(X_n)_{n\in[0,N]}$ une série temporelle en N étapes. Un record de cette série est un sous-ensemble $[n_1, n_2]$ de $[0, N]$ tel que $\forall n \in [n_1, n_2], X_n \leq X_{n_1}, X_{n_2+1} > X_{n_1}$ (si $n_2 < N$) et $X_{n_1} > X_{n_1-1}$ (si $n_1 > 0$). Par convention, $\{0\}$ est un record. La longueur d'un record $[n_1, n_2]$ est $card([n_1, n_2])$, sa hauteur (pour $n_1 > 0$) est $X_{n_1} - X_{n_1-1}$.

Voici la liste complète des « universalités » évoquées dans l'article :

1. Loi et espérance du nombre de records en N étapes : $\mathbb{E} \sim \sqrt{\frac{4N}{\pi}}$
2. Loi de l'écart-type du nombre de records en N étapes : $\mathbb{E} \sim \sqrt{(2 - \frac{4}{\pi})N}$
3. Loi de la durée du record le plus court en N étapes : $\mathbb{E} \sim \sqrt{N}\pi$
4. Loi de la durée du record le plus long en N étapes : $\mathbb{E} \sim c * N$,où c est une constante non trivale
5. Loi de probabilité de la durée moyenne d'un record en N étapes.

2.2 Vérification informatique

Après cette première phase, nous avons validé les formules informatiquement (en python) pour différentes lois vérifiant (P) (comme la loi normale ou

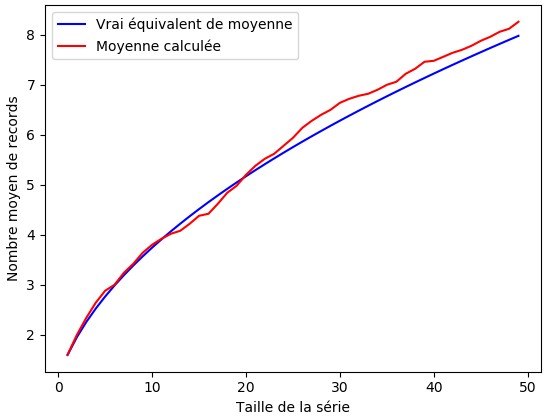
autre) en utilisant la loi des grands nombres pour estimer l'espérance des universalités par leur moyenne statistique sur un certain nombre de simulations de marches aléatoires.

Les résultats peuvent être observé sur les figures 1 13 14 15 présentes en annexe.

2.2.1 Une première approche

Dans le cas où l'universalité étudiée est le nombre de records, nous constatons que lorsque N (le nombre d'étapes de la marche aléatoire) augmente, la courbe « synthétique » (= générée informatiquement) oscille beaucoup autour de la courbe théorique.

FIGURE 1 – Nombre moyen de record



Après réflexion, nous avons compris d'où vient ce problème : quand nous écrivons l'inégalité la loi faible des grands nombres, il vient que pour contrôler l'écart entre les deux courbes, il faut adapter le nombre de simulations dans le calcul de la moyenne statistique à N, alors que dans notre cas le nombre de simulations était le même pour tout N. Autrement dit, la vitesse de convergence en probabilité de la moyenne statistique vers l'espérance du nombre de record est d'autant plus lente que le nombre d'étapes N est grand.

2.2.2 Raffinement de la courbe synthétique

D'après l'article de Satya N. Majumdar1and et Robert M. Ziff (6) :

$$< M > \sim \sqrt{\frac{4N}{\pi}} \tag{1}$$

$$V[M] \sim 2(1 - \frac{2}{\pi})N \tag{2}$$

On ne dispose pas de :

$$\lim_{N \to 0} M - E[M] = 0 \quad pp. \tag{3}$$

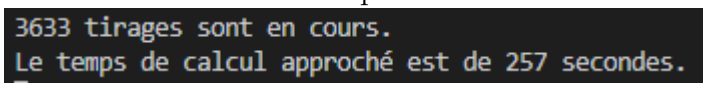
Cependant la loi faible des grands nombres nous assure sur un grand nombre de tirages de séries temporelles que le nombre moyen de nombre de records converge vers $< M >$. De façon plus précise, on dispose de l'inégalité :

$$P(|\frac{S_n}{n} - E[M]| \geq \epsilon) \leq \frac{V(M)}{n\epsilon^2} \tag{4}$$

$$\frac{V(M)}{n\epsilon^2} \sim \frac{2(1 - \frac{2}{\pi})N}{n\epsilon^2} \tag{5}$$

On peut donc approcher $E[M]$ à ϵ près informatiquement et obtenir un "n"convenable pour que la probabilité d'avoir des valeurs critiques ne soit pas trop éloignée de $E[M]$. Une analyse de la complexité des algorithmes permet ensuite de déterminer les paramètres et le temps de calculs associé.

FIGURE 2 – Maitrise du temps de calcul avec auto-adapt



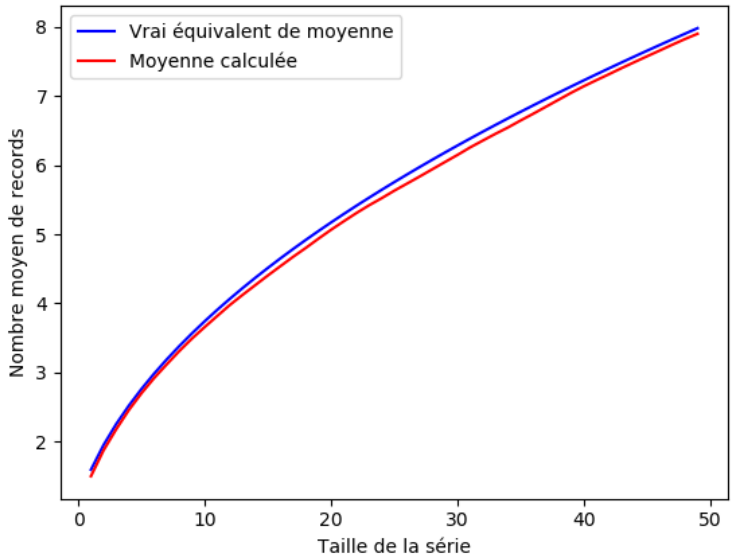
En mettant ces dernières considérations en pratique, nous obtenons une courbe plus lisse et plus proche de la courbe théorique 3.

2.3 Tests statistiques

2.3.1 Compréhension des tests statistiques

Nous nous sommes ensuite documentés sur les tests statistiques dans leurs généralités, en particulier en lisant la page wikipédia à ce sujet. Un test statistique permet donc de vérifier avec une certaine confiance si un

FIGURE 3 – Nombre moyen de record avec raffinement, $P(|X - E[X]| < 1) = 0.01$

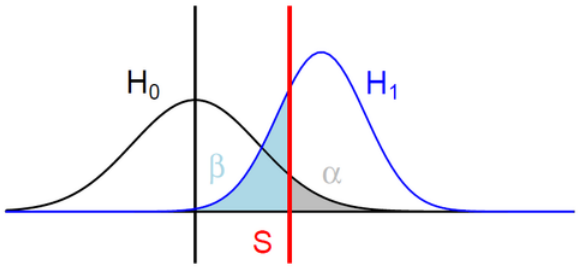


échantillon statistique vérifie ou non une hypothèse que l'on note H0, et le cas échéant il permet de dire avec quelle certitude cet échantillon vérifie une hypothèse alternative notée H1. Pour réaliser ce test, il faut disposer d'une quantité calculable à partir de l'échantillon (la statistique de test S) dont on connaît la distribution sous H0 et sous H1. La donnée de ces distributions et de la valeur de S(échantillon) permet d'évaluer avec quelle vraisemblance les hypothèses H0 ou H1 sont satisfaites ou rejetées.

2.3.2 Vers la réalisation de tests statistiques

Dans notre cas, l'hypothèse H0 est la suivante : « On peut modéliser le pas de la série temporelle donnée par une loi dont la distribution est symétrique centrée continue ». Cette hypothèse étant constituée de trois sous-hypothèses, on a plusieurs possibilités pour le choix de l'hypothèse alternative H1 en particulier les 3 suivantes : « la loi du pas n'est pas continue mais centrée et symétrique», « la loi du pas est non centrée mais symétrique par rapport à l'espérance et continue», « la loi du pas n'est pas symétrique mais centrée et continue». Nous disposons également de plusieurs possibilités pour

FIGURE 4 – Un schéma représentant le principe de fonctionnement



la statistique de test S qui sont en réalité les « universalités » listées plus haut dans ce rapport. Une grande difficulté est le choix de H1 qui peut se décliner d'une infinité de manière différente et également le choix de la statistique S qui permet de discriminer le mieux possible les deux hypothèses. Nous pouvons calculer la distribution de toutes les « universalités » sous H0, mais un autre problème rencontré est le calcul de la distribution de ces statistiques de test sous H1 qui n'est a priori pas universelle pour l'ensemble des lois vérifiant H1. En effet, par exemple, pour le cas où H1 = « la loi du pas est non centrée mais symétrique par rapport à l'espérance et continue », la distribution de S pour des lois vérifiant H1 peut varier énormément selon leur espérance qui peut être arbitrairement grande en valeur absolue.

Cas d'une variable de Rademacher

Théorème 1.

$$P(M,N) = \frac{U_{M,N}+1}{2^N}$$

En partitionnant en fonction de la montée ou de la descente au premier pas, on obtient :

$$U_{M,N} = U_{M-1,N-1} + U_{M+1,N-1} + \delta_{M,0}U_{O,N-1} \tag{6}$$

En raisonnant sur $f_M = \sum_{N=0}^{\infty} U_{M,N}Z^N$, on obtient $U_{M,N} = \binom{N}{\lfloor \frac{N-M}{2} \rfloor}$. La démonstration se fait par récurrence. En traçant la densité de probabilité, on ne fait pas vraiment la différence. Discriminer une densité discrète se fait en réalité en regardant si 2 sauts sont égaux, ce qui est impossible si la loi est continue.

3 Dévelopement du test statistique

3.1 Test de forte positivité de l'espérance

Dans cette partie nous proposons le développement d'un test statistique à l'aide des universalités de l'article (6). Dans un premier temps, nous utilisons le nombre de records qui est connu sous l'hypothèse H0.L'hypothèse H0 étant la loi peut être modélisée par une variable aléatoire dont la loi est centrée,symétrique et continue. On fait ici un test asymétrique, l'hypothèse H1 est alors : "la loi est symétrique, continue et non-centrée (de moyenne significativement positive)". Nous calculons alors la densité du nombre de record sous l'hypothèse H0 et sous plusieurs densités dans H1. On cherche ensuite à avoir un critère discriminant pour savoir si on accepte l'hypothèse H0 ou l'hypothèse H1. Pour cela, on fixe un risque de première espèce qui représente la probabilité d'accepter H1 alors que H0 est vraie et un risque de seconde espèce qui représente la probabilité de rejeter H0 alors que H1 est vrai.

Soit α le risque de première espèce, on cherche alors l'abscisse x_α tel que l'intégrale de la densité de probabilité sous H0 entre $-\infty$ et x_α soit égale à α.La zone à gauche de x_α correspond à la zone d'acceptation de l'hypothèse H0. De même, soit β le risque de seconde espèce, on cherche alors x_β tel que l'intégrale de x_β à $+\infty$ soit égale à β. On définit alors la zone d'acceptation de l'hypothèse H1 pour les abscisses supérieures à x_β (voir Figure 1).

Maintenant, une forte positivité de l'espérance se traduit par un indice $x_\alpha \leq x_\beta$, c'est à dire, $\beta\%$ des séries temporelles ayant une loi d'ésperance plus grande qu'un seuil μ_{lim} ont un nombre de record supérieur à ce $x_\alpha = x_\beta$ et $\alpha\%$ séries temporelles ayant une loi dans H0 ont un nombre de record inférieur à ce $x_\alpha = x_\beta$. On peut trouver cette valeur μ_{lim} informatiquement à partir de lois normales (densités centrales en théorie des probabilités). Par example pour $\alpha = \beta = 0.8, \mu_{\text{lim}} = 0.25$ (voir figure 3.1).

3.2 Test statistique assymptotique (TCL)

Soit $P^{(N)}$ un prix de N étapes généré comme vous le savez, je note $R(P^{(N)})$ son nombre de records.

On cherche à tester les hypohèses :

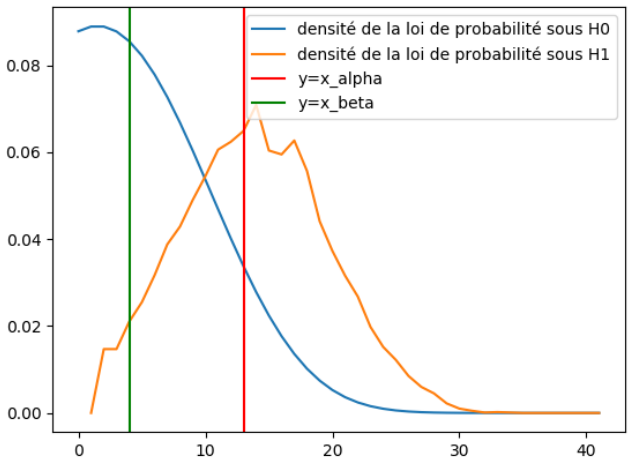


FIGURE 5 – Test de centralité pour $\alpha = \beta = 0.9$ la densité sous H1 étant réalisé grâce à une loi normale de variance et de moyenne égale à 1

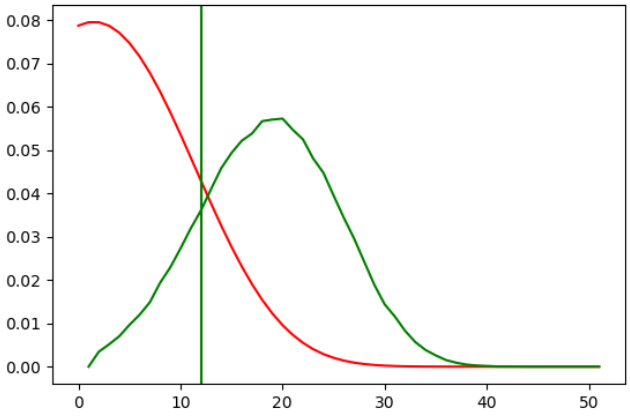


FIGURE 6 – Test de centralité pour $\alpha = \beta = 0.8$ pour une moyenne de 0.25

$$\begin{cases} H_0 : \text{La loi générant le prix est continue et symétrique.} \\ H_1 : \text{La loi n'est pas continue et symétrique.} \end{cases}$$

Nous disposons sous l'hypothèse H_1 de :

$$E(R(P^{(N)})) \sim \sqrt{\frac{4N}{\pi}} \tag{7}$$

$$\sqrt{(V(R(P^{(N)})))} \sim \sqrt{(2-\frac{4}{\pi})N} \tag{8}$$

Je considère alors (sur un échantillon de p prix, noté X) la statistique $T_{N,P}$:

$$T_{N,p} = \sqrt{p}\frac{\overline{R_p(P^{(N)})} - \mu_N}{\sigma_N}, \tag{9}$$

$$\tag{10}$$

Avec $\mu_N = E(R(P^{(N)}))$, et $\sigma_N^2 = V(R(P^{(N)}))$.

Alors d'après le TCL $T_{N,p}$ converge en loi par rapport à p vers $\mathcal{N}(0,1)$. Pour une taille α donné, on peut donc construire la région de rejet asymptotique (par rapport à N) :

$$R_\alpha = \{X, q_{\alpha/2}^{\mathcal{N}(0,1)} \leq T_{N,p} \leq q_{1-\alpha/2}^{\mathcal{N}(0,1)}\} \tag{11}$$

Ne connaissant pas μ_N et σ_N on essaie de plutôt utiliser leurs équivalents.

Proposition. $T_{N,p} = \sqrt{p}(\frac{\overline{R_p(P^{(N)})} - \sqrt{\frac{4N}{\pi}}}{\sqrt{(2-\frac{4}{\pi})N}} + o_N(1))$ *Avec* $o_N(1)$ *signifiant que la quantité est indépendante de p.*

Démonstration. On dispose en traduisant les équivalents :

$$\mu_N = \sqrt{\frac{4N}{\pi}} + o_N(\sqrt{N})$$

$$\sigma_N = \sqrt{(2-\frac{4}{\pi})N} + o_N(\sqrt{N})$$

On injecte ceci dans la quantité intéressante, alors :

$$\begin{aligned}
T_{N,p} &= \sqrt{p}\frac{\overline{R_p(P^{(N)})} - \sqrt{\frac{4N}{\pi}} + o_N(\sqrt{N})}{\sqrt{(2-\frac{4}{\pi})N} + o_N(\sqrt{N})} \\
&= \sqrt{\frac{p}{(2-\frac{4}{\pi})N}}\frac{\overline{R_p(P^{(N)})} - \sqrt{\frac{4N}{\pi}} + o_N(\sqrt{N})}{1 + o_N(1)} \\
&= \sqrt{\frac{p}{(2-\frac{4}{\pi})N}}(\overline{R_p(P^{(N)})} - \sqrt{\frac{4N}{\pi}} + o_N(\sqrt{N})(1 + o_N(1))) \\
&= \sqrt{\frac{p}{(2-\frac{4}{\pi})N}}(\overline{R_p(P^{(N)})} - \sqrt{\frac{4N}{\pi}} + o_N(\sqrt{N}))
\end{aligned}$$

□

On peut donc considérer la zone de rejet asymptotique lorsque $\sqrt{(}p)o_N(1)$ est négligeable : (je note $q = -q_{\alpha/2}^{\mathcal{N}(0,1)} = q_{1-\alpha/2}^{\mathcal{N}(0,1)}$)

La condition ne pouvant être vérifiée pour $p \gg N$, on ne peut pas donc réduire la zone autant que souhaité.

$$R_\alpha = \{X, q_{\alpha/2}^{\mathcal{N}(0,1)} \leq \sqrt{p}\frac{\overline{R_p(P^{(N)})} - \sqrt{\frac{4N}{\pi}}}{\sqrt{(2-\frac{4}{\pi})N}} \leq q_{1-\alpha/2}^{\mathcal{N}(0,1)}\} \tag{12}$$

$$= \{X, \sqrt{\frac{4N}{\pi}} - q\sqrt{\frac{(2-\frac{4}{\pi})N}{p}} \leq \overline{R_p(P^{(N)})} \leq \sqrt{\frac{4N}{\pi}} + q\sqrt{\frac{(2-\frac{4}{\pi})N}{p}}\} \tag{13}$$

Nous pouvons remarquer que connaissant la probabilité exacte d'avoir N records sur une série temporelle de taille M, nous pouvons calculer l'espérance et la variance exacte de notre fonction pivotale.

4 Recherches sur les permutations

4.1 Pourquoi utiliser des permutations ?

Comme expliqué dans l'introduction, bien souvent les observations se font sans connaissance de la loi sous-jacente. Cependant, un test statistique avec

une seule observation de record ne donne que très peu d'information. Heureusement, une hypothèse facilement vérifiable sans record (le développement d'un test pour cette hypothèse basé uniquement sur les records est un domaine à explorer) est implicitement supposée dans ce document. L'indépendance des sauts. Dès lors, des permutations de ces sauts permetttent de décupler, à priori, l'information apportée par une unique série temporelle.

Ainsi le but des permuations va être de diversifier la série temporelle, rendre les records des séries temporelles issues de l'application des permutations à la série initiale représentatif de la densité réelle des records sous la loi inconnue des sauts.

Ce qui va guider la reflexion des paragraphes suivant va être qu'une intervertion de deux sauts à de forte chances de mener à la destruction ou la construction d'un nouveau record. Ainsi nous allons chercher pour un groupe G_n à maximiser la valeur : $\tau(G_n) = \sum_{(\sigma_i,\sigma_j)\in G_n} \tau(\sigma_i, \sigma_j)$, où $\tau(f, g)$ est la distance de Kendall Tau qui compte le nombre de paires d'élements inversés sur $f \circ g^{-1}$. Par exemple, $\tau([5, 4, 3, 2, 1]) = \tau([5, 4, 3, 2, 1], \mathrm{Id}) = \binom{5}{2}$

4.2 Introduction

Comme nous l'avons expliqué, le but est de trouver une groupe G_n, maximisant le tau de Kendall : τ. La construction de se groupe va s'appuyer sur la détermination d'une partie génératrice assez petite ne générant pas toutes les permutations mais seulement certaines. Nous allons donc sans attendre introduire la permutation σ_5 : illustrée ci-dessous

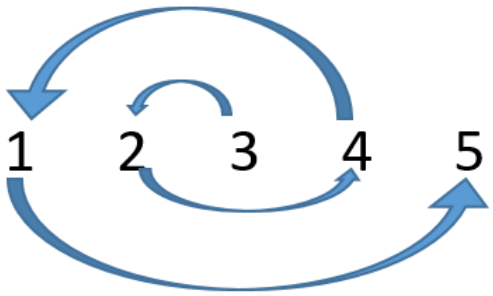


FIGURE 7 – σ_n illustrée

4.3 Analyse de σ_n

Dans un soucis de simplicité, je ne détaillerai les preuves que dans le cas de n impair. Les résultats sont aussi vrais dans les cas pair et les démonstration sont plus délicates mais similaires. Commençons par analyser $\tau(\sigma_n)$. Tout d'abord, l'ordre de σ_n est n. En effet un rapide exercice mental illustré par la figure 5 nous montre que nous avons affaire ici à un cycle. A présent, reformulons σ_n en l'analysant numériquement pour n petit :

$$\sigma_2 : [2, 1]$$
$$\sigma_3 : [2, 3, 1]$$
$$\sigma_4 : [4, 3, 1, 2]$$
$$\sigma_5 : [4, 3, 5, 2, 1]$$
$$\sigma_6 : [6, 5, 4, 1, 3, 2]$$
$$\sigma_7 : [6, 5, 4, 7, 3, 2, 1]$$

On distingue 2 blocs sur les permutations impaires qui sont inversés : $\sigma_7 : [6, 5, 4|7|3, 2, 1]$. Ici $6, 5, 4$ et $3, 2, 1$. On calcule facilement $\tau(\sigma_{2p+1}) = 2p^2$. Ceci peut se généraliser en un lemme sur les entiers quelconques :

Théorème 2. *Soit n un entier naturel, alors $\tau(\sigma_n) = \left\lceil \frac{(n-1)^2}{2} \right\rceil$*

$\tau(\sigma_n)$ n'est pas normalisé, et il n'est utile que lorsqu'il est comparé à l'entier $\binom{n}{2}$. On peut pour cela calculer le nombre de paires qui ne sont pas en inversion. On voit alors que ces paires sont les nombres du premier bloc avec le nombre du milieu. Par exemple sur σ_7, ce sont $(6, 7)$, $(5, 7)$ et $(4, 7)$. Dans ce cas on peut remarquer le résultat suivant sur un n quelconque.

Théorème 3. *Soit n un entier naturel, alors $\tau(\sigma_n) = \binom{n}{2} - \left\lfloor \frac{n-1}{2} \right\rfloor$. Ainsi, $\tau(\sigma_n) \sim \binom{n}{2}$*

Pour générer un groupe il faut à priori étudier les itérations de σ_n et on peut encore une fois représenter σ_n^2 pour des petites valeurs de n avant d'en donner une forme générale.

$$\sigma_2^2 : [1, 2]$$
$$\sigma_2^3 : [3, 1, 2]$$
$$\sigma_2^4 : [2, 1, 4, 3]$$
$$\sigma_2^5 : [2, 5, 1, 3, 4]]$$
$$\sigma_2^6 : [2, 3, 1, 6, 4, 5]$$
$$\sigma_2^7 : [2, 3, 7, 1, 4, 5, 6]$$

On peut remarquer que les deux nombres extremaux sont ramenés au centre et inversés. σ_n^2 n'a donc pas un tau de Kendall elevé, il tend même vers 0 si on le normalise. Il est intéressant de calculer la troisième itération de σ_n. Laissons donc défiler sous nos yeux les 3èmes itérations de σ_n.

$$\sigma_3^3 : [1, 2, 3]$$
$$\sigma_3^4 : [3, 4, 2, 1]$$
$$\sigma_3^5 : [3, 1, 4, 5, 2]$$
$$\sigma_3^6 : [5, 4, 6, 2, 1, 3]$$
$$\sigma_3^7 : [5, 4, 1, 6, 7, 3, 2]$$

On trouve aisément en appliquant σ_{2p+1} à σ_{2p+1}^2 la forme générale suivant :

$$\sigma_{2p+1}^3 : [2p-1, \ldots, p+1, 1, 2p, 2p+1, p, \ldots, 2]$$

C'est à partir de cette forme générale que l'on peut en déduire le théorème suivant.

Théorème 4. *Soit p un entier naturel, alors $\tau(\sigma_{2p+1}^3) = 2p(p-1)$*

Encore une fois, ce nombre n'est utile que lorsqu'il est comparé à $\binom{n}{2}$. Ainsi en calculant la différence de ces deux termes, on trouve une formule similaire à celle calculée lors du théorème 2.

Théorème 5. *Soit $p \geq 1$ un entier naturel, $\tau(\sigma_{2p+1}^3) = \binom{2p+1}{2} - 3p$.*

Toujours en construction, idée : étudier la forme génerale de σ_n^k. Cette question a été résolue, et une expression de la forme générale est possible. Cependant ce qui va suivre me pousse à croire qu'il n'est pas utile de détailler ces calculs. On trouve encore une forme du type $\binom{2p+1}{2} - k \cdot p$ avec k entier.

4.4 Génèse du groupe

Une première idée serait de considérer le groupe $< \sigma_n >$, de cardinal n. C'est un groupe malheureusement trop petit pour être vraiment utile. Dans un second temps, trouver une permutation commutant avec σ_n est impossible car σ_n est un n-cycle, donc les seules permutations commutant avec cette dernière sont les puissances de σ_n.

La difficulté majeure apparaît donc maintenant. Si nous voulons vraiment avoir un groupe, il faut réussir à le générer sans obtenir l'ensemble des permutations. Considérer le groupe engendré par σ_n et un retournement total ou une rotation circulaire ($s(k) = k\,mod\,[n] + 1$) génère presque toutes les permutations (un calcul informatique aide pour cette partie).

Devons-nous alors abandonner l'idée d'en faire un groupe à tout prix ? Existe-t-il une permutation simple permettant de trouver de manière explicite le groupe engendré par σ_n et cette dernière ?

Cette idée restera sans réponse par la suite, la notion de groupe n'étant pas nécessaire pour le reste.

4.5 Comparaison des ensembles et utilité

4.5.1 Définitions

Certaines questions se posent à ce stade. Comment trouver un critère plus général pour choisir l'un ou l'autre des ensembles de permutations. Pour cela, appelons arbitrairement G_n cet ensemble de permutation de tailles n, et notons dès à présent $U(G_n)$, l'utilité de cet ensemble. Cette utilité doit quantifier à quel point ces permutations permettent en partant d'une série temporelle quelconque de retrouver la densité du nombre de record d'une loi centrée symétrique continue à partir d'un marche respectant ces trois conditions.

Deux choix que je pense, valent le coup d'être explorés sont les suivants, on note D_X la densité de probabilité du nombre de record en utilisant H_0 ou G_n le groupe des permutations :

— En prenant $d(X,Y) = \sum_{i=1}^{n} |x_i - y_i|^k$, $U(G_n) = E[d(D_{H_0}, D_{G_n})]$
— En prenant $d(X,Y) = \sum_{i=1}^{n} x_i|x_i - y_i|^k$, $U(G_n) = E[d(D_{H_0}, D_{G_n})]$
— En prenant la distance de Kullback-Leibler

La première expression cherche à approcher uniformément la densité, tandis que la seconde cherche à l'approcher mieux aux valeurs les plus probables. Notons que la loi fortes des grands nombres nous permet de calculer informatiquement les utilités pour différents groupes.

4.5.2 Vérifications informatiques

Après tous ces calculs réalisés, voyons si en moyenne l'ensemble formé par $\{s = c^i \circ \sigma_n^j, (i,j) \in \mathbb{N}\}$ avec c une permutation circulaire, est meilleur qu'en

choisissant des permutations aléatoires. La difficulté principale repose sur le temps de calcul : le groupe G_n est de taille n^2, sur une série de taille 50 par exemple, avec une approximation de 100 séries dans la loi forte des grands nombres, il faut générer et traiter 250,000 séries et leurs records.

Pour $n = 30$ déjà, le résultat s'annonce mal. Peu importe la méthode utilisée parmi les 3 listées et k = 1 ou 2, l'utilité est dans une écrasante majorité des cas plus faible avec des permutations aléatoires (les écarts sont de l'ordre de 0.1%)

On pourrait croire que l'utilité ne révèle pas vraiment la qualité d'un groupe, ou bien que la distance est mal choisie. Cependant une seconde batterie de test, cette fois-ci en utilisant la théorie des tests statistiques, avec un niveau de risque de 10% et en prenant comme séries temporelles des bruits blancs a permis de fixer les idées.

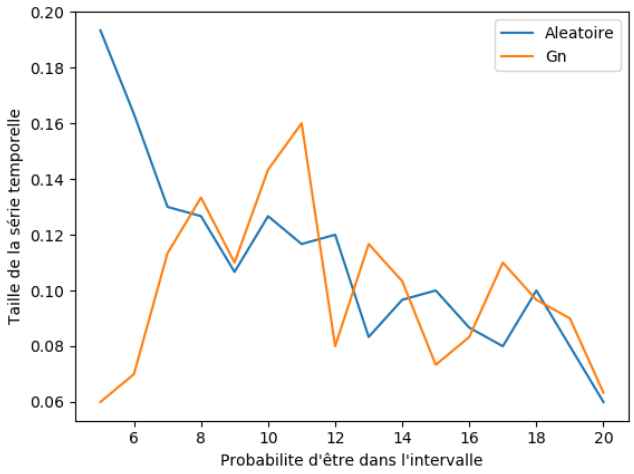


FIGURE 8 – Les titres des axes sont inversés mais les calculs sont longs

L'ensemble G_n n'est pas significativement meilleur que des groupes aléatoires peu importe la taille des séries temporelles.

Pour quantifier de manière plus précise l'écart entre ces densités de probabilités, nous avons pensé à utiliser la divergence de Kullback-leibler, aussi appelée entropie relative. Nous rappelons sa définition, en remarquant que nous nous plaçons ici dans le cas discret, puisque nous parlons de la densité de probabilité du nombre de records.

Définition 2 - Divergence de Kullback-Leibler.

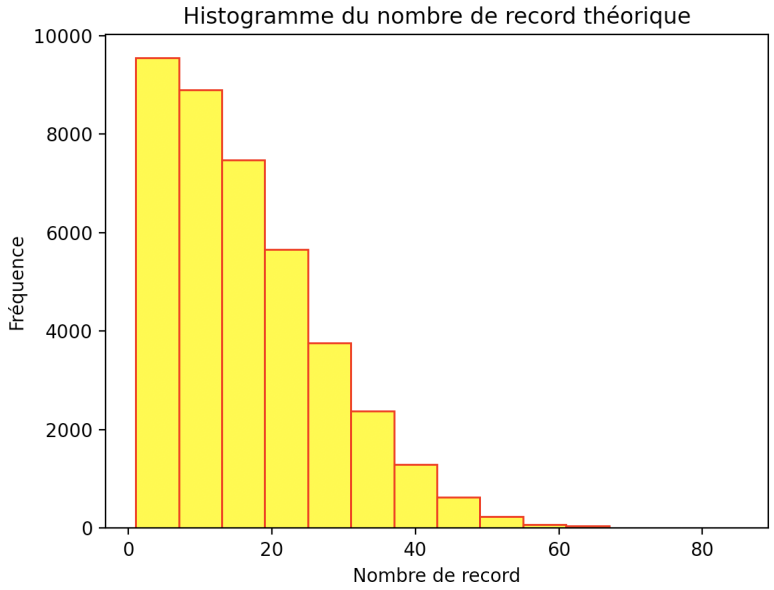
Soient P et Q deux densités de probabilités discrètes. La divergence de Kullback-Leibler se définit alors par :

$$D_{KL}(Q||P) = \sum_i Q(i) \ln\left(\frac{Q(i)}{P(i)}\right)$$

Méthodologie :

Dans un premier temps on calcule la densité de probabilité du nombre de records pour une loi normale centrée réduite, puis l'on trace son histogramme. Voici les résultats obtenus pour $n = 200$:

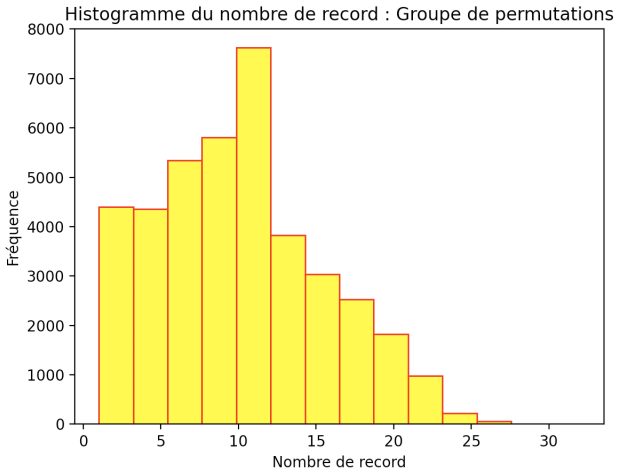
FIGURE 9 – Histogramme du nombre de records pour la loi normale centrée réduite



Dans un deuxième temps, on va calculer la densité de probabilité du nombre de records en appliquant cette fois les permutations de G_n. Pour cela on applique n fois la permutation σ_n aux sauts de notre série temporelle, puis au bout de n fois nous sommes revenus à la série temporelle de départ.
On décale alors tous les sauts d'un rang grâce à une permutation circulaire,

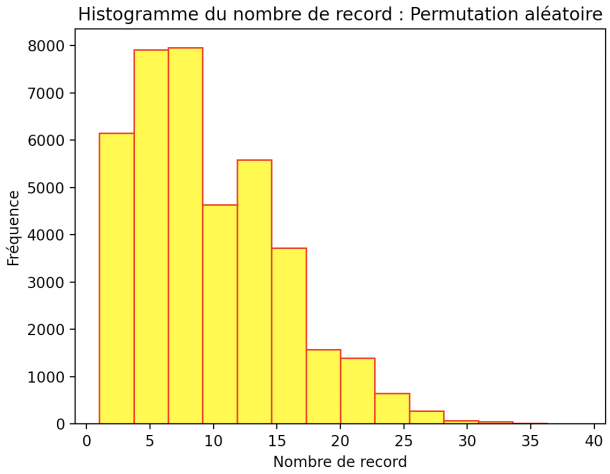
puis on recommence la première étape, n fois. On effectue alors un total de n^2 permutations. À chaque permutation des sauts, on note le nombre de records de la série temporelle. On note le tout dans une liste python puis on normalise pour obtenir notre densité de probabilité du nombre de record avec les permutations prises dans G_n. Voici l'histogramme obtenu :

FIGURE 10 – Histogramme du nombre de records en appliquant une permutation de G_n



Enfin nous allons comparer les permutations de G_n avec des permutations prises aléatoirement. Pour cela on permute les sauts n^2 fois avec une permutation aléatoire et on note à chaque étape le record obtenu dans une liste python, puis on normalise.
Voici les résultats que l'on obtient :

FIGURE 11 – Histogramme du nombre de records avec permutations aléatoires :



On peut alors déjà constater visuellement que prendre une permutation aléatoire est plus efficace que prendre une permutation de G_n puisque la densité de probabilité se rapproche d'avantage de la densité de probabilité théorique (sans permutation). Confirmons le alors avec la Divergence de Kullback Leibler que nous avons introduite un plus haut.
On note :
p_{th} : la densité de probabilité théorique
p_{rand} : la densité de probabilité avec une permutation random
p_G : la densité de probabilité avec une permutation de G_n
En effectuant plusieurs simulations numériques on obtient :

FIGURE 12 – Comparaison des divergences KL

```
1.8567594435472565
1.0662807400287997
(XXX) Francois-MacBook-Pro:~ francoisporcher$
2.2518398638589927
2.0315623147989146
(XXX) Francois-MacBook-Pro:~ francoisporcher$
1.394737838695663
0.9260231744015643
(XXX) Francois-MacBook-Pro:~ francoisporcher$
0.9956352632760817
0.6310308821584739
(XXX) Francois-MacBook-Pro:~ francoisporcher$
```

La première ligne correspond à $D_{KL}(p_{th}||p_G)$, la deuxième ligne est $D_{KL}(p_{th}||p_{rand})$ On constate alors que la Divergence KL est systématiquement plus grande pour les permutations de G_n. Cela vient confirmer ce que nous avons observé sur l'histogramme : une permutation aléatoire est plus efficace qu'une permutation de G_n. Malheureusement cela confirme également que le groupe auquel nous avions pensé est inefficace.

4.5.3 Analyse et prise de recul

Bien que l'analyse mathématiques nous pousse à penser que les permutations initiales forment une base solide et diversifiée, il ne faut pas perdre de vue que c'est avant tout les records que nous visons, et nous en savons toujours très peu sur comment ces derniers se comportent après permutations des sauts.

Cependant ces essais informatiques mettent en valeur la faible diversité de suites de sauts qu'apporte l'ensemble étudié. En effet, informatiquement, la moyenne du nombre de record sur une série temporelle en passant par les permutations est soit très faible, soit très forte. L'ensemble considéré accentue en réalité les différences. Ceci pourrait s'avérer utile si on voulait rejeter l'hypothèse H_1. Il faut donc utiliser des permutations aléatoires, et surtout essayer de trouver un bon ensemble. Il existe plusieurs méthodes pour cela.

Enfin, nous voyons apparaître en arrière plan l'influence des permutations sur la densité réelle du nombre de record. En réalité, si nous nous appuyons sur les permutations pour obtenir plus de données (et donc un intervalle d'acceptation de H0 plus étroit), il faut faire attention à la variance induite par ces dernières. En effet, la variance tend elle à augmenter la taille de l'intervalle, et elle dépend du groupe choisi. Une piste de recherche serait de relier cette variance et l'utilité d'un groupe. Ainsi, si la variance du groupe contrebalance l'avantage de plusieurs observations, il est préferable de ne choisir qu'un seul record et donc uniquement la série de base. Cependant, la perte d'information est alors grande, car nous n'exploitons qu'une valeur entière parmi les N-1 sauts.

4.6 Information et vraisemblance

Pour donner la confiance que l'on peut accorder à la décision du test vis-à-vis d'une série temporelle, il peut être intéressant de regarder une pseudo-

vraisemblance que l'on notera $I(P|L)$, c'est à dire, l'information de P sachant une loi/hypothèse L. Pour un test commun et sans grande conséquences comme le notre, cette notion se voit dénuée d'intérêt. Cependant, dans le développement d'un test financier, où les enjeux peuvent s'avérer très importants, il est important de quantifier la confiance du programme en sa décision. Toujours dans l'optique de se baser sur les records et les permutations, définissons dès à présent et plus formellement cette information.

$$I(P|L) = d(D_{G_n}, D_L)/U(G_n)$$

Ces notations ont déjà été explicitées dans la partie concernant l'utilité d'un groupe. On remarquera que le choix du groupe joue un rôle important, et son espérance est égale à 1 peu importe le groupe. On a donc un critère définissant une "bonne" série temporelle : si son information est inférieure à 1.

5 Conclusion et perspectives

Après une première phase d'appropriation du sujet, nous avons concentré nos efforts dans l'étude et la réalisation d'un test statistique permettant de tester le centrage de la loi des sauts d'une série temporelle (différence entre deux termes consécutifs de la série considérée). Bien que nous ne soyons pas parvenus à élaborer un test totalement fonctionnel et très performant dans le cadre de cette étude, les tests non paramétriques de ce type utilisant une fonction des records d'une série temporelle restent des outils très puissants de par leur caractère universel. En effet, il n'est pas nécessaire d'avoir des informations sur la loi des sauts pour appliquer ces tests, il suffit qu'elle vérifie (P). En revanche, lorsqu'on a des informations plus précises sur la loi des sauts, des tests paramétriques comme celui de Neyman-Pearson sont bien plus pertinents.
Une piste intéressante pour approfondir notre projet serait de réaliser un algorithme de prise de décision pour les investissements financiers basé sur le test de centrage mis à l'étude dans notre travail. L'objectif serait de pouvoir dire à un investisseur s'il faut investir ou non dans une action connaissant le cours de cette action dans le passé. L'historique du cours de l'action serait ici notre série temporelle à tester.

Références

[1] Benestad R. E. *How often can we expect a record event ?* 2003. http://regclim.met.no/results/Benestad03_CR25-1.pdf

[2] Douglas V. Hoyt *Weather 'records' and climatic change* 1981. https://link.springer.com/article/10.1007/BF02423217

[3] Chandler N. D. *The distribution and frequency of record values.* 1952. https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/j.2517-6161.1952.tb00115.x

[4] Barry C. Arnold, N. Balakrishnan, Haikady N. Nagaraja *Records.* 1998. https://b-ok.org/book/2818860/552c90

[5] M. S. Santhanam, Aanjaneya Kumar *Record Statistics of Equities and Market Indices.* 2017. https://link.springer.com/chapter/10.1007/978-3-319-47705-3_7

[6] Satya N. Majumdar1and, Robert M. Ziff *Universal Record Statistics of Random Walks and Levy Flights* 2008. https://arxiv.org/pdf/0806.0057.pdf

A Vérification informatique

Dans les trois figures ci-dessous, l'abscisse représente la longueur, ou nombre d'étapes, de la série temporelle, la courbe bleue est l'espérance théorique (donnée par l'article) de l'universalité considérée, et la courbe rouge l'espérance calculée par simulation numérique. Nous avons donc ici implicitement fait appel à la loi des grands nombres pour comparer l'espérance d'une universalité avec sa moyenne sur plusieurs séries temporelles.

FIGURE 13 – Écart type du nombre de record

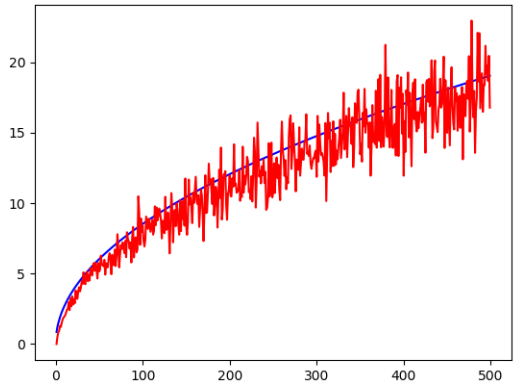


FIGURE 14 – Durée du plus court record

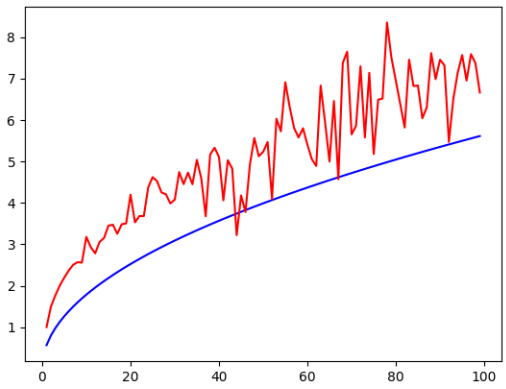


FIGURE 15 – Durée du plus long record