Une nouvelle façon de construire des statistiques non-paramétriques à partir de sommes cumulatrices

Damien Challet

Laboratoire de Mathématiques et Informatique pour la Complexité et les Systèmes, CentraleSupélec, Université Paris-Saclay, France,
damien.challet@centralesupelec.fr

Résumé.
La somme cumulatrice d'un échantillon de taille N définit une marche aléatoire de longueur N. Partant d'un résultat récent de Majumdar et al. (2008) sur les marches aléatoires sans tendance, cette contribution montre que le nombre de records supérieurs et inférieurs de la somme cumulatrice d'un échantillon d'espérance nulle mène à la définition d'un type de statistiques non-paramétriques pour des variables indépendantes et identiquement distribuées (i.i.d) et dont la distribution est symétrique autour d'une valeur donnée. Cela mène à des tests de même nature que ceux de Wilcoxon et Mann-Whitney et dont le temps de calcul d'ordre N ou $N \log N$ et de puissance comparable.

Les nouvelles statistiques ont ceci de particulier qu'elles dépendent de l'ordre dans lequel les valeurs échantillonnées apparaissent. Mais l'hypothèse de valeurs i.i.d permet de considérer n'importe quel ordre. Une question ouverte est donc de déterminer l'ordre optimal. Le choix de l'ordre est important : on montre par exemple que des permutations aléatoires sont meilleures pour des valeurs gaussiennes ou à support borné ; de plus un classement par ordre croissant et décroissant mène à des statistiques plus appropriées pour des échantillons exponentiels ou à queues grasses (Student, Weibull).

L'efficacité de la nouvelle statistique est illustrée par l'estimation des ratios de Sharpe d'actifs financiers. Elle permet d'estimer ce ratio (moyenne/déviation standard) sans calculer le moindre moment. On montre que contrairement aux méthodes habituelles, cet estimateur ne surestime pas les ratios de Sharpe en temps de crise.

Mots-clés. Statistiques non-paramétriques, sommes cumulatrices, nombre de records, drawdown, Sharpe ratio

Abstract.
The cumulative sum of N sample values is nothing else than a random walk of length N. Using a recent results of Majumdar et al. (2008) about unbiased random walks, this contribution shows that the number of upper and lower records of the cumulative sum of samples values of zero average leads to a new kind of nonparametric statistics for independent and identically distributed variables (i.i.d.) whose distribution is assumed to be symmetric around some location. This leads to tests i) of the same nature as Wilcoxon

and Mann-Whitney tests, ii) whose computation time scales as N or $N \log N$, and iii) of comparable power.

This new kind of statistics is quite peculiar in that the statistics depend on the order of sample values. However, the i.i.d. hypothesis allows us to choose how the sample values are indexed. An open question is what sample value order is optimal. The indexing order is clearly important. For example, this contributions shows that averaging the number of upper and lower records over several random index permutations leads to more powerful statistics for Gaussian and bounded variables, while sorting the sample values is more appropriate for (semi-)heavy-tailed variables (Exponential, Student, Weibull).

Estimating Sharpe ratios of financial assets is used as an illustration of the efficiency of this new kind of statistics, which makes it possible to estimate a ratio defined as the average price return divided by its standard deviation without computing any moment. This new method does not overestimate Sharpe ratios in times of crises, in contrast to moment-based methods.

Keywords. Nonparametric statistics, cumulative sums, number of records, drawdowns, Sharpe ratio

1 Introduction

Soit un échantillon x_i, $i = 1, \cdots, N$ tiré d'une distribution $P(x)$ et $\Xi = \{\xi_k = \sum_{i=1}^{k} x_i\}$ la somme cumulatrice de cet échantillon. Le nombre de records supérieurs de Ξ, noté par R_+, est le nombre de sauts du maximum courant, le premier point étant toujours considéré comme le premier record. Une façon équivalente de penser à R_+ est lié à la notion de drawdown en finance: le prix d'un titre donné est réputé être en drawdown s'il est inférieur au maximum courant. Etant donné qu'un nouveau point de données mène à un nouveau record ou à un drawdown, $N = R_+ T_{DD}$ où T_{DD} est le temps total passé en drawdown. Le nombre de records inférieurs est défini similairement par le nombre de sauts du minimum courant de Ξ. Si $P(x)$ est symétrique autour de 0, ($P(x) = P(-x)$) et est continue, alors la distribution $P(R_+ = k, N)$ est connue exactement pour toute valeur de N [1]:

$$P(R_+ = k, N) = \binom{2N - k + 1}{N} / 2^{2N-k+1}, \tag{1}$$

et ne dépend pas de la distribution $P(x)$. Pour des raisons de symétrie, $P(R_- = k, N)$ suit la même distribution. La nouvelle statistique est basée sur la quantité $R_0 = R_+ - R_-$ qui est par conséquent non-paramétrique. Elle permet de tester l'hypothèse

$$H_0 : P(\lambda + x) = P(\lambda - x), \quad \lambda \in \mathbb{R} \tag{2}$$

contre

$$H_1 : P(\lambda + x) \neq P(\lambda - x). \tag{3}$$

La distribution $P(R_0)$ est symétrique si H_0 est vrai, donc R_0 est non biaisé pour tout N. De plus, le nombre de records étant borné pour N donné, cette distribution converge vers une distribution normale centrée.

Le fait que R_0 dépende de l'indexation de l'échantillon va à l'encontre de l'hypothèse i.i.d: on est libre de mesurer R_0 après avoir choisi une permutation qui nous convient. Par conséquent, une question centrale (et non encore résolue) est de savoir quelle permutation est la plus appropriée. Deux possibilités seront présentées: d'une part, le moyennage de R_0 sur plusieurs permutations aléatoires et d'autre part le calcul de R_0 sur des échantillons triés par ordres croissant et décroissant. La première possibilité donne des statistiques plus puissantes pour des variables gaussiennes ou bornées, tandis que la seconde est plus appropriée pour des variables à queues semi-longues ou longues [2].

L'estimation d'un ratio de Sharpe donne une bonne illustration de l'efficacité de ce nouveau type de statistiques. Un tel ratio est défini comme μ/σ où μ est la moyenne de l'échantillon et σ sa déviation standard. Ce type de mesure est fort utilisé en finance, mais peu adapté à la nature des rendements des prix d'actifs financiers. La nouvelle statistique permet d'estimer ce ratio sans estimer le moindre moment. Autrement dit, le temps total passé en drawdown est une mesure robuste du ratio de Sharpe [3]. Le lecteur peut visiter la page web interactive écrite en Shiny `https://brillant.shinyapps.io/moment-free_Sharpe_ratio` pour en avoir une idée.

Bibliographie

[1] Majumdar SN, Ziff RM (2008), Universal record statistics of random walks and Lévy flights, Phys. Rev. Lett., 101(5):050601

[2] Challet D. (2015), Nonparametric record statistics of cumulative sample sums, arXiv:1502.05367, submitted

[3] Challet D (2015), Moment-free Sharpe ratio estimation from total drawdown durations, arXiv:1505.01333, submitted