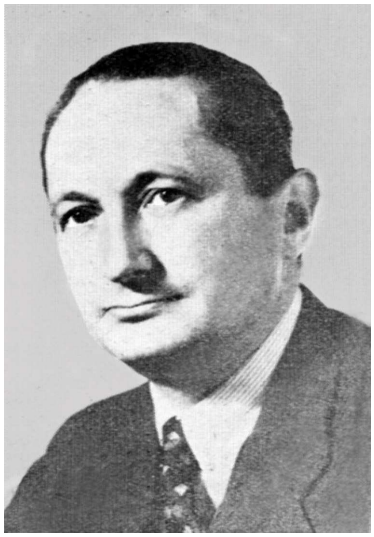
Statistique

Christophe Antonini[1], Olivier Teytaud[2], Pierre Borgnat[3], Annie Chateau[4], and Edouard Lebeau[5]

[1]Enseignant en CPGE, Institut Stanislas, Cannes
[2]Chargé de rechercher INRIA, Université d'Orsay, Orsay
[3]Chargé de recherche CNRS, ENS Lyon, Lyon
[4]Maitre de conférence, Université Montpellier-2, Montpellier
[5]Enseignant en CPGE, Lycée Henri Poincaré, Nancy

9 juillet 2023



Généralités sur les ensembles ordonnés.

1 Statistique

Cette très brève introduction aux statistiques ne peut remplacer la lecture d'ouvrages de référence. Nous introduirons ici simplement un peu de terminologie utile à la vie quotidienne. On pourra s'initier aux statistiques avec [**?**, **?**]. Pour un cadre plus financier, on pourra se pencher sur [1]. On pourra s'orienter vers la théorie des sondages avec [2].

1.1 Quelques notions élémentaires

1.1.1 Définitions

On considère ici x_i, pour $i \in [[1, n]]$, des nombres réels. Dans un grand nombre de cas, il sera intéressant de considérer le cas de n variables aléatoires, possiblement i.i.d.

On appelle **moyenne arithmétique** de n nombres réels x_1,...,x_n la quantité $\frac{\sum_{i=1}^n x_i}{n}$. On l'appelle aussi **moyenne** tout court lorsqu'il n'y a pas de risque de confusion, et on la note $\overline{x}$.
On appelle **moyenne géométrique** de n nombres réels x_1,...,x_n la quantité $\sqrt[n]{\Pi_i x_i}$ lorsqu'elle est définie.
On appelle **moyenne harmonique** des x_i l'inverse de la moyenne arithmétique des inverses des x_i :

$$\left(\frac{1}{n}\sum_{i=1}^{n}\frac{1}{x_i}\right)^{-1}$$

On appelle **moyenne quadratique** des x_i, lorsqu'ils sont positifs, la racine carrée de la moyenne arithmétique des carrés des x_i :

$$\sqrt{\frac{1}{n}\sum_{i=1}^{n}x_i^2}$$

On appelle **médiane** d'une mesure finie sur un espace ordonné tout élément x tel que la mesure de $\{y; y > x\}$ est égale à la mesure de $\{y; y < x\}$.
On appelle **effectif cumulé croissant** d'une distribution sur un espace ordonné la fonction qui à x associe la mesure de $\{y; y < x\}$, et **effectif cumulé décroissant** la fonction qui à x associe la mesure de $\{y; y > x\}$. Les effectifs cumulés croissants sont aussi appelés **effectifs cumulés** tout simplement. Ces notions sont définies lorsque les mesures correspondantes sont bien finies.
On appelle **k-ième percentile** d'une distribution sur $\mathbb{R}$ une valeur x telle que les effectifs cumulés en x représentent $k\%$ de la mesure de tout l'espace ; on parle aussi de quantile $k/100$ ou de quantile à $k\%$. On définit de même des **quartiles**, des **déciles** : premier quartile = quantile à 25 %, troisième quartile = quantile à 75 %, premier décile à 10 %, etc. On appelle **interquartile** la différence entre le troisième et le premier quartile.
On appelle **mode** ou **dominante** d'une distribution toute valeur x telle que la densité de probabilité en x soit localement maximale. S'il y a plusieurs modes la distribution est dite **plurimodale**.
On appelle **déviation** de x_i la valeur $x_i - \overline{x}$.
On appelle **écart moyen** la moyenne des $|x_i - \overline{x}|$; c'est donc $\overline{|x_i - \overline{x}|}$.
On appelle **variance** la moyenne des $(x_i - \overline{x})^2$; on la note souvent V ou σ^2. Pour des raisons de qualité d'estimation, on utilise en fait en général

$$\frac{1}{n-1}\sum(x_i - \overline{x})^2$$

comme variance approchée et non $\frac{1}{n}\sum(x_i - \overline{x})^2$.

En effet, l'équation 1.1.1 présente l'avantage d'être, si les x_i sont des variables aléatoires i.i.d., en moyenne égale à la variance de x_1, propriété que n'a pas l'équation 1.1.1 :

$$E_{x_1,\ldots,x_n}\frac{1}{n-1}\sum(x_i - \overline{x})^2 = E_{x_1}(x_1 - Ex_1)^2.$$

On dit alors que l'estimateur 1.1.1 est non-biaisé, alors que l'estimateur 1.1.1 est biaisé (il sous-estime la variance, à moins qu'elle soit nulle).
On appelle **écart type** ou **écart quadratique moyen** la racine carrée de la variance. On le note souvent σ ; $\sigma = \sqrt{V}$.
On procède à un **changement d'origine** lorsque l'on remplace les données x_i par les y_i définis par $y_i = x_i - C$, avec C une constante.
On procède à un **changement d'échelle** lorsque l'on remplace les données x_i par les y_i définis par $y_i = C.x_i$, avec C une constante.
On appelle **moment d'ordre** p des x_i par rapport à y la moyenne des $(x_i - y)^p$. Pour $p = 1$ et $y = 0$ il s'agit donc de la moyenne (arithmétique), pour $p = 2$ et $y = \overline{x}$ il s'agit de la variance.

1.1.2 Propriétés

On note les propriétés immédiates suivantes :
•Le logarithme de la moyenne géométrique est la moyenne arithmétique des $\log(x_i)$.
•Moyenne harmonique $\leq$ moyenne géométrique $\leq$ moyenne arithmétique $\leq$ moyenne quadratique.
•La moyenne arithmétique est peu sensible aux fluctuations d'échantillonnage.
•La médiane est peu sensible aux valeurs aberrantes.
•La somme des déviations est nulle.
•La variance V est aussi égale à $V = \overline{x^2} - \overline{x}^2$, avec $\overline{x^2}$ la moyenne arithmétique des x_i^2, et $\overline{x}^2$ le carré de la moyenne des x_i. On le prouve facilement en développant $\sum(x_i - \overline{x})^2$.
•Multiplier les données par C multiplie la moyenne arithmétique par C, la variance par C^2, et l'écart-type par C.
•Translater les données de C ajoute C à la moyenne arithmétique, et ne change ni la variance ni l'écart-type.

1.2 Applications des probabilités à l'échantillonnage

Cette partie ne se veut qu'une très brève introduction aux statistiques. Il est bien évident que dans le cadre de l'option probabilités de l'agrégation, il est indispensable de se référer à un livre plus complet. Pour une introduction concise on pourra consulter le livre "Thèmes de probabilité et statistiques" de P.S. Toulouse, Dunod 1999.

Soit X_1,..., X_n variables aléatoires indépendantes identiquement distribués L^1, ou du moins telles que le théorème central limite **??** sous une forme ou une autre est vérifié. Intuitivement, les X_i sont des mesures ; par exemple, on mesure la taille de 50 français pour évaluer la taille moyenne des français. L'intérêt des probabilités va être de fournir des bornes sur l'erreur commise par une telle évaluation.

On se donne donc $m = \frac{1}{n}(X_1 + X_2 + ... + X_m)$. On cherche $[a, b]$ tel que $M = E(X)$ soit compris dans $[a, b]$. Il faut alors noter que bien entendu, on ne peut être certain que M soit dans l'intervalle $[a, b]$, quel que soit l'intervalle que l'on donne, simplement au vu des X_i. Il est toujours possible que l'on ait été particulièrement malchanceux dans les tirages des X_i et que la moyenne soit très différente de ce que l'on suppose au vu des données. On doit donc plutôt donner α un réel (petit de préférence) et z tel que avec probabilité $1 - \alpha$, pour toute loi de X_1, $|m - M| \leq z$ soit vrai. a et b seront alors $m - z$ et $m + z$ respectivement.

Concrètement on procède comme suit :

•On évalue (empiriquement) l'écart type σ de X_i.

•On repère t_α tel que $P(|N| \leq t_\alpha) = 1 - \alpha$, avec N loi normale centrée réduite (espérance nulle et écart-type 1). Les valeurs de t_α sont tabulées (il s'agit simplement de la fonction de répartition de la loi normale). Le plus courant est de choisir $\alpha = 0.05$, t_α étant alors environ égal à 2.

•On détermine $a = m - t_a\sigma/\sqrt{n}$ et $b = m + t_\alpha\sigma/\sqrt{n}$.

•On peut alors écrire que, au **seuil de confiance** α, M est compris entre a et b. Ceci constitue un **intervalle de confiance**. Il faut bien noter le caractère approximatif (asymptotique) de cette conclusion. On pourrait s'affranchir de cette hypothèse asymptotique, en utilisant des inégalités exactes, par exemple en utilisant l'inégalité de Hoeffding, ou de Chernoff.

Il faut bien cerner la notion de seuil de confiance. On ne se trompe, au pire cas, que dans $100 \times (1 - \alpha)\%$ des cas en utilisant ce système (à l'approximation asymptotique près).

On peut ainsi dire que la moyenne arithmétique est un estimateur de l'espérance ; que la formule 1.1.1 est un estimateur non-biaisé de la variance ; que la formule 1.1.1 est un estimateur biaisé de la variance.

On peut citer les développements suivants :

— le cas des petits échantillons ($n < 30$). Il n'est alors pas adéquat d'utiliser la loi normale comme approximation asymptotique. Il faut alors utiliser la loi de Student, sous certaines hypothèses (hypothèse de normalité des x_i, i.e. hypothèse selon laquelle les x_i sont distribués selon une distribution normale).

— le cas où l'on ne s'intéresse pas à la probabilité pour que la moyenne soit *mal* évaluée, mais à la probabilité pour que la moyenne soit *sur*-évaluée. Il suffit, pour construire un intervalle de confiance de la forme $]-\infty, b]$, de constater que $P(N > t) = \frac{1}{2}P(|N| > t)$ pour toute variable aléatoire N symétrique, et en particulier donc la loi normale. On parle alors de test unilatéral (ou d'interval de confiance unilatéral), au lieu d'un test bilatéral.

— le cas de X_i à valeur dans $\{0, 1\}$, que l'on peut simplifier et étudier facilement sans hypothèse asymptotique ; plus généralement le cas de variables bornées peut aussi être commodément étudié sans hypothèse asymptotique (voir les inégalités de Hoeffding ou de Chernoff).

— le cas où l'on n'étudie pas la moyenne des X_i mais leur max.

— le cas de X_i non indépendants.

— le cas de X_i non identiquement distribués.

— le bootstrap, comme moyen d'évaluer des intervalles de confiance et des biais de manière très astucieuse.

— le test du χ^2 et celui de Kolmogorov-Smirnov sont deux développements indispensables des statistiques. Il permettent de tester le fait que deux échantillons proviennent d'une même distribution, ou qu'un échantillon est bien distribué suivant une certaine distribution de probabilité.

Ces études et d'autres encore constituent la théorie des tests et font appel à des variantes parfois beaucoup plus difficiles du théorème central limite (par exemple le bootstrap utilise des extensions difficiles de ce théorème). La façon d'échantillonner, de manière plus sophistiquée, est aussi un développement important des statistiques : on peut formaliser l'intuition selon laquelle il est plus important d'avoir un grand nombre de points dans les catégories les plus variables. Outre cette aspect, consistant à biaiser l'échantillonnage pour améliorer la précision d'estimateurs, il existe aussi des méthodes dites de quasi-monte-carlo, notamment pour les espaces continus : plutôt qu'échantillonner de manière aléatoire simple [1] et uniforme un domaine $[0, 1]^d$, pour calculer

1. Echantillonnage aléatoire simple = échantillonnage i.i.d.

l'espérance de $f(X)$ avec X une variable aléatoire uniforme sur $[0,1]^d$, on peut parfois choisir les $(x_i)_{i\in[[1,n]]}$ de manière "plus régulière" dans $[0,1]^d$ qu'en les tirant au sort. Ceci est le principe de base des méthodes dites de quasi-Monte-Carlo ; on parle de suites à faible-discrépance pour ces suites de points très régulières.

Références

[1] G. Demange, J.-C. Rochet, *Méthodes mathématiques de la finance*, Economica, 2ème édition, 1997.

[2] Y. Tillé, *Théorie des sondages. Echantillonnage et estimation en populations finies*, Dunod, 2001.