LA STATISTIQUE AVEC UN TABLEUR

Bernard Burtschy
ENST
46 rue Barrault
75013 PARIS

Le tableur a révolutionné l'informatique et l'informatique a révolutionné la statistique La relation est-elle transitive ?

Première application historique du micro-ordinateur et moteur essentiel sinon exclusif dans la conquête des utilisateurs professionnels, le tableur a été le cheval de Troie qui a permis de détrôner l'informatique centralisée, ses lourdeurs et ses pompes. Il n'a pas si longtemps que cela, il fallait une armée de programmeurs Cobol pour établir et maintenir un simple tableau de bord consolidé. Cette opération, comme tant d'autres, s'effectue maintenant en quelques heures, sur un coin de bureau. De fait, le tableur est devenu l'outil à tout faire, et même le symbole, du cadre moderne.

Que l'informatique ait révolutionné l'analyse des données, composante importante (mais non exclusive) de la statistique, il n'y a pas besoin de faire un dessin au lecteur assidu de la revue de MODULAD. Il en est intimement persuadé. La vague de fond commence à toucher la statistique en général.

Quelles sont les relations entre le tableur et la statistique ? Les deux travaillent sur des tableaux de données qui sont leurs matières de base. L'esprit est différent. Le principe de l'analyse des données est de prendre un tableau de données et de « l'essorer » à fond. Les logiciels spécialisés proposent moult méthodes de traitements, mais font peu de cas de l'entrée et de la gestion des données. A quelques exceptions près, les manipulations des lignes ou des colonnes relèvent du parcours du combattant. Les logiciels spécialisés de statistique sont des pachydermes en matière de gestion de données. Dans bien des cas, il vaut mieux savoir programmer pour se créer un logiciel jetable *ad hoc* qui ne servira qu'une seule fois.

Le tableur quant à lui, est spécialisé dans la gestion des données sous toutes ses formes. Saisie des données, importations, créations de formules et de liaisons, outils de mise en forme, tris, filtrages, création de tableaux croisés, n'ont pas de secrets et s'effectuent aisément par des simples clicks de la souris ou à l'aide d'assistants. Le traitement statistique n'est pas son fort. Mais, fait nouveau, ces logiciels ne se cantonnent plus à la gestion des données. Version après version, les outils graphiques et statistiques s'amoncellent, sans bruit et sans que les statisticiens s'en aperçoivent.

Le marché des outils logiciels pour statisticiens (Tab.1) se partage maintenant en deux. D'un côté, il existe, depuis quelque temps déjà, des logiciels riches en méthodes statistiques. Réservés à une élite qui est et se sent dépositaire d'un savoir-faire, ils leur permettent de soumettre les données à un intense pilonnage de méthodes afin qu'elles expriment jusqu'à la dernière goutte, toute la substantifique information qu'elles recèlent. Ces logiciels sont à confier aux mains de statisticiens professionnels aguerris.

Par ailleurs, il existe tout un ensemble de données qui ne nécessitent qu'une analyse statistique rapide et légère. Les logiciels professionnels de statistique sont des marteaux-pilons pour ces données. Ils sont aussi beaucoup trop lourds pour une utilisation pédagogique. Les élèves passent plus de temps à apprendre le jargon des logiciels statistiques qu'à assimiler les méthodes. Faute de logiciels adaptés, beaucoup de données ne sont pas analysées du tout et restent en jachère.

Les tableurs tentent de combler ce vide que les statisticiens n'ont pas su combler avec l'adage suivant : mieux vaut une analyse rapide et succincte que pas d'analyse du tout. Pour ce faire, et c'est là que réside la clé de leur réussite, ils utilisent ce support naturel de gestion des données qui est le tableur. Les logiciels spécialisés nécessitent des conversions acrobatiques des données entre logiciels, conversions qui sont pleines d'embûches et chronophages. Ces conversions sont acceptables (pour le moment) lors d'une utilisation intensive des données, parce que l'investissement est compensé par la qualité et la variété des méthodes statistiques. Mais l'investissement n'est pas rentable pour une utilisation ponctuelle.

Type d'analyse	Logiciel
intensive	spécialisé
rapide	tableur

Tab. 1 Le marché des tableurs

Cette approche par le tableur ouvre la statistique à un public beaucoup plus large. Cet aspect n'a pas échappé aux développeurs des tableurs qui facilitent de plus en plus la liaison entre les données et les outils. D'ailleurs, seule la peur de la loi antitrust empêche une entreprise protéiforme comme Microsoft de trop s'immiscer dans les logiciels statistiques. Dominant déjà largement les tableurs, au point de faire de son tableur vedette Excel la référence en la matière, la tentation est d'autant plus forte que son appétit est féroce.

Le paysage de la statistique avec un tableur, inexistant il y a encore quelques années, a évolué en profondeur. Le déclic a probablement été la mise à la disposition de Visual Basic, le langage de programmation d'Excel, qui permet de créer des macros. Une macro est une série d'instructions qui commande à un tableur des tâches bien définies, en utilisant le riche environnement (données, graphique) du tableur. Ainsi, une analyse en composantes principales se prête aisément à une macro. Il existe maintenant des ensembles de macros toutes prêtes pour les statisticiens. La stratégie de Microsoft est d'intégrer les macros les plus demandées dans des fonctions internes.

Dans l'état actuel, le statisticien dispose d'un ensemble d'outils sous formes de fonctions, de macros et d'ouvrages (Tab. 2). La palette est beaucoup plus riche en statistique qu'en analyse des données au sens français, prééminence anglo-saxonne oblige.

Domaines	Disponible
Statistique	Fonctions Excel
	Macros spécialisées
	Ouvrages pédagogiques
Analyse de données	Macros spécialisées

Tab. 2 Disponibilités des outils spécialisés pour Excel.

Les outils statistiques

Le vocable statistique comprend les lois statistiques usuelles, le théorème de la limite centrale, les tests d'hypothèses (χ^2 compris), la régression et l'analyse de variance. En gros, c'est ce qu'on attend d'un bon cours d'initiation à la statistique.

La gestion des données

Les données peuvent être importées dans le tableur à l'aide d'un large ensemble de procédures. Dans sa dernière version, le module d'importation d'un fichier ASCII avec son module de visualisation immédiate, est particulièrement réussi. La plupart des utilisateurs bloquent dès l'importation des données et se découragent. Nombre de logiciels statistiques ont été irrémédiablement abandonnés à ce niveau. Tout a été fait dans le tableur pour que cela ne soit plus le cas. L'importation des données est un régal.

Une cellule contient trois types d'informations : valeur, texte et formule. Les formules donnent toute la puissance au tableur et une grande souplesse. Le simple fait de démarrer par le signe = indique qu'il s'agit d'une formule. Les adresses des cellules sont connues en coordonnées absolues, mais aussi relatives. Prenons l'exemple suivant :

= somme(B1..E1)

Cette fonction fait la somme des cases de B1 à D1. Ces formules couvrent l'ensemble des fonctions mathématiques, mais aussi statistiques.

La manipulation des données avec leurs caractéristiques (labels, dates) a aussi considérablement progressé. Les données qualitatives peuvent être entrées en tant que telles et la création de tableaux croisés s'est beaucoup simplifiée. La gestion de données à trois dimensions est aussi très naturelle. Les données se compulsent à la manière d'un livre. Les logiciels additionnels se servent bien de cette propriété. Ainsi, pour une analyse en composantes principales, valeurs propres, vecteurs propres, graphiques, se sélectionnent par onglets.

Graphiques

La communauté des cadres financiers et de gestion représente l'essentiel du marché du tableur. Tout a été fait pour faciliter leurs tâches. Indépendamment des évitables camemberts, les histogrammes et graphiques X-Y sont particulièrement simples à utiliser.

Le statisticien regrette l'absence des boites à moustaches, des graphiques en tiges et feuilles et des graphiques en semis de points. Mais il est assez facile de se procurer les macros correspondantes qui sont publiées et maintenant livrées sur disquette ou sur Internet avec les ouvrages spécialisés sur les tableurs statistiques (Middleton, 1995).

Les assistants ont aussi fait beaucoup de progrès. La création d'un graphique est maintenant immédiate et Windows permet de disposer de tous les pilotes imprimantes permettant d'en obtenir un support papier. Tous ceux qui rageaient de voir leurs graphiques sur leur écran sans pouvoir les imprimer, apprécieront.

La liaison dynamique entre données et graphique ouvre des horizons nouveaux. En enlevant un point atypique des données, la droite de régression bouge en proportion sur le graphique. Spectaculaire ! De même, la suppression d'un point sur le graphique modifie directement le calcul de la corrélation. L'illustration d'un cours sur la robustesse est particulièrement facile.

Les fonctions statistiques

Excel est livré avec soixante-dix fonctions statistiques (Tab.3). On y trouve toutes les mesures statistiques usuelles, les diverses utilisations des lois statistiques ainsi que tous les éléments de la régression. Les assistants permettent l'utilisation de ces fonctions sans avoir à les mémoriser.

Dans la pratique, le tableur peut être utilisé pour un cours d'initiation à la statistique, en prenant toutefois que certaines options prennet des libertés avec les principes théoriques. Ainsi le premier quartile est défini par l'observation de rang (n+3)/4 et le troisième quartile par l'observation de rang (3n+1)/4.

Pour toute utilisation un peu approfondie des fonctions statistiques, on aura intérêt à se procurer un des ouvrages spécialisés (Middleton, 1995; Pelosi et al., 1996). Ils sont d'excellents conseils et donnent une foule de « trucs » plus utiles les uns que les autres.

Quelques « add-ins » statistiques sont livrés avec le tableur. Ils couvrent la régression, l'analyse de variance, les tests. Ils sont moins fiables que les fonctions, mais bénéficient de l'environnement du tableur. Comme ce sont des macros qui sont exécutées à part, il n'y a pas de liaison dynamique avec les données. Ces macros ne sont pas exemptes d'erreurs. Les macros corrigées peuvent être téléchargées à partir du site de Microsoft (http://www.microsoft.com/).

Pour une utilisation plus sûre, il vaut mieux acquérir des add-ins bien documentés (Cooney, 1995). Pour une utilisation pédagogique, on pourra aussi acquérir les feuilles de calculs toutes prêtes (Hunt et Tyrrell, 1995).

Enfin, sachez que qu'une association des utilisateurs statistiques d'Excel s'est formée sous le nom d'ASSUME (Association of Statistics Specialists Using Microsoft Excel). Elle a pour objectif d'améliorer les fonctions statistiques d'Excel. Le Web bruisse de toutes les applications possibles.

Pour une utilisation plus sophistiquée, commencent à apparaître des outils plus complets d'analyse des données, soit sous forme Shareware (X-lstat), soit sous forme de logiciels commerciaux (Statbox). Ils intègrent analyse en composantes principales, analyse de correspondances et classifications. Nous en ferons le point ultérieurement.

Conclusion

En quelques années, le tableur Excel s'est enrichi de nombreuses fonctions statistiques. Des ouvrages, des macros permettent d'utiliser le tableur pour les travaux courants de statistique descriptive au sens large. Alors, à vos tableurs !

Bibliographie

Callender, J.T. and Jackson, R (1995). Exploring Probability and Statistics using Spreadsheets. Prentice Hall.

Cooney, J (1995). ASTUTE, add-ins for Microsoft Excel. DDU Software, University of Leeds.

Hunt N. (1996). Teaching Statistics with Excel 5.0. Newsletter Maths&Stats.

Hunt, D N and Tyrell S.E. (1995). Discus - Discovering Important Satistical Concepts Using Spreadsheets. Coventry University Enterprises Ltd.

Middleton M.R. (1995). Data Analysis using Microsoft Excel 5.0. Belmont : Duxbury Press.

Pelosi M.K., Dandifer T.M., Letkowski J.J. (1996). Doing Statistics with Excel 5.0 for Windows. New York : John Wiley and Sons.

Tab.3 Les fonctions statistiques d'Excel 7.0

BETA.INVERSE	Renvoie, pour une probabilité donnée, la valeur d'une variable aléatoire suivant une loi Bêta
CENTILE	Renvoie le k-ième centile des valeurs d'une plage
CENTREE.REDUITE	Renvoie une valeur centrée réduite
COEFFICIENT.ASYMETRIE	Renvoie l'asymétrie d'une distribution
COEFFICIENT.CORRELATION	Renvoie le coefficient de corrélation entre deux séries de données
COEFFICIENT.DETERMINATION	Renvoie la valeur du coefficient de détermination R^2 d'une régression linéaire
COVARIANCE	Renvoie la covariance, moyenne des produits de deux variables centrées sur leurs espérances mathématiques
CRITERE.LOI.BINOMIALE	Renvoie la plus petite valeur pour laquelle la distribution binomiale cumulée est supérieure ou égale à une valeur critère
CROISSANCE	Renvoie les valeurs de y en fonction d'une courbe exponentielle
DROITEREG	Renvoie les paramètres d'une tendance linéaire
ECART.MOYEN	Renvoie la moyenne des écarts absolus des observations par rapport à leur moyenne arithmétique
ECARTYPEP	Calcule l'écart-type d'une population à partir de la population entière
ECARTYPE	Evalue l'écart-type d'une population en se basant sur un échantillon de cette population
ERREUR.TYPE.XY	Renvoie l'erreur-type de la valeur y prévue pour chaque x de la régression
FISHER.INVERSE	Renvoie la transformation de Fisher inverse
FISHER	Renvoie la transformation de Fisher
FREQUENCE	Renvoie une distribution fréquentielle sous forme de matrice verticale
GRANDE.VALEUR	Renvoie la k-ième plus grande valeur d'une série de données
INTERVALLE.CONFIANCE	Renvoie l'intervalle de confiance pour la moyenne d'une population
INVERSE.LOI.F	Renvoie, pour une probabilité donnée, la valeur d'une variable aléatoire suivant une loi F

KHIDEUX.INVERSE	Renvoie, pour une probabilité unilatérale donnée, la valeur d'une variable aléatoire suivant une loi du Khi-deux
KURTOSIS	Renvoie le kurtosis d'une série de données
LNGAMMA	Renvoie le logarithme népérien de la fonction Gamma, G(x)
LOGREG	Renvoie les paramètres d'une tendance exponentielle
LOI.BETA	Renvoie la probabilité d'une variable aléatoire continue suivant une loi de probabilité Bêta
LOI.BINOMIALE.NEG	Renvoie la probabilité d'une variable aléatoire discrète suivant une loi binomiale négative
LOI.BINOMIALE	Renvoie la probabilité d'une variable aléatoire discrète suivant la loi binomiale
LOI.EXPONENTIELLE	Renvoie la probabilité d'une variable aléatoire continue suivant une loi exponentielle
LOI.F	Renvoie la probabilité d'une variable aléatoire suivant une loi F
LOI.GAMMA.INVERSE	Renvoie, pour une probabilité donnée, la valeur d'une variable aléatoire suivant une loi Gamma
LOI.GAMMA	Renvoie la probabilité d'une variable aléatoire suivant une loi Gamma
LOI.HYPERGEOMETRIQUE	Renvoie la probabilité d'une variable aléatoire discrète suivant une loi hypergéométrique
LOI.KHIDEUX	Renvoie la probabilité d'une variable aléatoire continue suivant une loi unilatérale du Khi-deux
LOI.LOGNORMALE.INVERSE	Renvoie l'inverse de la probabilité pour une variable aléatoire suivant la loi lognormale
LOI.LOGNORMALE	Renvoie la probabilité d'une variable aléatoire continue suivant une loi lognormale
LOI.NORMALE.INVERSE	Renvoie, pour une probabilité donnée, la valeur d'une variable aléatoire suivant une loi normale
LOI.NORMALE.STANDARD.INVERSE	Renvoie, pour une probabilité donnée, la valeur d'une variable aléatoire suivant une loi normale standard (ou centrée réduite)
LOI.NORMALE.STANDARD	Renvoie la probabilité d'une variable aléatoire continue suivant une loi normale standard (ou centrée réduite)
LOI.NORMALE	Renvoie la probabilité d'une variable aléatoire continue suivant une loi normale

LOI.POISSON	Renvoie la probabilité d'une variable aléatoire suivant une loi de Poisson
LOI.STUDENT.INVERSE	Renvoie, pour une probabilité donnée, la valeur d'une variable aléatoire suivant une loi T de Student
LOI.STUDENT	Renvoie la probabilité d'une variable aléatoire suivant une loi T de Student
LOI.WEIBULL	Renvoie la probabilité d'une variable aléatoire suivant une loi de Weibull
MAX	Donne le plus grand nombre de la liste d'arguments
MEDIANE	Renvoie la valeur médiane des nombres
MIN	Renvoie la valeur minimale des nombres
MODE	Renvoie la valeur la plus fréquente d'une série de données
MOYENNE.GEOMETRIQUE	Renvoie la moyenne géométrique
MOYENNE.HARMONIQUE	Renvoie la moyenne harmonique
MOYENNE.REDUITE	Renvoie la moyenne de l'intérieur d'une série de données
MOYENNE	Renvoie la moyenne des nombres
NBVAL	Détermine combien de valeurs sont comprises dans la liste des arguments
NB	Détermine combien de nombres sont compris dans la liste des arguments
ORDONNEE.ORIGINE	Renvoie l'ordonnée à l'origine de la droite de régression linéaire
PEARSON	Renvoie le coefficient de corrélation d'échantillonnage de Pearson
PENTE	Renvoie la pente d'une droite de régression linéaire
PERMUTATION	Renvoie le nombre de permutations pour un nombre donné d'objets
PETITE.VALEUR	Renvoie la k-ième plus petite valeur d'une série de données
PREVISION	Renvoie une valeur suivant une tendance linéaire
PROBABILITE	Renvoie la probabilité que des valeurs d'une plage soient comprises entre deux limites
QUARTILE	Renvoie le quartile d'une série de données
RANG.POURCENTAGE	Renvoie le rang en pourcentage d'une valeur d'une série de données

RANG	Renvoie le rang d'un nombre dans une liste d'arguments
SOMME.CARRES.ECARTS	Renvoie la somme des carrés des écarts
TENDANCE	Calcule les valeurs par rapport à une tendance linéaire
TEST.F	Renvoie le résultat d'un test F
TEST.KHIDEUX	Renvoie le test d'indépendance
TEST.STUDENT	Renvoie la probabilité associée à un test T de Student
TEST.Z	Renvoie la valeur bilatérale P du test Z
VAR.P	Calcule la variance d'une population en se basant sur la population entière
VAR	Estime la variance d'une population en se basant sur un échantillon de cette population