JOURNAL DE LA SOCIÉTÉ STATISTIQUE DE PARIS

CHARLES A. BICKING
SIMONE LEMARIÉ

Importance de la rapidité de traitement des données pour la statistique

Journal de la société statistique de Paris, tome 104 (1963), p. 68-73
<http://www.numdam.org/item?id=JSFS_1963__104__68_0>

IMPORTANCE DE LA RAPIDITÉ DE TRAITEMENT DES DONNÉES POUR LA STATISTIQUE

Introduction

A la naissance des calculateurs électroniques, l'utilité de ceux-ci pour l'ingénieur et le physicien en ce qui concerne la résolution de problèmes numériques était une considération de première importance. Bien que leur multiplication astronomique récente soit due en grande partie à leur emploi en comptabilité commerciale, les applications scientifiques continuent également à se développer. Ce sont les utilisations des calculateurs dans le domaine scientifique, que ce soit en engineering ou dans les sciences sociales, qui ont le plus d'importance pour la statistique. Ceci parce que les scientifiques s'intéressent à la résolution de problèmes et que la statistique est une science de résolution de problèmes.

C'est surtout dans le domaine scientifique que les matériels électroniques de mesures et d'enregistrement de données ont progressé. Le développement de l'instrumentation a précédé celui des calculateurs et contribué à nous l'imposer. Il est significatif, pour les statisticiens, que l'enregistrement des données et le calcul aient été automatisés en même temps. En conséquence, nous traiterons, dans cet article, à la fois d'un système d'enregistrement automatique des données et des programmes de statistique pour calculateurs. Toutes les fois qu'on doit analyser de grandes quantités de données il est souhaitable de savoir comment celles-ci ont été recueillies et d'avoir quelque idée sur leur validité avant de commencer de longs calculs. C'est pourquoi on prêtera une attention particulière au type d'examen préliminaire qui devrait être fait avant de passer au calcul principal.

Nous décrirons d'abord plus spécialement un des premiers appareillages mobiles d'enregistrement, prévu pour recueillir des données venant d'expériences à l'échelle usine ou à l'échelle laboratoire ou encore des données venant d'une application de contrôle de qualité par une méthode statistique. Ensuite nous décrirons la composition d'un programme général

d'analyse statistique pour calculateur à grande vitesse. Enfin, nous donnerons la composition d'une bibliothèque de programmes permettant une large gamme d'analyses statistiques.

PROBLÈME GÉNÉRAL DE L'EXAMEN DES DONNÉES

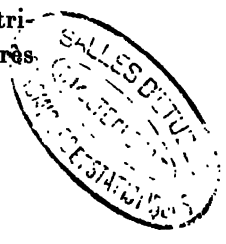
En premier lieu, dans une analyse statistique, on devrait s'inquiéter de la manière dont ont été élaborées les données. L'analyse de celle-ci apporte souvent des déceptions si le statisticien n'a pas pris la précaution de se renseigner auprès de l'ingénieur ou du savant quant à la signification technique des variables dans une expérience. Par exemple, en cherchant une corrélation entre les impuretés chimiques et la dureté Brinell, on découvrit trop tard que les données n'avaient pas toutes été prises au même point de test. Il était trop tard car tout le travail sur calculateur électronique avait été fini avant la découverte de cette non-homogénéité d'origine. On avait publié un rapport qu'il fallut corriger non sans quelque ennui pour ses auteurs. Ceci ne serait probablement pas arrivé si tout le travail avait été exécuté par un même individu. Cependant, au stade de complexité actuel au moins, beaucoup de personnes y sont mêlées : des ingénieurs, des expérimentateurs, des statisticiens, des programmeurs et des opérateurs de calculateurs. Il serait peut-être bon de recommander, dès le début, que le statisticien lise au moins les carnets de notes de l'expérimentateur.

Une fois que l'on a les données en mains, la première étape devrait être une opération rapide d'examen, soit visuel soit électronique. On comprendra mieux ce que nous entendons par là si les éléments de cette opération sont expliqués en termes de techniques familières d'analyse sans calculateur. On peut parcourir simplement et rapidement les données pour détecter les erreurs avant de passer un temps précieux à les analyser formellement. Ceci suppose expérience et pratique de la part du statisticien, qui peuvent être mises en parallèle avec l'expérience et la pratique de l'ingénieur ou du savant dans le domaine de qui les données ont été recueillies. Il existe un certain nombre de techniques rapides qui peuvent être utilisées, au besoin successivement.

Une possibilité qui vient d'elle-même à l'esprit consiste à reporter les données sur des graphiques pour détecter toute association proche entre des couples de variables (porter les valeurs d'une variable sur l'axe des x et les valeurs correspondantes de l'autre variable sur l'axe des y). Dans la corrélation entre les impuretés et la dureté Brinell citée plus haut, des graphiques furent établis et ils étaient en désaccord avec les informations fournies par les ingénieurs sur des relations connues. Ceci aurait dû faire soupçonner une erreur. Les statisticiens partagèrent le blâme avec les ingénieurs qui auraient aussi dû montrer un peu plus d'esprit critique à l'égard de leurs données. Le péché du statisticien, entre parenthèses, fut un péché par omission, celui des ingénieurs par commission car ils s'étaient en fait écartés des spécifications pour prendre une partie des échantillons à un autre stade du processus où ils étaient plus facilement accessibles à la mesure.

Pour en revenir à notre sujet, cependant, un graphique rapidement établi, sur lequel une courbe approchée peut être tracée à l'œil, peut conduire, s'il met en évidence des relations suffisamment étroites, à des corrélations simples ou à une corrélation multiple entre toutes les variables. Une des tentations du temps présent, avec ses calculateurs à grande vitesse capables de donner les résultats en quelques minutes, est de continuer tout d'une traite. L'expérience conseille de s'arrêter et de réfléchir entre les différentes étapes, même si l'on peut tout faire en une seule fois compte tenu des limitations de coût.

Il peut parfois être indiqué d'utiliser une autre technique : les graphiques des distributions de fréquences. On peut vouloir vérifier simplement que les données sont à peu près



normales. Ou bien on peut obtenir réellement des indications précieuses si la distribution est bimodale (données mélangées), ou rectangulaire (données chaotiques), ou oblique. Dans le dernier cas, on peut se demander si les erreurs de machine et les erreurs instrumentales n'obéissent pas à une loi linéaire, de même que les logarithmes des erreurs humaines, ou bien si l'on se trouve en présence d'un grand nombre de petites sources d'erreurs qui se multiplient plutôt qu'elles ne s'ajoutent.

Les lecteurs familiarisés avec les techniques statistiques de contrôle de qualité sauront comment interpréter les graphiques de contrôle pour l'analyse des données (par opposition à leur utilisation pour le contrôle). Des règles en usage dans le monde entier traitent de cette application des graphiques de contrôle. Les graphiques d'écarts, par exemple, peuvent être utilisés pour étudier l'homogénéité de jeux de données expérimentales; on peut encore déduire d'un graphique de moyenne l'influence du temps sur les données.

Le fait d'avoir un calculateur à sa disposition pour examiner les données ouvre d'intéressantes possibilités. La première est celle d'enregistrer automatiquement les données de manière à ce que celles-ci soient déjà sous une forme directement utilisable par le calculateur.

Enregistrement automatique des données

Le matériel électronique d'enregistrement de données nous oblige à considérer non seulement le recueil de ces données mais aussi les techniques d'analyse à appliquer et la conversion de ces données en une forme utilisable par le calculateur.

Les convertisseurs analogique/digital permettent de convertir directement les sorties d'un certain nombre d'instruments en des résultats numériques. La position d'une valve, par exemple, peut être transmise comme une pression d'air, convertie en une tension et affichée sur un voltmètre digital sous forme de nombre compris entre 000 et 090, lequel correspond de manière unique à une position angulaire entre 0 et 90°. On peut enregistrer la sortie numérique sur carte, sur bande perforée ou sur bande magnétique en vue de son utilisation ultérieure comme entrée sur un calculateur.

Pour différentes raisons, ce genre d'enregistrement automatique de données est particulièrement séduisant. Il élimine toute intervention humaine entre la source de données et le calculateur. Il augmente la probabilité d'utilisation des données correctes pour toute l'expérience et d'élimination des erreurs de transcription. Il permet les mesures simultanées de plusieurs grandeurs et, par affichage des résultats sur des appareils indicateurs ou enregistreurs, le contrôle préliminaire de la validité des données, ceci avant toute analyse formelle plus longue. Enfin, il est susceptible d'accélérer tout le processus de traitement et d'analyse des données.

Un système mobile a été construit pour recevoir simultanément des impulsions électriques venant de huit sources primaires de données, convertir cette information sous forme digitale et enregistrer les données sur bande perforée. L'installation de connexions externes supplémentaires permettrait de commander plus de huit canaux, si besoin était.

La lecture est contrôlée par un chronomètre à un rythme pouvant varier entre : un test sur les huit canaux toutes les seize secondes et un test analogue par heure.

Outre la perforation sur bande de papier les données sont enregistrées sous forme de séries de points sur un graphique, chaque canal correspondant à une série identifiée par un chiffre et une couleur.

Les détails du codage, de la conversion et de la traduction sont extrêmement intéressants. La précision des résultats dépend du type de codeur et de système traducteur. Pour

illustrer la traduction d'une mesure codée typique, supposons que nous lisions 5,13 mv sur l'échelle de l'enregistreur à points multiples. Ceci correspond à la position 513 du commutateur du codeur, en notation normale, ou à la position 586 en notation décimale cyclique. Pour passer du système arabe au système cyclique il suffit de prendre le complément à 9 (9 — le chiffre) pour les chiffres qui suivent un chiffre impair. La traduction comprend le décodage de la représentation binaire du nombre décimal cyclique 586 et la conversion de ce nombre à l'échelle de l'enregistreur. Cette méthode permet d'éliminer les erreurs car changer un seul chiffre sur le codeur équivaut à changer un seul contact. La perforation de bande utilise deux codes à cinq canaux distincts pour enregistrer le numéro du point et la donnée.

Il y a deux stades entre la donnée contenue sur bande perforée et la donnée analysée sur calculateur. Premièrement, au cours d'un transfert des données de bande perforée sur bande magnétique, des blocs de séparation sont introduits et les données sont regroupées conformément au plan de calcul préparé. Au cours d'un deuxième stade on ramène en unités d'origine, telles que psi ou degré centigrade, les données comprises entre 000 et 999, on recherche les erreurs de perforation ou autres et on convertit les résultats en une forme décimale acceptable par le calculateur.

La simplicité avec laquelle les sorties de l'enregistreur peuvent être utilisées comme entrées sur le calculateur montre l'intérêt du système combiné.

Programme général d'analyse statistique

Avec de tels moyens d'enregistrement automatique des données la nécessité de programmes d'analyse statistique s'impose. On voit aussi combien il est souhaitable d'effectuer un examen préliminaire quelconque des données.

Une bibliothèque de programmes statistiques pour calculateur doit comporter d'abord, par conséquent, un programme d'analyse générale ou un jeu de programmes calculant pour toute la série de données tout ou partie des quantités suivantes : moyenne, variance, écart type, et coefficient de variation; la carré moyenne des écarts successive; un décompte des lectures au-dessus ou au-dessous de la moyenne avec test de significance; la régression des nombres par rapport à leur ordre chronologique et un test de non-normalité. Ce programme général pourrait aussi comprendre un sous-programme de tracé de courbe pour les distributions de fréquences avec des largeurs choisies. Il pourrait également permettre l'arrangement des données en sous-groupes et le calcul des moyennes et écarts dans les sous-groupes, la moyenne mobile et l'écart ($\underline{n} = 2$) mobile; la moyenne totale et l'écart à la moyenne; et les limites de contrôle pour $\bar{X}$ et R. Ceci serait surtout utile si le programme contenait un sous-programme de tracé de graphique de contrôle.

Toutes ces possibilités seraient à la disposition du statisticien qui ne les utiliserait pas forcément toutes dans tous les cas. Le calcul principal à faire sur la machine serait déterminé d'après les analyses préliminaires ci-dessus. Il existe maintenant pour la plupart des calculateurs un certain nombre de programmes de statistiques.

Programmes de statistiques types pour les calculateurs

Une bibliothèque-type de programmes statistiques comprend, outre les possibilités envisagées ci-dessus pour l'analyse en général, les programmes de détail suivants :

1. *Régression simple :* Calcul de pente, d'intercept; moyenne des y et moyenne des x; erreur type sur la pente; et coefficient de régression.

2. *Régression multiple :* (Limitations générales : 20 variables y compris la variable dépendante; pas plus de 2 000 observations sur chaque variable) calcule les pentes, linéaires ou curvilignes; les carrés moyennes résiduelles; le coefficient F de réduction en sommes de carrés; la matrice-moment; la matrice variance-covariance; et les coefficients de régression multiple.

3. *Compilateur d'algèbre matricielle :* Permet les calculs classiques d'algèbre matricielle (addition, multiplication, inversion, etc.); calcule les coefficients de régression et les tests de significance.

4. *Analyse de variance :* Analyse factorielle à six facteurs; calcule la table type de l'analyse de variance, c'est-à-dire les sommes des carrés, les degrés de liberté, les carrés moyennes et les coefficients F.

5. *Analyse de variance :* Analyse hiérarchique — analyse jusqu'à dix facteurs avec « n » boucles par facteur.

6. *Analyse de variance :* Modèle mixte — analyse un modèle mixte à cinq facteurs dont deux sont croisés avec jusqu'à trois ou six niveaux respectivement, et dont trois peuvent avoir « n » niveaux, plus « n » boucles.

7. *Tracé de la surface de réponse :* Calcule l'équation et trace les courbes pour deux ou trois variables.

8. *Analyse canonique :* Calcule les valeurs propres et les vecteurs propres d'une matrice réelle symétrique.

9. *Test T^2 de Hotelling :* Calcule T^2 et les moyennes et variances à utiliser pour les tests individuels « t » si T^2 n'est pas significatif.

10. *Intervalles de confiance simultanés :* Calcule l'intervalle de confiance de deux coefficients.

Conclusion

Des considérations multiples et variées doivent influencer nos réflexions en matière d'applications des statistiques lorsque le recueil des données est automatique et qu'un calculateur électronique est à notre disposition. En premier lieu, un planning doit être établi bien avant qu'une expérience soit montée ou qu'une production soit prête à commencer. Cela tient en partie à ce que la conversion de résultats de mesures en signaux électroniques exige un grand nombre d'intermédiaires et que l'installation de ceux-ci peut demander un temps considérable.

Pour éviter de nous trouver dans la situation de quelqu'un qui reçoit plus de données qu'il n'en peut traiter, même avec un calculateur, nous devons réduire le rythme d'échantillonnage au minimum absolu imposé par les conditions particulières de l'expérience. Même lorsqu'il a été bien programmé un enregistreur automatique de données peut produire beaucoup plus de résultats qu'une armée d'ingénieurs munis de carnets et crayons. Ce sera toujours un problème de réduire le rythme d'échantillonnage à la quantité adéquate et économique dans notre cas particulier. Des études et des analyses complémentaires sont nécessaires pour établir des règles permettant de décider *a priori* des rythmes appropriés aux différents systèmes industriels.

L'un des aspects de ce recueil de la quantité correcte de données est lié au genre d'analyse désiré. Donc, notre troisième et dernière considération aura pour but de souligner la nécessité de savoir d'avance avec précision ce qu'on fera des données. C'est une erreur de recueillir

des données simplement parce que l'on en a la possibilité si on ne doit pas les utiliser immédiatement, ou si l'on n'a aucune idée de ce que l'on pourra en faire plus tard.

Il n'est pas exagéré de dire que les développements du traitement des données sont de la plus haute importance pour le statisticien. Ils ouvrent des perspectives d'utilisations qu'on commence seulement à comprendre mais dont l'efficacité dernière est déjà hors de question.

Charles A. BICKING

Membre titulaire de la Société de Statistique de Paris
The Carborundum Company, Niagara Falls, New York

et Simone LEMARIÉ

Centre national d'études spatiales