

STATISTIQUE ET ANALYSE DES DONNÉES

NORBERT VICTOR

Statistique informatique science ou outil ?

Statistique et analyse des données, tome 9, n° 1 (1984), p. 76-101

http://www.numdam.org/item?id=SAD_1984__9_1_76_0

© Association pour la statistique et ses utilisations, 1984, tous droits réservés.

L'accès aux archives de la revue « Statistique et analyse des données » implique l'accord avec les conditions générales d'utilisation (<http://www.numdam.org/conditions>). Toute utilisation commerciale ou impression systématique est constitutive d'une infraction pénale. Toute copie ou impression de ce fichier doit contenir la présente mention de copyright.

NUMDAM

Article numérisé dans le cadre du programme
Numérisation de documents anciens mathématiques

<http://www.numdam.org/>

STATISTIQUE INFORMATIQUE

SCIENCE OU OUTIL ?

Norbert VICTOR

Institut für Medizinische
Dokumentation
Universität Heidelberg
Im Neuenheimer Feld 325
D-6900 HEIDELBERG (R.F.A)

Résumé : *La Statistique Informatique est définie comme le domaine de la statistique dont les objectifs sont de contribuer par l'informatique à la résolution totale ou partielle des problèmes posés aux statisticiens.*

Un rappel historique montre que ce domaine est plus vieux que le matériel électronique de traitement de l'information lui-même. Les différentes organisations qui ont contribué et qui contribuent au développement de la Statistique Informatique sont rappelées ainsi que les revues qu'elles ont diffusées.

Les différents domaines de la statistique qui ont été stimulés par Statistique Informatique (Méthodes Multidimensionnelles, Recherche de Tests exacts, Méthodes non-paramétriques, construction de tables statistiques, Simulations, Analyse Exploratoire) sont ensuite explorés et l'aptitude de Statistique Informatique à donner une vue globalisante des tâches du statisticien est soulignée.

Enfin les perspectives d'avenir sont tracées : Amélioration de l'accès aux logiciels statistiques et de leurs qualités, intégration des tâches du statisticien, recherche de solutions pour des problèmes non encore résolus, systèmes experts.

Ce texte voudrait montrer l'importance actuelle et future de la Statistique Informatique (STAT.INFO.) pour le développement de la statistique dans son ensemble.

Abstract : *The principal aim of Computational Statistics is to provide computerized methods to aid in the resolution -partial or complete- of statistical computation.*

Historically speaking this area already existed prior to the electronic age. The different organizations which have contributed to the development of Computational Statistics are presented as well as the journals used for their diffusion.

The main areas of Statistics stimulated by Computational Statistics (multidimensional methods, search of exact tests, non-parametric methods for construction of statistical tables, simulation, exploratory analysis) are studied as well as the global contribution as applied to statistical work in general.

Finally we outline the main areas of future development : improvement of Access Methods and quality of statistical software, integration of statisticians needs, research in the area of unresolved problems and expert systems.

Above all this text clearly underlines the importances of Computational Statistics as a driving force in the development of Statistical research in general.

i - INTRODUCTION

J'envisage avec mon exposé de donner une vue d'ensemble sur le domaine particulier de la statistique connu sous l'appellation anglaise de "Computational Statistics" et qui est une branche en plein essor ces dernières années. Nous utilisons dans cet article la dénomination "STATISTIQUE INFORMATIQUE", en abrégé "STAT.INFO". Le but principal est de montrer l'importance et l'utilité de ce domaine dans le cadre

de la statistique en général. C'est dans ce sens que je vais m'appliquer -après un bref aperçu historique- à donner des exemples concrets mettant en exergue des champs d'action de "STAT.INFO" qui ont eu ou promettent d'avoir du succès. J'étofferai la liste des exemples par une esquisse des perspectives d'avenir dans ce domaine.

Permettez-moi de donner une définition de "STAT.INFO" qui me soit propre. En effet, la compréhension linguistique de "Computational Statistics" ne me paraît pas claire. Ce manque de clarté est d'ailleurs mis en lumière par la multiplicité des appellations comme par exemple "Statistical Computing" ou "Statistical Computation" dans la terminologie anglophone, "Statistique Informatisée" ou "Calcul Statistique sur Ordinateur" dans la terminologie française.

La définition que je vais donner nous aidera, je l'espère, à répondre de manière précise à la question de savoir si "STAT.INFO" constitue une science autonome ou pas.

Je dirais d'abord que par "Statistique" je n'entends pas seulement la conception axiomatique de la statistique mathématique, mais aussi l'outil de la recherche empirique qui contribue, par ses méthodes d'investigation et d'évaluation objective des observations et des données d'expérimentation, à élargir considérablement nos connaissances. Sa démarche commence avec le développement de modèles, englobe le dépouillement et finit avec l'interprétation adéquate des résultats en relation avec les problèmes posés, en un mot l'analyse exploratoire ainsi que l'analyse confirmatoire des données.

II - DEFINITION

Dans ce cadre, je dirais alors que "STAT.INFO" est le domaine de la statistique dont les objectifs sont :

- soit d'apporter des solutions aux problèmes posés aux statisticiens,
 - soit de faciliter la tâche aux statisticiens
- par l'utilisation de l'ordinateur et/ou de l'informatique.

Cette définition me permettra par la suite d'utiliser la notion de "STAT.INFO" sans pour autant faire référence à son contenu anglais qui n'est pas aussi précis qu'on le croit d'habitude; elle me permettra en plus de répondre par un NON catégorique à la question de savoir si "STAT.INFO" est une science autonome au sens de la philosophie des Sciences.

En effet une science se caractérise par des méthodes d'approche qui lui sont spécifiques et/ou par un domaine de recherche propre. Pour "STAT.INFO." le champ d'action est celui de la statistique et les méthodes supplémentaires viennent de l'Informatique. Ainsi, manquent à "STAT.INFO." les deux traits fondamentaux qui caractérisent une science en tant que telle. Cette constatation est selon moi importante et pourra éviter, si l'on en tient suffisamment compte, de tomber dans le piège du développement de nouvelles disciplines universitaires baptisées pour l'occasion de disciplines scientifiques mais qui en réalité ne sont que de pseudo-sciences. Ceci ne nie pas l'importance de "STAT.INFO.". Au contraire, j'espère pouvoir montrer par la suite, l'importance de ce domaine pour la statistique et la manière dont il en a stimulé le développement ces dernières années. J'espère aussi pouvoir montrer qu'aujourd'hui, "STAT.INFO." est devenu un outil indispensable pour le statisticien praticien.

Le fait d'avoir dénié à "STAT.INFO." le caractère d'une science autonome ne diminue en rien la valeur des efforts scientifiques fournis dans ce domaine. Bien au contraire. Les travaux accomplis dans ce domaine sont d'un niveau scientifique élevé. C'est d'ailleurs un domaine qui demande au chercheur de posséder une large vue d'ensemble ainsi que des connaissances approfondies dans plusieurs disciplines complexes.

La statistique en tant que telle est un outil au service d'autres sciences. Que "STAT.INFO." ait aussi ce caractère ne doit pas à mon avis déranger.

En résumé disons que "STAT.INFO." n'est pas une science autonome, mais plutôt une partie de la statistique et un outil indispensable entre les mains du statisticien.

III - SURVOL HISTORIQUE

"STAT.INFO." est un domaine plus vieux que le matériel électronique de traitement de l'information. Il est aussi vieux que la statistique elle-même, car sans calcul, une analyse des données n'est pas possible. L'immensité des calculs pour les besoins de la statistique a poussé très tôt le statisticien à simplifier son travail au moyen de méthodes efficaces de calcul.

Le calculateur humain utilisant sa mémoire, des techniques de calcul avancées et (éventuellement) une machine à calculer mécanique, constituait autrefois ce que nous appelons aujourd'hui ordinateur.

Les fondateurs de la statistique moderne comme K. PEARSON, FISHER, EDGEWORTH, YATES, TIPPET ainsi que bien d'autres, avaient fait preuve d'une très grande habileté de calcul et devaient perdre une partie importante de leur temps à calculer et à chercher des méthodes de calcul simplificatrices.

Beaucoup de statisticiens ont passé toute leur vie à élaborer des tables statistiques et nous devons constater, que ces tables ont permis d'élargir pour la première fois l'utilisation de tests et ont permis à la statistique ses premières avancées.

Tous ces faits sont connus de tous, même s'il y a bien souvent la tendance à sous-estimer cet effort immense accompli avant l'ère de l'ordinateur. Il est notamment bien moins connu que déjà aux époques lointaines, "STAT.INFO." comme domaine propre dans la statistique, avait aussi son organisation, ses institutions ainsi que ses organes de vulgarisation.

Le but avoué de la recherche scientifique était d'alléger la tâche du statisticien en lui fournissant des méthodes de calculs et d'investigation (par exemple : tables statistiques, nomogrammes, utilisation adroite des méthodes d'approximation et d'analyse numérique). Selon ma définition ce but est resté celui de "STAT.INFO.". Il est aujourd'hui courant de se reporter aux développements en série et autres découvertes géniales en analyse numérique faites aux époques antérieures.

Laissez-moi vous citer comme témoignage seulement deux institutions anciennes ayant rigoureusement contribué au développement de "STAT.INFO".

La première de ces institutions est la collection "TRACTS FOR COMPUTERS" éditée par le Department of Statistics, University College, London (publisher : Cambridge University Press) avec une publication totale de 30 volumes entre 1919 et 1973. Des statisticiens connus comme K et E.S. PEARSON, H.E. SOPER, J. BROWNLEE, J.O. IRWIN, L.H.C. TIPPETT et M.C. KENDALL étaient parmi les auteurs de cette publication.

La deuxième de ces institutions est le "COMMITTEE ON APPLIED MATHEMATICAL STATISTICS" (ce nom a subi quelques variations au cours du temps), allant de 1936 à 1962, ayant son siège aux USA et dont la présidence fut assurée par A.A. BENNETT, R.C. ARCHIBALD, C. EISENHART et S.S. WILKS. On peut citer entre autres membres T.W. ANDERSON, D. BLACKWELL, R.A. BRADLEY, W.G. COCHRAN, W. FELLER, H.O. HARTLEY, P.C. HOEL, I.R. SAVAGE et J. WOLFOWITZ.

Ce comité a entretenu une publication dénommée "Mathematical Tables and Aids to Computation" (en abrégé M.T.A.C.), rebaptisée plus tard du nom de "Mathematics of Computation", entre 1943 et 1962. Une oeuvre supplémentaire de ce comité portait le nom de "Guide to Tables in Mathematical Statistics" (GREENWOOD and HARTLEY, 1962) dont l'avant propos informe en détail sur la naissance et le développement du comité. L'introduction de l'ordinateur dans le travail statistique a quelque peu rendu superflu l'utilisation de beaucoup de tables, sans pour autant les supprimer complètement ; cet index reste donc valable.

Les revues citées ci-dessus sont aujourd'hui encore exploitées par beaucoup de Statisticiens-Informaticiens qui bien souvent se limitent tout simplement à traduire les méthodes décrites en programmes.

Les dates les plus importantes dans le développement de STAT.INFO. sont schématisées par la figure 1. Elles se caractérisent par deux faits :

1. L'axe de ce développement est intimement lié au développement et à la diffusion de l'ordinateur. Corrélativement au développement des techniques du traitement de l'information se résorbe lentement une bonne partie des activités d'autrefois.
2. Une période du développement supplémentaire de STAT.INFO a connu ses débuts au moment de l'inclusion des méthodes informatiques dans le travail ainsi que de la collaboration fructueuse entre statisticiens et informaticiens.

Si l'on exclut l'ère précédant l'ordinateur, les praticiens étaient plutôt des avant-coureurs qui, de manière pragmatique, créaient des outils importants pour l'analyse statistique et ce sous forme de programmes-produits (anglais : program packages). Cette activité, à savoir la mise à la disposition de méthodes statistiques sous forme de logiciel d'une manière souple et maniable constitue aujourd'hui encore l'axe principal de "STAT.INFO.". C'est à cet axe que nous dédions la colonne gauche de la figure 1.

BMD avait fait ses débuts en 1961 comme étant le premier logiciel à être diffusé. Depuis cette époque, le nombre de logiciels et de systèmes a augmenté considérablement et augmentera à l'avenir. Le faisceau de méthodes dans chaque logiciel est en train d'augmenter. Les systèmes sont adaptés aux différents types d'ordinateurs et des types spéciaux de systèmes sont créés pour de nombreux domaines du processus global de l'analyse des données. Aux "general purpose packages", dont des exemples significatifs sont donnés dans les lignes du milieu, viennent s'ajouter d'autres systèmes pour des tâches spéciales comme acquisition et manipulation des données (par exemple CENTS), pour la conservation et la manipulation des données (par exemple SIR), pour la production de tableaux (par exemple TPL), pour la représentation graphique des résultats (par exemple SAS-GRAPH) ainsi que des systèmes interactifs (par exemple CS et SCSS).

Il y a un très grand afflux de ces produits sur le marché au point qu'il est très difficile de faire un choix. Dans son deuxième rapport FRANCIS (1981) étudie 127 logiciels. Actuellement, la tendance au développement des logiciels et à leur implémentation sur les micros, a donné une dimension nouvelle à STAT.INFO.

La colonne du milieu de la figure 1 est dédiée aux congrès spécialisés dans "STAT.INFO." de même qu'aux fondations des associations scientifiques s'occupant de ce domaine. Signalons que c'est en Amérique du Nord qu'a commencé aussi ce genre d'activités. A ma connaissance, c'est le "Symposium on the Interface of Computer Science and Statistics" qui eut lieu pour la première fois en 1967 qui jeta les bases de ce genre d'activités. Une année après, l'"American Statistical Association" (ASA) consacrait une session à "Statistical Computation" lors de sa réunion annuelle. En 1969, l'ASA créa un comité particulier qui devait s'occuper uniquement de ce domaine. Ce comité se transforma en 1972 en section autonome et, depuis, il compose un courant autonome du congrès de l'ASA. Les activités de ce comité se sont élargies par la fondation de nombreux sous-comités, par exemple le comité de l'évaluation des logiciels statistiques.

En Europe, la Société COMPSTAT a été fondée en 1973. Elle tien son premier congrès en 1974 à VIENNE. Depuis lors ce congrès se tient tous les deux ans et élargit considérablement son audience et son champ d'action.

En 1977, l'"International Association for Statistical Computing" filiale de l'Institut International de la Statistique, était fondée, en tant qu'organisation mondiale. En 1980, COMPSTAT était reçue en son sein en tant que section européenne.

On pourrait établir une liste encore plus longue d'organisations nationales qui s'occupent actuellement de STAT.INFO. Permettez-moi de citer en exemple le cas du groupe de travail pour les logiciels statistiques de la Société Allemande pour l'Informatique et la Statistique Médicale. Ce groupe de travail a tenu, à partir de 1974, de nombreux "workshops" ainsi que des congrès qui ont eu beaucoup de succès et réuni beaucoup de participants et qui continuent à le faire. J'aimerais aussi citer la série des congrès tenus par l'I.N.R.I.A. sur l'Analyse des Données et l'Informatique respectivement en 1977, 79, 83.

La colonne de droite de la figure 1 comprend les activités de STAT.INFO. dans le domaine de la publication.

On remarquera que les efforts d'autrefois ont connu, sans difficulté, une transition dans l'ère de l'ordinateur.

Sur le côté droit, des dates sont mentionnées, dates au cours desquelles des revues statistiques influentes ont soit élargi leurs thèmes, ou alors créé des rubriques propres pour STAT.INFO. Un cas spécial est la rubrique des algorithmes du "Journal on Applied Statistics" (JRSS (C)), qui s'était limité à cet aspect partiel de STAT.INFO. Nous l'avons cité ici car c'était la première activité de ce genre ; elle continue à avoir du succès aujourd'hui.

Le côté gauche de cette colonne montre quelques exemples d'éditions de rapports de congrès, d'autres rapports importants ainsi que des revues dédiées spécialement à STAT.INFO.

A mon avis, les Proceedings de la conférence de "Statistical Computation" de l'Université de Wisconsin (MILTON and NELDER, 1969) constituent ici une pierre angulaire. L'année suivante débuta l'édition des "Proceedings" du "Symposia on the Interface" (cf. MANN, 1978). En 1972 débuta l'édition des Proceedings de la Section "Statistical Computation" de l'ASA. En 1974 était publié le rapport d'évaluation de l'ASA (cf. FRANCIS, HEIBERGER et VELLEMAN, 1975). Ce rapport impulsa de nombreuses activités pour l'évaluation des logiciels statistiques et fut à l'origine des "Comparative Reviews" de l'IASC et FRANCIS fréquemment utilisées (cf. FRANCIS, 1977, 1981).

J'aimerais citer comme revue spécialisées uniquement dans "STAT.INFO." le "Journal of Statistical Computation and Simulation" (première parution: 1972). Permettez-moi de citer aussi le "Statistical Software Newsletter (SSN), une revue spécialisée créée par moi-même et qui s'est donné comme objectif la standardisation et l'évaluation des logiciels statistiques. Cette revue est depuis 1975 publiée par la CSF de Munich et est maintenant devenue l'organe officiel de l'IASC. Il est probable que d'autres organes de vulgarisation ou d'autres publications de même genre existent sans que

J'en sois informé. Entre 1980, 1983 et 1984 sont apparues trois revues éditées par des maisons d'édition commerciales qui semblent avoir le même objectif que "STAT.INFO.". La revue "Journal of Scientific Statistical Computing" (SIAM) n'est pas, malgré son nom plus adaptée à "STAT.INFO." que d'autres revues informatiques. Il est encore trop tôt pour apprécier la valeur des deux autres revues, en particulier "Computational Statistics Quaterly" dans la mesure où aucun numéro n'a été diffusé jusqu'ici.

Ce coup d'oeil rétrospectif nous a permis de constater que STAT.INFO. est plus vieux que l'ordinateur. Cependant, il doit l'avancée significative de son développement au développement prodigieux des techniques informatiques. Les développements de STAT.INFO. sont bien souvent dus à des collègues travaillant dans des domaines marginaux de la statistique. Les sociétés établies de la statistique mettent bien souvent du retard à accepter ces développements.

IV - DEVELOPPEMENTS EN STATISTIQUE STIMULES PAR STAT.INFO.

Le volume des activités indiquées en figure 1 est un témoignage de la vitalité de STAT.INFO. ainsi que de l'importance de l'emploi de l'ordinateur pour la statistique.

STAT.INFO. n'est pas seulement un outil entre les mains du statisticien, outil qui lui permet de faire des calculs précis. Cette branche a grandement contribué -elle contribue encore- à la diffusion des méthodes statistiques dans presque toutes les sciences utilisant des méthodes d'approche empiriques. Cette liaison fertilisante entre les méthodes de l'informatique et celles de la statistique, a contribué à l'amélioration et au perfectionnement des méthodes statistiques. Elle a eu un effet bénéfique sur le développement méthodologique.

STAT.INFO. c'est beaucoup plus que des calculs statistiques. Le statisticien-informaticien n'est pas l'esclave calculateur au service du statisticien. C'est pourquoi je trouve l'expression "Statistical Computing" que je traduirais par "Calcul Statistique sur l'Ordinateur", quelque peu malheureuse. Pour justifier ma remarque, il me semble urgent d'établir une liste -incomplète- des domaines statistiques influencés positivement par STAT.INFO.

a) Méthodes multidimensionnelles

La condition nécessaire à l'application étendue des méthodes d'analyse multidimensionnelle, est l'utilisation d'un ordinateur de même que l'existence d'un logiciel. Ceci est valable pour les modèles linéaires et non linéaires, aussi bien que pour les variables quantitatives et les variables qualitatives. Sans aucun doute, l'application permanente des méthodes multidimensionnelles et l'acquisition d'expérience qui en découle, ont énormément favorisé le perfectionnement de ces méthodes.

C'est bien connu -nous l'avions signalé plus haut- que les statisticiens d'avant l'ère de l'ordinateur avaient eu d'étonnantes performances dans leurs calculs. Cependant, le volume de ces calculs restait très modeste comparé à la masse des analyses multidimensionnelles exécutées maintenant : on est aujourd'hui en mesure d'inverser de grandes matrices en quelques fractions de secondes, ce qui prenait autrefois une journée entière à un calculateur expérimenté.

Non seulement les statisticiens se voient débarrassés d'un temps de calcul stupide -ce qui leur permet de s'occuper des problèmes proprement dits de la statistique- mais aussi et surtout la réduction de la marge des erreurs commises est un élément important à prendre en compte.

Dans sa conférence célèbre, tenue en 1948 devant la "Royal Statistical Society" sur les problèmes de classification, RAO donna des fonctions de discrimination pour la séparation de trois castes indiennes à l'aide de quatre variables aléatoires et parla du temps énorme nécessaire aux calculs, un point central dans les discussions qui suivirent son rapport. ANDERSON (1951) utilisa dans son travail de base sur les fonctions discriminantes l'exemple de RAO et corrigea à cette occasion toute une série d'erreurs de calculs commises par RAO. Dans son oeuvre de base sur l'analyse multidimensionnelle parue sept ans plus tard, ANDERSON (1958) reprit à nouveau l'exemple de RAO et corrigea alors les erreurs commises dans son travail de 1951. Pourtant ces calculs avaient été exécutés par des praticiens expérimentés à l'aide cette fois de calculatrices mécaniques hautement perfectionnées.

b) Tests exacts

Le progrès de STAT.INFO. a sans aucun doute comme influence la tendance à l'usage plus fréquent de tests exacts. Les performances extraordinaires des ordinateurs de même que le développement d'algorithmes nouveaux rapides et efficaces, en particulier pour les problèmes d'analyse combinatoire, nous permettent aujourd'hui de conduire des tests de manière exacte là où autrefois on procédait par approximation ou par méthodes asymptotiques. Par ailleurs, la théorie de la complexité -une branche relativement jeune de l'informatique- nous permet de faire des pronostics sur la calculabilité des problèmes statistiques (Problèmes de classe P, de classe NP, NP complet, NP difficile...), sur le volume des calculs, les limites de la performance, ainsi que sur les garanties de la performance des algorithmes concernés. Ainsi peut-on décider à quel moment la recherche d'algorithmes plus efficaces ou bien de solutions exactes, a un sens, de même qu'on peut apprécier l'ampleur de la perte résultant du remplacement des algorithmes éprouvés par des algorithmes d'approximation rapide. On trouve l'importance de cette branche de l'informatique pour la statistique dans le travail panoramique de BENTLEY (1978).

c) Méthodes non paramétriques, méthodes robustes et autres

STAT.INFO. a aussi été à l'origine de la diffusion des méthodes statistiques non paramétriques et des méthodes robustes. Elle impulsa par la même occasion le développement théorique correspondant. On peut citer comme exemple l'estimation des courbes et surfaces non paramétriques (développements en série orthogonaux, méthodes du noyau, spline-approximation, etc.) qui serait absolument impensable -surtout dans le cas multidimensionnel- sans l'utilisation de l'ordinateur. Entre autres exemples on peut citer les techniques de 'Jackknife', les méthodes de 'Bootstrap' ainsi que l'analyse de sensibilité qui contribuent à la diminution des marges d'erreurs des fonctions d'estimation, de l'influence de données entachées d'erreurs et à l'identification des données aberrantes et qui, pour cela, occupe une place de choix dans la pratique du statisticien. Il était impossible de

mener ce genre d'activités autrefois, car en effet, seule l'utilisation de l'ordinateur permet aujourd'hui de renouveler n-fois une analyse statistique afin de déterminer l'influence de chaque valeur de l'échantillon sur le résultat de l'analyse. Les statisticiens-informaticiens ne se sont pas contentés de mener une analyse n-fois. Ils sont allés plus loin en cherchant des algorithmes adéquats qui permettent de réduire de manière sensible la durée des calculs nécessaires. Nous pouvons mentionner comme exemple typique les propositions de FUKINAGA et de KESSEL (1971) au sujet de l'analyse discriminatoire "One-Hold-Out".

En règle générale, le développement d'algorithmes numériques efficaces et stables est à mettre aussi au compte des efforts de STAT.INFO. L'article de CHAN, GOLUB et LE VEQUE (1983) sur les calculs de variance des échantillons met en lumière le gain dans l'efficacité et dans la précision qu'on peut en attendre même pour des procédures standards. Citons aussi la masse d'algorithmes nouveaux pour les estimations qui permettent aujourd'hui une utilisation adéquate de toute la palette des procédés existants, par exemple : les estimateurs de REML, MINQUE, I-MINQUE et MIVQUE des composantes de la variance et l'algorithme de Deming-Stephan pour des modèles log-linéaires des tableaux de contingence.

d) Tables statistiques

. L'élaboration de tables statistiques occupait autrefois, bien souvent pour plusieurs années, des groupes de statisticiens et des calculateurs. Ces tables peuvent de nos jours être élaborées en peu de temps à l'aide de l'ordinateur qui -c'est important- en assure immédiatement l'impression sans erreur. Ce n'est pas seulement la production de ces tables qui s'en trouve par là facilitée. Il y a aussi le fait que STAT.INFO. a accentué la tendance à une meilleure utilisation de ces tables. Bien des éléments des tables restent inefficaces, car malgré le grand volume de calculs qu'ils ont nécessité, la fréquence d'accès à ces éléments est bien minime. C'est pourquoi les tables sont de plus en plus remplacées par des algorithmes d'accès facile qui, en cas de besoin, peuvent fournir la valeur

voulue avec la précision souhaitée. Cette conception évite au statisticien les transformations difficiles et bien souvent entachées d'erreurs des valeurs des tables au moment des applications pratiques car il ne lui est plus imposé de se limiter à un petit nombre de tables de base.

Il n'y a pas que la capacité de calcul de l'ordinateur qui a une place de choix dans les progrès accomplis, mais aussi le développement continu de l'analyse numérique, en particulier l'intégration numérique dans les domaines multidimensionnels.

Même s'il est facile de coucher sur le papier une intégrale de la distribution normale dans le cas multidimensionnel, des cas simples concernant des domaines peu complexes conduisent à des calculs difficiles et longs. Je me suis permis de faire cette remarque banale car j'ai bien l'impression que beaucoup de théoriciens, en livrant ce genre d'intégrales comme résultat de leurs recherches ne se rendent pas compte bien souvent des difficultés rencontrées par les praticiens chargés de transporter la théorie dans la pratique concrète.

e) Simulations

Des procédés de simulation -absolument impensables sans l'ordinateur- ont, dans les temps récents, rendu un service inestimable à beaucoup de domaines de la statistique. La signification de ces procédés est devenue si grande, qu'aujourd'hui les techniques de simulation constituent elle-même un domaine important de l'informatique et de la statistique. La simulation intervient toujours là où la théorie pure est incapable d'aider. C'est ainsi que beaucoup de problèmes analytiquement insolubles se sont trouvés de cette manière résolus de façon satisfaisante. De bonnes études de simulation permettent de faire des pronostics pour les limites à partir desquelles certains résultats asymptotiques sont utilisables pour des échantillonnages d'une grandeur limitée. Le nombre de ces simulations dans la statistique n'est, à l'heure actuelle, pas à déterminer. En effet on simule -on simulera encore- un peu trop inutilement.

Les mauvais exemples ne doivent cependant pas diminuer la valeur de tout le reste. Comme exemple concret d'une étude de simulations de grande valeur nous pouvons citer le cas de la "Princeton-Study" (ANDREWS et al., 1972) dans laquelle la robustesse de différents estimateurs pour les paramètres de localisation est étudiée de manière empirique.

De bons générateurs de nombres (pseudo) aléatoires sont la condition sine qua non pour tout modèle de simulation. Les aléas et dangers de la construction de générateurs de nombre aléatoires restent inconnus à beaucoup de statisticiens. Bien souvent, des constructeurs de petits calculateurs lancent sur le marché des générateurs entachés d'erreurs que les utilisateurs, en toute bonne foi, mettent en oeuvre (cf. RIPLEY, 1983).

Une coopération assidue entre statisticiens et informaticiens et l'intégration à leur travail d'une connaissance précise de l'arithmétique de l'ordinateur, sont nécessaires pour développer un instrument adéquat pour les simulations et son contrôle. Il suffit de jeter un coup d'oeil sur la littérature pour se rendre compte que cette problématique est depuis longtemps au centre des préoccupations de STAT.INFO. et le sera encore à l'avenir.

f) Analyse exploratoire des données (EDA)

Nous avons abordé jusqu'ici seulement des domaines de la "statistique traditionnelle". Ces dernières années cependant, le domaine de l'analyse exploratoire des données qui, dans sa conception, va au delà de l'architecture axiomatique de la statistique mathématique, a pris de l'importance au sein de la statistique et trouve de plus en plus de partisans parmi les statisticiens. Si nous prenons au sérieux notre travail de statisticiens, travail qui consiste à aider les sciences empiriques, à élargir leurs connaissances, alors nous ne devons pas nous limiter à vérifier la validité d'hypothèses préalablement établies. Nous devons parallèlement nous occuper à trouver des procédés adaptés à la découverte de phénomènes nouveaux, même si l'ensemble de nos méthodes traditionnelles ne suffit pas à faire des énoncés quantifiables sur la validité de nos nouvelles découvertes.

Le développement parallèle de STAT.INFO. et de EDA ne peut passer inaperçu et est facile à expliquer. Les méthodes exploratoires opèrent sur des données empiriques et multidimensionnelles et ont pour but de découvrir, dans ces données, des structures non triviales. On réalise progressivement cette tâche de la manière suivante :

1. Caractérisation de la distribution (empirique) des données.
2. Elimination des informations redondantes.
3. Recherche et description des interdépendances entre les attributs de même que la recherche des hétérogénéités de l'ensemble des objets par rapport à ces attributs -respectivement la recherche de sous groupes homogènes.

Les moyens employés pour cela peuvent être classifiés de la manière suivante :

1. Techniques descriptives, en particulier méthodes graphiques pour données multidimensionnelles. Ajustement de fonctions ou de structures (par exemple : calculs de régression et méthodes de la théorie des graphes).
2. Analyse des composantes principales et méthodes de positionnement multidimensionnel.
3. Analyses canoniques (dont un cas particulier est l'analyse de correspondance très utilisée en France) et classification automatique.

Tous ces procédés prennent un temps de calcul énorme et seraient, sans l'emploi de l'ordinateur, utilisables seulement avec de grandes restrictions.

BOCK (1980) dont l'article laisse poindre aussi l'importance de l'emploi de l'ordinateur, donne une bonne vue d'ensemble sur la panoplie des méthodes de EDA, sans pour autant se laisser enfermer dans une "école" déterminée. De même TUKEY, avec son appréciation singulière de l'EDA et son beguin pour les méthodes du "papier-crayon" n'a pu s'empêcher, à la fin de son oeuvre fondamentale sur l'EDA, d'écrire "You couldn't see the computer on the 661 pages that have gone before, but its shadow lay heavily on many of them" (TUKEY (1977), p. 664).

g) Une vue globalisante des tâches du statisticien

Il n'est pas nécessaire de parler de l'influence que les logiciels statistiques généraux, ont eu sur la diffusion et l'utilisation intenses des méthodes statistiques dans les sciences empiriques. J'aimerais considérer d'un peu plus près les tentatives reposant sur une vision globale des tâches que doit remplir le statisticien et qui ont débuté avec l'emploi et le développement des logiciels.

Les tâches du statisticien dans la recherche empirique comprennent : la planification des expérimentations, l'établissement des questionnaires, la collecte des données, leur saisie sur un support permettant leur introduction sur l'ordinateur, les traitements des données, le dépouillement des résultats, leur présentation et leur interprétation relative au problème. Se limiter uniquement à une des tâches ci-dessus (par exemple dépouillement) signifierait une restriction de l'importance du statisticien et un amoindrissement de sa responsabilité. La participation du statisticien à toutes ces étapes d'une étude empirique lui confère une responsabilité dans la garantie et la qualité d'un dépouillement.

L'expérience acquise, que seulement environ 10 % des activités d'un statisticien praticien sont consacrées au dépouillement pendant que d'autres tâches, telles que l'acquisition et le traitement des données, constituent la plus grande partie de ses activités, a provoqué avec le temps l'introduction de logiciels pour d'autres tâches que les traditionnelles tâches de dépouillement. De même qu'on s'est donné la peine au début de s'occuper des possibilités de liaison entre les différents systèmes de dépouillement dans le but d'élargir le faisceau des méthodes statistiques utilisables, de même prennent aujourd'hui de l'importance, les liaisons avec les banques de données, les systèmes de manipulation de données, et les systèmes de représentation graphique. Un pas supplémentaire réside dans le développement de systèmes spéciaux de traitement des données pour les structures de données courantes en statistique. En effet, les banques de données usuelles s'avèrent mal adaptées à la statistique.

On est sûrement encore loin de systèmes universels cohérents ou de cadres généraux pour systèmes partiels. La tendance à avoir une approche globale des tâches du statisticien commencent à porter ses fruits : on cherche une description formelle pour toutes les tâches dans le processus global d'une analyse de données et on commence à considérer globalement les structures de données pour la statistique en liaison avec les algorithmes d'application correspondants. (Systèmes de banques de données - banques de méthodes, cf. HAUX (1983/84).

V - PERSPECTIVES D'AVENIR

De même que dès le début il était impossible à la statistique d'exister sans méthodes de calcul, de même il est impensable d'envisager la statistique de l'avenir sans STAT.INFO. Cela veut dire que, non seulement tout l'outillage obtenu sera employé par les statisticiens, mais aussi que, dans ce domaine, il y aura continuellement un développement et une recherche intense. Je me contente ici de donner une liste de tâches, de défis et de perspectives d'avenir, liste appropriée pour compléter l'exposé antérieur de CHAMBERS (1980).

a) Amélioration de l'accès aux logiciels statistiques

Dans le futur, l'accès aux logiciels statistiques devra être amélioré dans deux directions :

1. Utilisation de micro-ordinateurs par transfert des logiciels existants ou création de logiciels spécifiques.
2. Amélioration du dialogue Homme-Machine.

Le développement de logiciels adéquats pour micros sera certainement une des principales activités de STAT.INFO. dans les années à venir. Le problème clé sera, à ce sujet, de maintenir le niveau qualitatif qu'offre la palette des produits destinés aux gros ordinateurs. Il est à prévoir que les problèmes de l'analyse des données seront résolus presque exclusivement par le "personal computing".

Dans le dialogue Homme-Machine, je note deux points qui demandent à être améliorés : la communication entre l'utilisateur et le système dans les deux directions (langages de commandes simples et adaptés aux problèmes ; réponses claires et suffisante du système) et une amélioration de la présentation des résultats surtout dans le domaine graphique.

b) Amélioration de la qualité du logiciel statistique

Une amélioration de la qualité est souhaitable en particulier dans trois directions :

- l'ampleur, l'actualité et l'exactitude des méthodes disponibles.
- l'évaluation des logiciels existants et
- une conception meilleure des programmes et des systèmes du point de vue informatique.

Dans le domaine méthodologique, je mentionnerai quelques élargissements souhaitables : l'estimation non paramétrique de densités, le dépouillement des temps de survie censurés, les procédés de tests multiples séquentiels dans le sens des propositions de MARCUS, PERRITZ, GABRIEL (1976), une considération renforcée des méthodes robustes et surtout de l'EDA. J'aimerais citer aussi l'adaptation de procédés pour ensembles incomplets de données avec des valeurs entachées d'erreurs et l'application de meilleurs algorithmes numériques.

L'évaluation du logiciel existant représente un grand défi pour le statisticien-informaticien. Là où il y a de la lumière existe aussi l'ombre et on doit malheureusement considérer une bonne partie du logiciel existant comme défectueuse. Protéger l'utilisateur, à l'aide d'une bonne information sur la qualité du logiciel mis à sa disposition doit être un des objectifs fondamentaux pour le futur. Des catalogues avec critères d'évaluation (FRANCIS, HEIBERGER et VELLEMAN, 1975), des listes d'exigences (HULTSCH et al, 1978), les catalogues d'évaluation déjà cités (FRANCIS, 1977, 1981) et quelques publications dispersées de comparaisons de fond fournissent déjà des éléments d'appréciations ; cela n'est qu'un

début. Il manque encore des méthodes générales pour mesurer la qualité de logiciels statistiques. L'avancée pénible des informaticiens dans le domaine du "software metrics research" montre, qu'il ne suffit pas de transposer uniquement les méthodes de l'informatique, mais qu'il est nécessaire que STAT.INFO impulse ses propres développements.

Nous devons beaucoup plus que par le passé tenir compte, au cours de la révision des systèmes existants et du développement de systèmes nouveaux, des développements nouveaux de l'informatique dans les domaines de la conception des systèmes, des techniques de programmation et de l'implémentation des systèmes. Au cours de ces développements nous devons tenir compte des exigences issues de l'expérience accumulée jusqu'ici dans la pratique. Un coup d'oeil sur le code des logiciels statistiques mis aujourd'hui à notre disposition montre souvent que ceux-ci ont été créés sans l'aide des méthodes de programmation modernes (par exemple structuragrammes) et sans l'aide des nouveaux moyens de conception de programmes (pré-processeur, générateurs-Parser, programmes prototypes avec convertisseurs etc..).

Permettez-moi d'énumérer quelques uns des mots clefs d'une telle démarche :

Prise en compte renforcée des structures des données concernées ; élargissement du nombre des types de données et introduction de types de données abstraites ; allègement de l'extensibilité des systèmes par la possibilité d'écriture de procédures spécifiques à partir des composants des systèmes disponibles ; amélioration de la portabilité ; amélioration de la documentation pour l'agent chargé du développement des méthodes ; aide à l'édition de "log-books" destinés à l'utilisateur ; meilleur contrôle et comptes rendus des erreurs et meilleur guide pour l'utilisateur (help-systems).

Une tâche principale de STAT.INFO. pour les années à venir sera la tentative de trouver des logiciels de qualité -selon les règles de

l'informatique-, adaptés aux problèmes posés. Adaptés aux problèmes posés veut dire ici : introduction facile et pratique des données dans le système (éventuellement à partir des banques de données), structures des données flexibles, faisceau des méthodes adéquat et possibilités simples de traitement des irrégularités dans les données (par exemple les données manquantes) ; parmi les critères de qualité des bons logiciels nous pouvons citer : une bonne structuration, la maintenance simple, la portabilité et une documentation complète.

c) Faciliter le développement de nouveaux logiciels

Permettez-moi de citer, à côté des possibilités d'extensions des systèmes par écriture de procédures spécifiques, les bibliothèques de sous-programmes pour des domaines spéciaux (par exemple : analyse numérique et graphique) ainsi que l'usage renforcé d'instruments logiques pour le développement de logiciels . L'utilisation de bibliothèques spéciales ne diminuent pas seulement le volume du travail, elle contribue aussi la plupart du temps (numérique !) à l'amélioration de la qualité des logiciels.

d) Algorithmes pour des problèmes non encore résolus jusqu'ici

Bien sûr la recherche continuera pour l'obtention d'algorithmes adaptés à l'obtention des solutions de problèmes non résolus à ce jour. Elle continuera de même pour l'obtention d'algorithmes d'approximation pour des problèmes dont la complexité ne permet pas d'espérer des solutions exactes dans des temps raisonnables. Bien sûr des algorithmes, connus aujourd'hui pour bons, seront améliorés à l'avenir. Il faut malheureusement constater que dans ce domaine qui nécessiterait normalement une coopération intense entre spécialistes de l'analyse combinatoire, de l'informatique et de la statistique, le courant d'informations entre scientifique des diverses disciplines est extrêmement lent.

e) Intégration des tâches du statisticien

Ce point, cité plus haut, ne peut nullement être considéré comme clos ; bien au contraire, car l'intégration de tout le processus de l'analyse des données dans un cadre cohérent reste un des principaux défis lancés à STAT.INFO. pour l'avenir. Le livre de WIRTH "Algorithms + Datastructures = Programs" (WIRTH, 1976), qui fait époque en informatique, peut être considéré ici, transposé sur le plan des systèmes statistiques globaux (banque de données + banque de méthodes) comme indiquant une direction ; ainsi serions-nous dans l'obligation d'adopter un langage pour la description formelle des tâches du statisticien durant tout le processus de l'analyse des données. Toutefois j'aimerais attirer l'attention sur les délais d'obtention d'un langage statistique unique, les progrès lents depuis les premières propositions de GOWER et al. (1967) nous incitant à plus de précaution.

f) Intégration des connaissances des experts dans l'analyse des données

Le plus grand défi lancé à STAT. INFO. pour l'avenir est cependant la création des concepts et des systèmes qui, au moment du dépouillement de l'ensemble des données, profiteraient des connaissances existant au dehors des expériences conçues (knowledge based software). Nous pouvons indiquer comme première tentative d'inclusion des connaissances heuristiques dans l'analyse des données, le développement des systèmes experts qui, aujourd'hui déjà, trouvent leur application dans le domaine de l'intelligence artificielle et qui existent comme prototypes pour les problèmes statistiques (cf. GALE et PRÉGIBON, 1982).

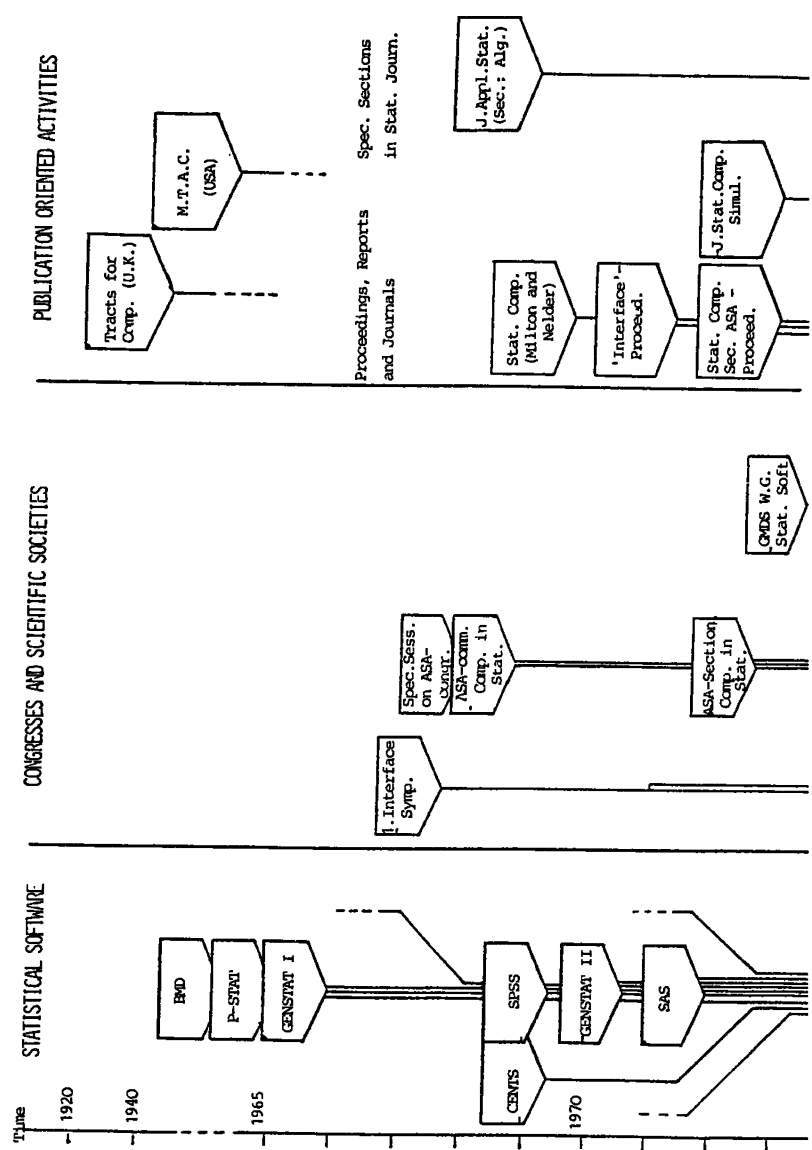
Ce défi lancé aux statisticiens avait été formulé de manière nette par ZELEN (1983) et CHAMBERS (1983) dont les propositions conduisirent à des discussions très riches aux congrès mondiaux de la Biométrie et de l'ISI. Les statisticiens sont invités à voir au delà du modèle choisi et au delà de l'ensemble des données au moment de prendre leur décision. Cette manière de procéder ressemble fort à l'introduction de l'information a priori dans l'analyse statistique.

VI - CONCLUSION

J'espère être parvenu à montrer l'importance actuelle et future de STAT.INFO. pour le développement de la statistique en général. J'espère aussi avoir montré combien est injuste le jugement rabaissant que portent certains collègues théoriciens sur cette branche de notre discipline. J'espère avant tout avoir montré que le statisticien praticien d'aujourd'hui ne pourrait pas, sans cet outil, se tirer d'affaire et c'est pourquoi STAT.INFO. doit prendre obligatoirement une plus grande importance dans la formation du statisticien.

REMERCIEMENTS

Mes remerciements vont d'une part à mon collègue, Y. ESCOUFIER pour avoir consacré un temps précieux pour son travail à la lecture critique de mon article, d'autre part à mes collaborateurs M. MECKESHEIMER et D. COLY pour s'être associés à la traduction de l'allemand en français de mon document.



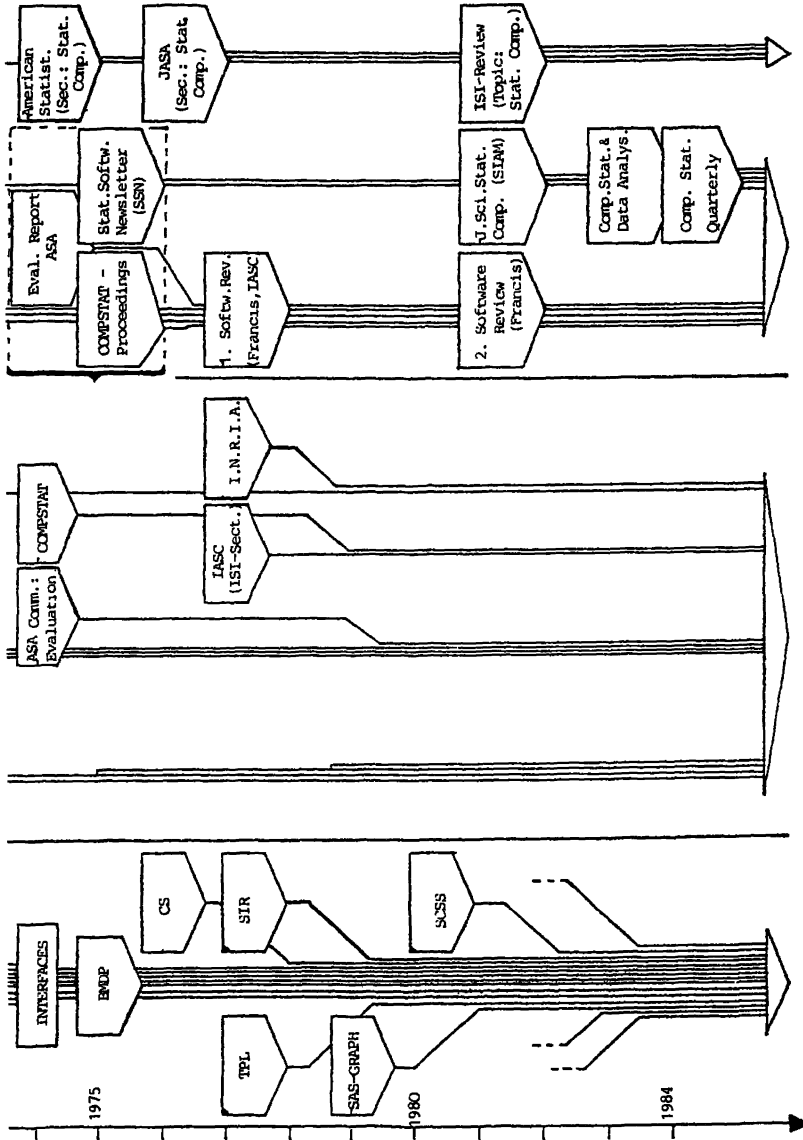


FIGURE 1: TIME TABLE OF ACTIVITIES IN COMPUTATIONAL STATISTICS

BIBLIOGRAPHIE

- ANDERSON, T.W. (1951) : Classification by multivariate analysis
Psychometrika 16, 31-50.
- ANDERSON, T.W. (1958) : An introduction to Multivariate Statistical
Analysis.
Wiley, New York pp. 374.
- ANDREWS, D.F., BICKEL, P.J. HAMPEL, F.R., HUBER, P.J., ROGERS, W.H.
and TUKEY, J.W. (1972) : Robust Estimates of Location :
Survey and Advances.
Princeton University Press pp. 373.
- BENTLEY, J.L. (1978) : The Whys and Wherefores of Algorithm Design.
In : Gallant, A.R. and GERIG, T.M. (eds.) : Proc. Comp. Sci.
and Statistics : 11th Ann. Symp. Interface. Inst. Statist.
N.C.-State Univ., Raleigh pp. 286-296.
- BOCK, H.H. (1980) : Explorative Datenanalysis.
In : Victor, N. LEHMACHER, W. and VAN EIMEREN, W. (eds) :
Explorative Dateanalyse. Springer, Heidelberg pp. 6-37.
- CHAMBERS, J.M. (1980) : Statistical Computing : History and Trends.
Amer. Stat. 34, 238-243.
- CHAMBERS, J.M. (1983) : The New Future of Data Analysis.
Bull. I.S.I. 50,1, 97-103.
- CHAN, T.F., GOLUB, G.H. and LE VEQUE, R.J. (1982) : Updating Formulae
and a Pairwise Algorithm for Computing Sample Variances.
In : CAUSSINUS, H., ETTINGER, P. and TOMASSONE, R. (eds) :
COMPSTAT 1982, Physika Verlag Wien pp. 30-41.

- FRANCIS, I. (ed.) (1979) : A comparative Review of Statistical Software.
Int . Ass. Stat. Comp., Voorburg pp. 658.
- FRANCIS, I. (1981) : Statistical Software -A Comparative Review.
North Holland, New-York pp. 542.
- FRANCIS, I., HEIBERGER, R.M. and VELLEMAN, P.F. (1975) : Criteria and
Considerations in the Evaluation of Statistical Packages.
Amer. Statist. 29, 52-56.
- FUKUNAGA, K. and KESSELL, D.I. (1971) : Estimation of Classification Error.
IEEE Trans. Comp. 20, 1521-27.
- GALE, W.A. and PREGIBON, D. (1982) : An expert System for Regression
Analysis. In : HEINER, K.W., SACHER, R.S. and WILKINSON, J.W.
(eds.) : Proc. Comp. Sci. and Statistics : 14th Ann. Symp.
Interface. Springer, New York pp. 110-117.
- GOWER, J.C., SIMPSON, H.R. and MARTIN, A.H. (1967) : A Statistical Program-
ming Language.
Appl. Statistics 16, 89-99.
- GREENWOOD, J.A. and HARTLEY, H.O. (1962) : Guide to Tables in Mathematical
Statistics.
Princeton Univ. Press pp. 1014.
- HAUX, R. (1983/84) : Statistical Analysis Systems - Construction and
Aspects of Methods Design.
Part I : Stat. Softw. News1. 9, 106-115,
Part II: Stat. Softw. News1. 10, 14-27.
- HULTSCH, E., JANNASCH, H., KRIER N., SUND, M. and VICTOR, N. (1978) :
Anforderungen an Programmsysteme zur statistischen Datenanalysis.
Stat. Softw. News1. 4, 2-30.

- MANN, N.R. (1978) : Everything You always Wanted to Know about the History of Computer Science and Statistics : Annual Symposia on the Interface -- and more.
In : GALLANT, A.R. and GERIG, T.M. (eds.) : Proc. Comp. Sci. and Statistics. : 11th Ann. Symp. Interface. Inst. Stat., N.C. State Univ., Raleigh pp. 2-5.
- MARCUS, R., PERITZ, E. and GABRIEL, K.R. (1976) : On Closed Testing Procedures with Special Reference to Ordered Analysis of Variance. *Biometrika* 63, 655-660.
- MILTON, R.C. and NELDER, J.A. (1969) : Statistical Computation. Academic Press, New York pp. 462.
- RAO, C.R. (1948) : The Utilization of Multiple Measurements in Problems of Biological Classification (with discussion). *J. Roy. Stats. Soc. B* 10, 159-203.
- RIPLEY, B.D. (1983) : Computer Generation of Random Variables : A Tutorial. *Int. Stat. Review* 51, 301-319.
- TUKEY, J.W. (1977) : Exploratory Data Analysis. Add.-Wesley Publ. Comp., Reading pp. 688.
- WIRTH, N. (1976) : Algorithms + Data Structures = Programs. Prentice Hall Inc., Englewood Cliffs pp. 366.
- ZELEN, M. (1983) : Biostatistical Science as a Discipline : A look into the Future (with discussion). *Biometrics* 39, 827-837.