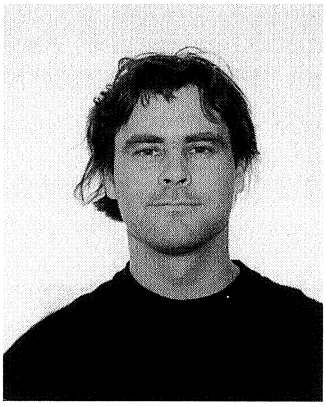
IN THIS ISSUE

B-SPLINES WITHOUT DIVIDED DIFFERENCES

Regression splines is one of the nonparametric method employed when the relationship between two random variables is not linear. A desirable basis for the implementation of regression splines is the B-spline basis, because it is almost orthogonal. As B-splines are usually defined with divided differences, Todd Mackenzie and Michal Abrahamowicz in the first article of this issue present a new definiion of B-splines without divided differences.

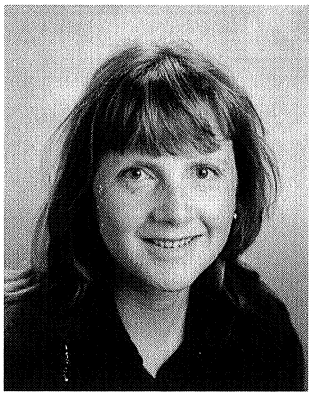


Todd Mackenzie is a Ph.D. candidate at McGill University, Montreal. He has been employed as a research assistant at the Division of Clinical Epidemiology of the Montreal General Hospital for over 8 years, where he has worked predominantly with rheumatologist, John Esdaile, and the biostatistician Michal Abrahamowicz.

A NEW IMPUTATION METHOD BASED ON THE SAMPLE VECTOR CORRELATION RV

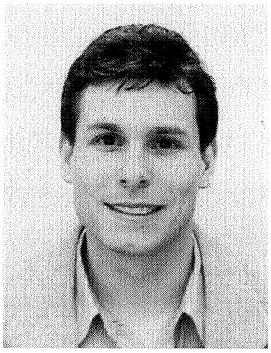
Missing observations occur quite frequently in designed experiments and observational studies. Analyzing data of this nature without appropriate knowledge may lead to inappropriate inferences about the hypothesis and, consequently, to wrong conclusions about the results of the experiments. For a multivariate data set, when certain observations are missing at random, Fabienne Crettaz von Roten proposes a new imputation method based on the correlation structure of the data.

Fabienne Crettaz von Roten was born November 2, 1962 in Sierre, Switzerland. From 1982 to 1987, she studied Mathematics at the Swiss Federal Institute of Technology in Lausanne. In 1993, she obtained her Ph.D. from the same institute. She is currently working as teacher and consultant at the Social Sciences faculty of the University of Lausanne.



CONSTRUCTION OF STRATEGIC GROUPS

In order to find strategic groups adapted to the study of services strategies, Olivier Furrer mixes analysis of variance and principal component analysis. This original approach seems efficient since the strategic groups are very well identified and very distinct from each other.



Olivier Furrer was born on December 18, 1968 in Lausanne. He obtained his Bachelor of Science in Business Administartion from the University of Neuchâtel in 1992. Mr. Furrer is currently a research assistant with the Management Group at the University of Neuchâtel.

FROM THE EDITOR'S NOTEBOOK

The *minmad* (the minimizer of the median of the absolute deviations) is sometimes presented as an alternative to the *mean* and to the *median* in order to describe the center of a distribution or of a data set. In comparing the *influence functions* of these three statistics we learn that the *minmad* is not so nicely behaved and makes one mad.

GEORGE A. BARNARD

FRAGMENTS OF A STATISTICAL AUTOBIOGRAPHY

We are pleased to present in this issue a paper written by Professor George A. Barnard for *Student*. Comparing the 50 years parallelism between the history of logic and statistics, he hopes for a cheerful consensus between the Fisherian, the Jeffreys, the Neyman-Pearsonian, and the Bayseian. In his "Fragments of a Statistical Autobiography" he writes that "Any professional statistician now needs to be familiar with all four approaches, and their relations to each other, so that an appropriate choice of approaches, or combination of approaches, can be made to problems as they arise."

DATA & STATISTICS

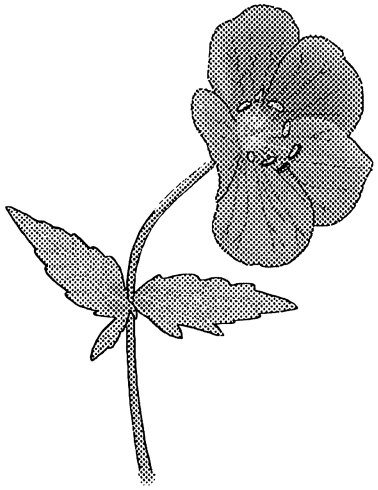
A.S.C. Ehrenberg explains his vision via seven do's and don'ts for a successful data analysis using two data sets published in *Data & Statistics* (1996, Vol.1, No.3). Comments are then made on Ehrenberg method of data analysis by David Hand and Joe Whittaker. A new data set is also introduced by L. Gusmão, M. do C. Pinheiro-Alves and J.T. Mexia, on "Total Fat Content of Olive Pastes by Nuclear Magnetic Resonance".

Acknowledgment of Referees' Services

The following individuals served as referees of papers for *Student* from January 1995 to September 1996. We are grateful for their outstanding and indispensable contribution.

The Editor

Gérard Antille	Hans Bock	David Cox
Miklos Csörgö	Yves Escoufier	Riccardo Gatto
Robert Gray	David Hand	Samad Hedayat
Jean-Marie Helbling	Jochen Jesinghaus	Horva'th Lajos
Farhad Mehran	Vijayan N. Nair	Hans T. Nyquist
Jonathan D. Parker	Simon Sheather	Stephen Senn
Bernard Van Cutsem	Arthur Vogt	Joe Whittaker



Student (1996), Vol.1, No.4, 223-230

B-Splines Without Divided Differences

Todd Mackenzie and Michal Abrahamowicz

Department of Mathematics and Statistics, McGill University, Montreal, Canada

Division of Clinical Epidemiology, Montreal General Hospital, Canada

Abstract: Many approaches to nonparametric regression, for instance, smoothing splines and regression splines involve estimators that are splines (piecewise polynomials). These approaches are generally implemented by first constructing a spline basis. The B-spline basis is appealling because it is almost orthogonal and easy to compute. Typically the B-spline is defined using the calculus of divided differences (de Boor, 1978, Curry and Schoenberg, 1966, Schumacher, 1981). This definition is opaque to those unfamiliar with this calculus. This article presents an alternative derivation of the B-spline aimed at statisticians. Beginning with the desire of obtaining a spline basis which is as orthogonal as possible and using the reasoning that any two functions whose supports do not intersect are orthogonal we define a B-spline to be a spline of minimal support. Using elementary tools from linear algebra we show, in succession, that (i) splines of finite support do indeed exist, (ii) an r-th degree spline must contain at least r knots in its support, (iii) B-splines are 'bumps' and (iv) derive a formula for the B-spline.

Key words: Nonparametric modelling, regression splines, smoothing splines, penalized regression splines.

1 Introduction

Nonparametric methods are employed when it does not seem suitable to impose the rigid assumptions associated with parametric models. An example of nonparametric regression is the estimation of a conditional expectation $\alpha(x) = E[Y|X = x]$ without assuming a linear model, $\alpha(x) = \alpha_0 + \alpha_1 x$. Eubank (1988), Hastie and Tibshirani (1990) and Green and Silverman

(1994) are excellent textbooks devoted to nonparametric regression. Three popular methods are *local likelihood*, of which *kernel smoothing* is a special case, *smoothing splines* and *regression splines*. Smoothing splines arise as the mathematical solution to the optimization problem of maximizing a penalized likelihood, when the penalty, the term to be subtracted from the log-likelihood, is of the form $\lambda \int \alpha''(x)^2 \mu(dx)$. The method of *regression splines* generalizes linear regression by enlarging the space in which the function α is assumed to exist, from two dimensions, as with the model $\alpha(x) = \alpha_0 + \alpha_1 x$, to an arbitrary dimension that may depend on the sample size. It is similar to polynomial regression but is more *local* (Ramsay, 1988). The method of *penalized regression splines*, is a hybrid of smoothing splines and regression splines, (Gray, 1992).

Implementation of smoothing splines or regression splines generally involves the construction of a spline basis. The *B-spline* basis is desirable because it is almost orthogonal and it is easy to compute. Typically B-splines are defined using *divided differences* (de Boor, 1978, Curry and Schoenberg, 1966, Schumacher, 1981). This definition is opaque to those unfamiliar with divided difference.

We propose an alternative definition of the B-spline. We begin with the desire of obtaining a spline basis that is as close to orthogonal as possible. Two real-valued functions f_1 and f_2 are orthogonal if their inner product, $\int f_1(x) f_2(x) \mu(dx)$, is zero, where μ is some measure. Two functions are orthogonal if their supports do not intersect. This suggests that one strategy for obtaining an orthogonal basis is to find a basis whose elements have support that intersect each other as little as possible. This reasoning guides our definition of a B-spline found below.

2 Spline preliminaries

A (polynomial) spline is a piecewise polynomial subject to continuity constraints. All the pieces have the same degree, say r, and the first $r-1$ derivatives are continuous at the junctures, called *knots*, of any two adjacent pieces. Specification of more than $r-1$ derivatives would force the piecewise polynomial to be a polynomial. A quadratic spline is continuous and has continuous slope whereas a cubic spline also has continuous curvature.

In what follows $t_0 < t_1 < \ldots < t_m < t_{m+1}$ shall indicate the knot sequence. We refer to $t_1, \ldots, t_m$ as *interior knots*. The points t_0 and t_{m+1} are *boundary* knots and define the domain of the spline. We shall denote the set of splines with these knots as $S^r(t_1, \ldots, t_m)$. This set is a linear space. An elementary basis is generated, in part, by *plus* (also known as *truncated*)

powers, $(t-t_i)_+^r = \{\max[0,(t-t_i)]\}^r$. The $r+1$ powers, $1, t, \ldots, t^r$ together with the m plus powers $(t-t_1)_+^r, \ldots, (t-t_m)_+^r$ is known as the *plus powers basis.* If m is large, the plus power basis is numerically unstable because the plus powers are highly collinear. Orthonormalization of the plus power basis is not a satisfactory solution as it requires $O(m^2)$ steps.

3 An alternative B-spline derivation

3.1 Minimal support

A simple approach to non-parametric regression consists of the following two steps: (i) partition the range of the predictor into intervals, by choosing some knots (cutpoints), (ii) estimate the mean, or other characteristics, of the primary variable within each interval. In this setting one is regressing onto a linear space of piecewise constant functions. This is the same idea as density estimation by histograms. The indicator functions, $I_{[t_i,t_{i+1}]}(x)$, one for each interval, are an ideal basis because they are orthogonal, easy to calculate and they yield estimates which are easy to understand. This suggests that an ideal basis for $S^r(t_1,\ldots,t_m)$ consists of functions of small support. Such bases are almost orthogonal for any two basis elements whose supports do not overlap are orthogonal. This motivates the following definition:

Definition 3.1: A B-spline is a non-zero spline of minimal support.

In other words, a B-spline is a non-zero spline whose support contains a minimal number of knots. We shall show that this minimal number is r. Note that the boundary of the support of a spline, if it is finite, consists of knots (because in between knots it is a polynomial). This definition applies easily to $S^1(t_1,\ldots,t_m)$: in this space the 'triangle' functions, which equal zero outside (t_i, t_{i+2}), are B-splines.

For splines of arbitrary degree, $r > 1$, we need to show that splines of finite support do exist. Let $s(t) = \sum_{k=0}^{p} \beta_k (t-t_{i+k})_+^r$, which is 0 for $t < t_i$. For t **larger** than t_{i+p}, the spline $s(t)$, is a $(p+1)$-fold linear combination of r-th degree polynomials. Therefore if $p+1 > r+1$ there exists some $\boldsymbol{\beta} \neq \mathbf{0}$ for which $s(t) = 0$ for $t \geq t_{i+p}$. In particular, take $p = r+1$. There exists an *r-th* degree spline whose support is contained in (t_i, t_{i+r+1}) and therefore contains at most r knots. Can we go lower than r?

For $t \geq t_{i+p}$, $s(t)$ is the polynomial in t for which t^l has coefficient

$$(-1)^{r-l}\binom{r}{l}\left\{\sum_{k=0}^{p}\beta_k t_{i+k}^{r-l}\right\}$$

for $l = 0, \ldots, r$. If $s(t) = 0$ for $t > t_{i+p}$ then each of these coefficients must be 0. Therefore $\boldsymbol{\beta}$ satisfies the matrix equation $\mathbf{V_p}\boldsymbol{\beta} = \mathbf{0}$ where matrix $\mathbf{V_p}$ consists of the first $(1+p)$ columns of the *Van der Monde* matrix, $\mathbf{V}$ whose entries are $\mathbf{V}_{kl} = t_{i+l}^k$, for $k = 0, 1, \ldots, r$ and $l = 0, 1, \ldots, r$. Van der Monde matrix $\mathbf{V}$ is full rank, because we have assumed that the knots t_i are distinct. Therefore for $p \leq r$ there exists no $\boldsymbol{\beta} \neq \mathbf{0}$ for which $\mathbf{V_p}\boldsymbol{\beta} = \mathbf{0}$. The implication of this is that the support of a B-spline, 'beginning' at t_i must end at t_{i+r+1} and contains exactly r knots!

B-splines are *bumps*. By a bump we mean a function which is constant 0, then increasing, then decreasing, then constant 0 again. For instance it has exactly one local extremum. If it had more than one local extrema then the derivative $DB_i^r(t)$ which is a degree $r-1$ spline would have more than two, $D^2B_i^r(t)$ which is a degree $r-2$ spline would have more than three, ..., and $D^{r-1}B_i^r(t)$ which is a linear spline, and whose support equals (t_i, t_{i+r+1}) would have more than r local extrema which is impossible.

3.2 A B-spline formula

Let $B_i^r(t) = \sum_{k=0}^{r+1} \beta_k (t - t_{i+k})_+^r$ be the B-spline whose support begins at t_i. The vector of coefficients, $\boldsymbol{\beta}$, satisfies

$$\begin{pmatrix} 1 & 1 & \ldots & 1 \\ t_i & t_{i+1} & \ldots & t_{i+r} \\ \vdots & \vdots & \ddots & \vdots \\ t_i^r & t_{i+1}^r & \ldots & t_{i+r}^r \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_r \end{pmatrix} = -\beta_{r+1} \begin{pmatrix} 1 \\ t_{i+r+1} \\ \vdots \\ t_{i+r+1}^r \end{pmatrix}.$$

Cramer's Rule (for solving matrix equations) yields the solution

$$\beta_j = -\beta_{r+1} |\mathbf{V}_j^*| / |\mathbf{V}|$$

where $\mathbf{V}_j^*$ is the matrix obtained by substituting $(1, t_{r+1}, \ldots, t_{r+1}^r)'$ for column $(1, t_{i+j}, \ldots, t_{i+j}^r)'$ in $\mathbf{V}$. Matrix $\mathbf{V}_j^*$ is also Van der Monde. The determinant of a Van der Monde matrix with entries u_j^i is $\Pi_{i<j}(u_j - u_i)$. After some cancellation we obtain

$$\beta_j = -\beta_{r+1} \frac{\Pi_{\substack{k=0 \\ k \neq j}}^{r} (t_{i+k} - t_{i+r+1})}{\Pi_{\substack{k=0 \\ k \neq j}}^{r} (t_{i+k} - t_{i+j})} \qquad j = 0, 1, \ldots, r.$$

The choice of $\beta_{r+1} = -\rho / \Pi_{k=0}^r (t_{i+k} - t_{i+r+1})$, for some $\rho \neq 0$ whose presence is explained below, simplifies this expression to

$$\beta_j = \frac{\rho}{\Pi_{\substack{k=0 \\ k \neq j}}^{r+1} (t_{i+k} - t_{i+j})} \qquad j = 0, 1, \ldots, r, r+1.$$

This solution yields the following formula for the B-spline,

$$B_i^r(t) = \rho \sum_{j=0}^{r+1} \frac{(t - t_{i+j})_+^r}{\Pi_{\substack{k=0 \\ k \neq j}}^{r+1} (t_{i+j} - t_{i+k})}. \tag{1}$$

The final term of the sum above can be dropped if instead we specify that $B_i^r(t)$ is 0 outside of (t_i, t_{i+r+1}). This is the plus powers expansion of a B-spline.

The choice of $\rho = t_{i+r+1} - t_i$ yields the B-splines defined by de Boor (1978) and Schumacher (1981). They have the attractive property of being bounded below by 0 and above by 1. The choice $\rho = r + 1$ yields the *M-splines* which are density functions. They have been employed in density estimation (Abrahamowicz, Ciampi and Ramsay, 1992).

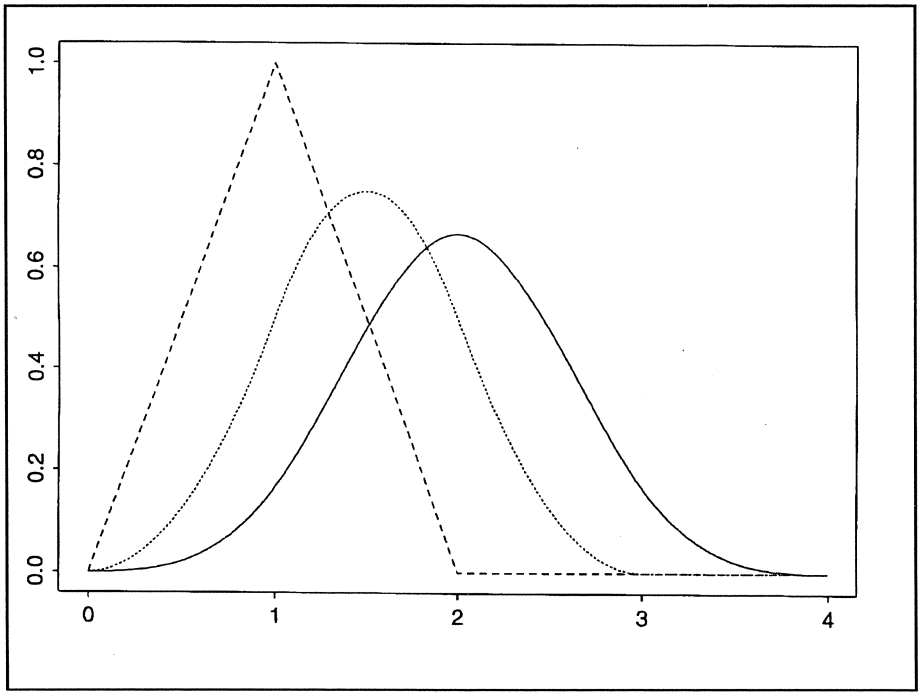


Figure 1: Superimposed, from left to right, a linear, quadratic and cubic M-spline (standardized cubic B-spline); these are also, respectively, the 2, 3 and 4-fold convolutions of $U(0, 1)$

3.3 Example: convolutions of $U(0,1)$

If $t_i = i$ the M-spline, whose support begins at 0, is

$$\frac{1}{(r-1)!}\sum_{k=0}^{r-1}(-1)^k\binom{r}{k}(t-k)_+^{r-1}.$$

This is the density (proof by induction) of the sum of r independent standard uniformly distributed random variables! Figure 1 superimposes the 2, 3, and 4-fold convolutions of $U(0,1)$.

3.4 Recursion formula for B-splines

If $t_i = i$, each term in the plus powers expansion of the B-spline (1) contains a combinatorial term, $\binom{r+1}{k}$. The presence of this term hints that a recursive formula, similar to Pascal's identity, might exist. Indeed, for B-splines,

$$B_i^r(t) = \frac{t-t_i}{t_{i+r}-t_i}B_i^{r-1}(t) + \frac{t_{i+r+1}-t}{t_{i+r+1}-t_{i+1}}B_{i+1}^{r-1}(t) \qquad r \geq 1$$

This is the *recursion formula* for B-splines (de Boor, 1978). The recursion formula for M-splines is slightly different (Ramsay, 1988).

The recursion formula has two advantages over the plus powers expansion. First, the recursion formula does not require the knots to be distinct. Second, the recursion formula yields less rounding error (de Boor, 1978).

3.5 B-spline basis

By introducing r knots $t_{-r},\ldots,t_{-1}$ to the left of t_0 and r knots $t_{m+2},\ldots,$ t_{m+1+r} to the right of t_{m+1} a basis for $S^r(t_1,\ldots,t_m)$ can be constructed out of B-splines. Together the $m+r+1$ B-splines, $B_{-r}^r,\ldots,B_m^r$ form a basis (de Boor, 1978). This basis is well conditoned because B_i^r and B_j^r are orthogonal unless $|i-j| \leq r$. Figure 2 depicts a cubic B-spline basis.

Acknowledgments: This work was supported in part by the Fonds pour la Formation de Chercheurs à l'Aide à la Recherche and the Natural Sciences and Engineering Research Council. The authors are indebted to Alain Vandal and Roxane du Berger.

Résumé: Beaucoup d'approches en régression non paramétrique, telles que le lissage spline et la regression spline supposent des estimateurs splines (polynômes par pièces). Ces approches nécessitent généralement la construction préalable d'une base spline. La base B-spline est avantageuse du fait qu'elle est presque orthogonale et facile à calculer. Elle est définie à

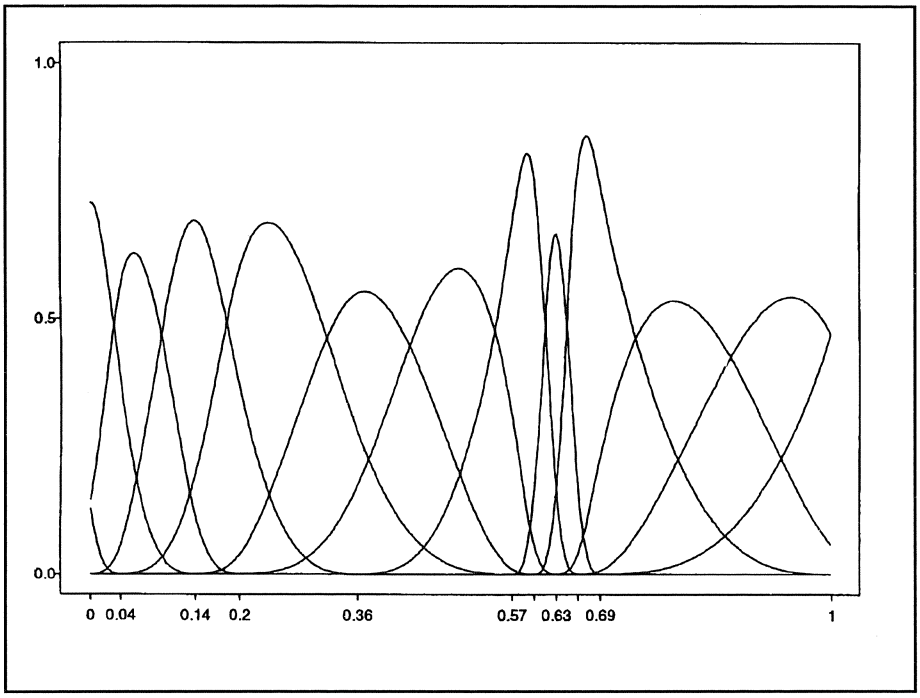


Figure 2: A B-spline basis for a space of cubic splines with 9 interior knots; the knots locations are indicated by ticks

l'aide de calculs de différences divisées (de Boor, 1978; Curry et Schoenberg 1966; Schumacher, 1981). Cette définition est cependant difficile pour les personnes non familiarisées à ce type de calcul. Cet article présente une dérivation alternative de la B-spline à l'attention des statisticiens. En essayant d'obtenir une base spline aussi orthogonale que possible et en utilisant le raisonnement affirmant que deux fonctions dont les supports sont disjoints sont orthogonales, nous définissons une B-spline comme étant une spline de support minimal. En utilisant des outils mathématiques élémentaires d'algèbre linéaire, nous montrons successivement que (i) une spline de support fini existe vraiment, (ii) une spline de r-ième degré doit contenir au moins r noeuds dans son support, (iii) les B-splines sont 'bossues' et (iv) on peut dériver une formule pour le B-spline.

Mots clés: Modélisation non-paramétrique, régression spline, lissage de splines, régression spline avec pénalité.

References

[1] Abrahamowicz, M., Ciampi, A. and Ramsay, J.O. (1992), "Nonparametric Density Estimation for Censored Survival Data: Regression Spline Approach", *Canadian Journal of Statistics*, **2**, 171-175.

[2] deBoor, C. (1978), *A Practical Guide to Splines*, Springer-Verlag, New York.

[3] Curry, H.B. and Schoenberg, I.J. (1966), "On Polya Frequency Functions. IV: The Fundamental Spline Functions and Their Limits", *J. Analyse Math.*, **17**, 71-107.

[4] Eubank, R.L. (1988), *Smoothing Splines and Nonparametric Regression*, Marcel Dekker, New York and Basel.

[5] Gray, R.J. (1992), "Flexible Methods for Analyzing Survival Data Using Splines, With Applications to Breast Cancer Prognosis", *J. of the Amer. Statist. Assoc.*, **87**, 942-951.

[6] Green, P.J. and Silverman, B.W. (1994), *Nonparametric Regression and Generalized Linear Models*, Chapman and Hall, London.

[7] Hastie, T.J. and Tibshirani, R.J. (1990), *Generalized Additive Models*, Chapman and Hall, New York.

[8] Ramsay, J.O. (1988), "Monotone Regression Splines in Action", *Statistical Science*, **3**, 425-461.

[9] Schumacher, L.L. (1981), *Spline Functions: Basic Theory*, John Wiley, New York.

Student (1996), Vol.1, No.4, 231-240

A New Imputation Method Based on the Sample Vector Correlation *RV*

Fabienne Crettaz von Roten

Applied Mathematical Institute, University of Lausanne, Switzerland

Abstract: In this paper, we propose a new imputation method in the context of multivariate analysis. Our method relies on the correlation structure of the data and imputes with the value that maximises the sample vector correlation *RV* (Escoufier, 1973). This article involves first the development of the method based on the *RV* coefficient, then a comparison with eight existing methods on the basis of simulations and real data.

Key words: Buck's method, imputation methods, method based on the *RV* coefficient, missing data, vector correlation.

1 Introduction

In the context of multivariate statistical inference and data analysis, the missing data problem happens very frequently, as in the case of mechanical breakdowns unrelated to the experimental process, of respondents refusal to report an item, and so on. The statistician should analyse the mechanism leading to missing values since the performance of any missing-data techniques depends heavily on this mechanism. In order to simplify we frequently assume that the data are completely missing at random (MCAR in Rubin, 1976) or missing at random (MAR). We judge the tenability of this assumption either empirically (data and missing-data mechanism knowledge) or theoretically (test for MCAR). Essentially all the literature on multivariate incomplete data assumes that the data are MCAR or MAR, so we assume it too here.

Missing data methods can be roughly grouped into the following two approaches:

1) parameters estimation from incomplete data with appropriate formula followed by analyses based on an initial reduction of the data to the sample mean vector and sample covariance matrix of the variables,

2) filling in the missing values ("imputation") followed by statistical analysis of the completed data by standard methods.

The EM algorithm lies in the boundaries of the two approaches since iteratively the E step imputes the missing values and the M step finds maximum likelihood estimates of the parameters (Dempster, Laird and Rubin, 1977). Meng and Rubin (1991) propose an important supplement to EM algorithm (the SEM algorithm) that produces asymptotic variance-covariance matrices for maximum likelihood estimates. The EM algorithm can also provide robust estimation when the data are described as a mixture of normal distributions with the same mean and different variances (Little and Rubin, 1987).

We focus on the second approach because of its various advantages: use of data collector knowledge, use of standard complete-data methods, possible comparison of multi-user results, and absence of problem related to not positive definite covariance matrix or to correlation that lies outside the range $[-1, 1]$.

The literature on imputation methods is prolific (see Little et Rubin, 1987). We will make a distinction between:

- Methods based on statistical models:

a) the method based on the EM algorithm sets a normal multivariate model and imputes iteratively with the conditional mean of the missing values given the observed data and current estimated parameters,

b) the Buck's method sets a regression model and estimates the missing values by predicted values from linear regressions of the missing variables on the present variables (one or many) case by case (Buck, 1960); for normal data, this method is closely related to the EM algorithm but it is widely applied to non-normal data.

- Data analysis methods:

a) the mean method imputes missing values by an estimate of the mean of the variable (Wilks, 1932),

b) the "hot deck" procedures use distances between individuals and impute missing values by values from responding units close to the unit with the missing value (Ford, 1983).

Filled-in data method can be evaluated on the imputation quality (if we know the true value) or on the quality of estimates from the filled-in data. Efron (1993) has proposed two non-parametric bootstrap methods for judging the effectiveness of an estimate after imputation. In general there is no superior imputation method; simulations and real data have tried to distinguish situations when the methods work from situations when they fail.

At first, the imputation techniques seem very different but a deep analyse shows a common denominator: most methods minimise a distance (Crettaz von Roten, 1992). Our idea was to develop a new data analysis method based on the minimisation of a distance between two clouds of points or equivalently on the maximisation of the link between two groups of variables. This idea is at the origin of the RV method that we will expose in this article.

We first define the vector correlation ρV and the sample vector correlation RV and develop the method based on RV. After that we compare the RV method with eight existing methods on the basis of simulations and real data.

2 Methodology

2.1 The RV coefficient

Let $\mathbf{X} = (\mathbf{X}^{(1)}, \mathbf{X}^{(2)})'$ be a random vector partitioned into two subvectors $\mathbf{X}^{(1)}$: $p \times 1$ and $\mathbf{X}^{(2)}$: $q \times 1$. The partition of $\mathbf{X}$ into subvectors happens naturally for example in the case of endogenous and exogenous variables, or physical and intellectual measurements.

The population mean vector $\mathbf{u}$ and the covariance matrix $\mathbf{\Sigma}$ are:

$$\mathbf{u} = \begin{pmatrix} \mathbf{u}^{(1)} \\ \mathbf{u}^{(2)} \end{pmatrix} \text{ and } \sum = \begin{pmatrix} \mathbf{\Sigma}_{11} & \mathbf{\Sigma}_{12} \\ \mathbf{\Sigma}_{21} & \mathbf{\Sigma}_{22} \end{pmatrix}.$$

Escoufier (1973) introduces the following vector correlation ρV between $\mathbf{X}^{(1)}$ and $\mathbf{X}^{(2)}$:

$$\rho V = \rho V\left(\mathbf{X}^{(1)}, \mathbf{X}^{(2)}\right) = \frac{tr(\mathbf{\Sigma}_{12}\mathbf{\Sigma}_{21})}{\sqrt{tr(\mathbf{\Sigma}_{11}^2)tr(\mathbf{\Sigma}_{22}^2)}}$$

where $tr(\cdot)$ is the trace operator.

This coefficient is a generalisation of the simple square correlation coefficient in the case of two groups of variables; it varies between 0 (no association between the two groups of variables) and 1.

Given a sample $\mathbf{X}_1, \mathbf{X}_2, ..., \mathbf{X}_n$, Escoufier defines the sample vector correlation by replacing in the expression of ρV the parameters by the usual estimators:

$$RV = RV\left(\mathbf{X}^{(1)}, \mathbf{X}^{(2)}\right) = \frac{tr(\mathbf{S}_{12}\mathbf{S}_{21})}{\sqrt{tr(\mathbf{S}_{11}^2)tr(\mathbf{S}_{22}^2)}}$$

where $\mathbf{S}_{ij} = \frac{1}{n-1} \sum_{\alpha} (\mathbf{X}_{\alpha}^{(i)} - \overline{\mathbf{X}}^{(i)})(\mathbf{X}_{\alpha}^{(j)} - \overline{\mathbf{X}}^{(j)})' \quad i,j = 1,2$ and $\overline{\mathbf{X}}^{(i)}$, $\overline{\mathbf{X}}^{(j)}$denote the i^{th} and j^{th} part of the sample mean vector of the elements $\mathbf{X}_a^{(i)}$ and $\mathbf{X}_a^{(j)}$ $(1 \leq \alpha \leq n)$.

RV is a measure of similarity between two clouds of points; more the coefficient RV is close to 1, more one cloud can be used as a substitute for the other cloud.

2.2 The RV method

Let

$$\mathbf{X} = (\mathbf{X}_1, \mathbf{X}_2, ..., \mathbf{X}_n) = \begin{pmatrix} \mathbf{X}_1^{(1)} & \mathbf{X}_2^{(1)} & ... & \mathbf{X}_n^{(1)} \\ \mathbf{X}_1^{(2)} & \mathbf{X}_2^{(2)} & ... & \mathbf{X}_n^{(2)} \end{pmatrix}$$

be a data matrix with n cases and $(p+q)$ variables.

Suppose that there are missing data in the first part of the vector $\mathbf{X}_n$, we denote this missing vector $\mathbf{x}$ and we write:

$$\mathbf{X}_n = \begin{pmatrix} \mathbf{x} \\ \mathbf{y} \\ \mathbf{z} \end{pmatrix}$$

where $\mathbf{y}$ is the complete part of the first group of variables and $\mathbf{z}$ the second group of variables.

If the index $n-1$ added to the sample covariance matrix $\mathbf{S}$ and the mean vector $\overline{\mathbf{X}}$ denotes that the first $n-1$ individuals are used for estimation (that is the individuals without missing data), $\mathbf{S}_n$ is a function of $n, \mathbf{x}, \mathbf{y}, \mathbf{z}, \overline{\mathbf{X}}_{n-1}$ and $\mathbf{S}_{n-1}$. Hence we can write the coefficient $RV(\mathbf{X}^{(1)}, \mathbf{X}^{(2)})$ as a function of these elements (the exact form of this function may be found in Crettaz von Roten and Helbling, 1991):

$$RV\left(\mathbf{X}^{(1)}, \mathbf{X}^{(2)}\right) = f\left(n, \mathbf{x}, \mathbf{y}, \mathbf{z}, \overline{\mathbf{X}}_{n-1}, \mathbf{S}_{n-1}\right).$$

Our method imputes with the value $\mathbf{x}^*$ maximising the function $RV\left(\mathbf{X}^{(1)}, \mathbf{X}^{(2)}\right)$ (i.e., maximising the link between the two groups of variables). Other idea of imputation based on the RV coefficient were left out of consideration since they induce the same well-known bad properties of

the existing imputation methods. See (Crettaz von Roten, 1993) for the proof of the existence of a maximum.

The RV method imputes values such that the two clouds of points, corresponding to the two groups of variables, are as close as possible; so this method is applicable only if there is a sufficient similarity between the two clouds of points; this point will be analysed in Section 2.3.

This method is characterised by the following properties:

a) a new imputation principle based on similarity between two clouds of points,

b) no hypotheses about model or distribution, which is a great advantage since we rarely know the correct model specification,

c) an easy implementation (simple matrix operations, no inversion matrix),

d) the global treatment of the missing values of an unit: when a unit has m missing values, Buck's method has the disadvantage of using different regression models for imputation (each variable with missing value requires a model that explains it in terms of the available variables); contrarily the RV method imputes simultaneously the m missing values of a unit by maximising the RV function of m variables.

Generalisations:

1) This method is easily generalised when missing values lie in the two groups of variables: the formula become more complicated but the principle remains the same.

2) If r observations have missing values, the RV method will treat separately and sequentially each one of the r incomplete vectors with the $(n-r)$ complete vectors (the two groups of variables remain the same).

We find many real data with missing values and outliers, but most imputation methods are sensible to outliers in the complete set of observations. The RV coefficient may be deeply influenced by an extreme value: an outlier can create an artificial structure in the data and induce an erroneous imputation by the RV method. In case of outliers and missing data, two different approaches are available: the first approach detects the outliers, eliminates them and imputes the missing values, the second approach imputes with a technique that minimises the influence of outliers (for example, regression methods using the norm L_1 or the least median of squares). Following this second approach, we propose a robust version of the RV method that uses robust estimators of the sample mean vector and covariance matrix evaluated on the complete observations. The principle remains the same but we use robust estimation technique. The analysis of real data with outliers and missing data yields the superiority of the robust

version to the standard version of the *RV* method in this case (Crettaz von Roten, 1993).

2.3 Performances of the *RV* method

In order to compare the *RV* method with existing methods, we generate multivariate normal variables of dimension four (partitioned into two groups of two variables) and we allocate randomly missing values. We adopt the four following factors:

A = percent of missing data (we choose three values 3%, 6% and 15%),
B = size of the sample (we choose two values 50 and 100),
C = pattern of the missing values (we choose here two patterns of missing values set in the first group of variables: in the first pattern the second variable is more observed than the first variable, in the second pattern all is missing in the first group of variables),
D = the covariance matrix of the generated data (we choose 8 matrices with vector correlation lying between 0.60 and 0.94, and with variances equal to 1).

For each of the 96 ($= 3 \cdot 2 \cdot 2 \cdot 8$) combinations of the four factors, we generate 1000 samples. We decide to compare 9 methods: 3 mean methods, 5 regression methods (with the *RM* method based on maximum likelihood estimation) and the *RV* method.

The performances of each method is evaluated with 3 different types of criteria found in Gleason and Staelin (1975):

a) a distance between the real data (known here since we have randomly deleted some data) and the imputed data, named Q_α:

$$Q_\alpha = \sqrt{\sum_{i=1}^{p+q}\sum_{j=1}^{n} \frac{(X_{ij} - X_{ij}^\alpha)^2}{(p+q)AB}}$$

where X_{ij}^α is the filled-in data by α method,

b) a distance between the real sample covariance matrix (before generating missing data) and the sample covariance matrix after imputation, named D_α:

$$D_\alpha = \sqrt{\sum_{i=1}^{p+q}\sum_{j=1}^{n} \frac{(S_{ij} - S_{ij}^\alpha)^2}{(p+q)(p+q-1)}}$$

where S_{ij}^α is element of the sample covariance matrix of the filled-in data by means of the α method,

c) bias and mean square of the estimates of the mean and the covariance matrix (the formula used for theses criteria may be found in Crettaz von Roten, 1993).

A small value for the first two criteria represents a successful imputation method. The last group of criteria allows to analyse the quality of the parameters estimation after the imputation (means, variances and covariances) and to detect imputation methods that systematically impute by a lower or upper value or that systematically underestimate the variance-covariance matrix.

Each method has different significant factors for the first criteria. The performances of the *RV* method depends on the factors C and D, the mean methods on factors A, B and C and the regression methods on factors A, C and D.

Generally, the smallest bias are found for the methods *RV* and *RM*. All methods except *RV* have estimated bias for the variance-covariance matrix negative since they underestimate systematically the variability. The *RV* method performs differently from the other methods since it does not impute systematically in the centre of the distribution, in that way it does not underestimate the variance. The smallest mean squares are found for the mean methods; we knew already this result since these methods impute systematically in the centre of the distribution; in what follows, we will not allude to this criteria since it does not teach us anything.

In order the simplify the results, we choose 3 methods representative of the 3 different families of imputation methods.

A level of a factor corresponds to a set of simulations; each method is performed on this set of simulation and gets the 16 averaged criteria (2 distances plus 14 bias). We repeat this for each level of each factor. Table 1 indicates how many times the 3 selected methods were the best in regard of the 16 criteria (ties may occur).

The choice of factor A was relevant since we observe changes when percents vary. The mean methods lose their efficiency when the percent grow (less optimal criteria), whereas some regression methods stay stable and some regression methods as well as the *RV* method become a little bit better.

In combining factor A and B, we observe that the performances of the methods *RM* and *RV* are opposite: at the same percent of missing data, *RV* is better (at the level of bias) for a small size and *RM* for a big size.

The size of sample is the least significant factor: it seems that a size of 50 is large enough to show no great differences compared with a size of 100 for data with dimension 4.

In general, most mean methods and *RV* method perform better under

	factor	mean = MA	regression = RM	RV
	3%	2	9	4
A	6%	1	7	5
	15%	0	8	5
B	50	0	8	5
	100	0	9	4
C	1	0	7	4
	2	0	3	9
	1 ($\rho V = 0.94$)	0	1	13
	2 ($\rho V = 0.86$)	0	4	11
	3 ($\rho V = 0.86$)	0	11	3
D	4 ($\rho V = 0.78$)	0	12	2
	5 ($\rho V = 0.73$)	1	9	4
	6 ($\rho V = 0.66$)	1	7	3
	7 ($\rho V = 0.66$)	1	9	2
	8 ($\rho V = 0.60$)	1	2	1

Table 1: Number of times that the 3 selected methods were the best

the second pattern (where all data are missing in one group for some cases) but most regression methods perform better under the first pattern.

The factor D reveals in which case the RV method is a highly useful technique: as expected RV yields superior criteria to the other methods when the ρV coefficient of the covariance matrix is high. The more ρV decreases, the more the optimal criteria spread over all methods. Two different covariance matrices may yield the same value of ρV (this is the case for $D = 2$ and 3 or for $D = 6$ and 7); in this case, the RV method is better for the data with a strong correlation structure in the group of variables without missing data.

The correlation structure of the data is an important factor: the mean methods perform better under a low correlation structure and a small sample size (as noted in Gleason and Staelin, 1975) and the regression methods need a good correlation structure to be optimal (as noted in Frane, 1976).

As reported by Beale and Little (1975), we observe among the other imputation techniques that the regression method based on maximum likelihood yields superior criteria, this was to be expected since the data were generated from a multivariate normal distribution.

A crucial concern is whether the normality (the simulations were based on multivariate normal data of dimension four) and the number of variables influence the performances of the RV method or not. Analysis of real data with many variables have been undertaken to that extent; we found no

significant changes with the conclusions of the simulations; more precisely, the RV method performs as well on real data as it does on simulated ones.

The analysis of the RV method reveals the following points:

a) the vector correlation between the two groups of variables should be sufficient for using RV method; we drove an empirical study based on the asymptotic distribution of $n \cdot RV$ (Cléroux and Ducharme, 1989) and on the simulation of 4000 samples with RV coefficient lying between 0.2 and 0.6. We found 0.4 as an empirical limit of RV coefficient: when the complete data have a similarity less than 0.4, the RV method should not be used. The need for a sufficiently strong correlation structure for using the regression methods is equivalent to the limit of 0.4 for using method RV.

b) if there are large differences between the variances of the variables, we have observed that the RV method has a low imputation quality on the variables with small variances. A solution to this problem may be to first standardise the data.

3 Conclusion

This article shows that the RV method is worth mentioning with respect to

a) its imputation principle which is completely different from the existing methods (use of the similarity between two clouds of points),

b) its use of this supplementary information in the presence of two natural groups of variables: the RV method gives imputations with a pertinent interpretation,

c) its data analysis approach requiring no hypotheses,

d) its good performances especially in certain contexts: high coefficient RV based on complete data, pattern where all data are missing in one group, and so on.

Résumé: Dans cet article, nous proposons une nouvelle méthode d'estimation des données manquantes de type analyse des données. Notre imputation utilise la structure multivariée des données en maximisant le coefficient de corrélation vectoriel échantillonal RV (Escoufier 1973). Dans un premier temps nous dévelopons la méthode basée sur le coefficient RV, puis nous étudions les performances de la méthode RV en la comparant sur des données réelles et simulées avec huit méthodes existantes.

Mots clés: Coefficient de corrélation vectoriel, données manquantes, méthode de Buck, méthodes d'imputation, méthode RV.

References

[1] Beale, E.M.L. and Little, R.J.A. (1975), "Missing Values in Multivariate Analysis", *J. Roy. Statist. Soc. B,* **37**, 129-145.

[2] Buck, S.F. (1960), "A Method of Estimation of Missing Values in Multivariate Data Suitable for use With an Electronic Computer", *J. Roy. Statist. Soc. B,* **22**, 302-306.

[3] Cléroux, R. and Ducharme, G.R. (1989), "Vector Correlation for Elliptical Distributions", *Commun. Statist. Theory Meth.*, **18**, 1441-1454.

[4] Crettaz von Roten, F. and Helbling, J.M. (1991), "Une estimation de données manquantes basée sur le coefficient *RV*", *Rev. Statist. Appliquée,* **39**, 47-57.

[5] Crettaz von Roten, F. (1992), *Les méthodes d'estimation des données manquantes vues comme minimisation de distances*, Compte rendu des XXIV^{es} Journées de Statistiques, Bruxelles.

[6] Crettaz von Roten, F. (1993), *Données manquantes en statistique multivariée: une nouvelle méthode basée sur le coefficient RV*, Thèse n°1111 de l'Ecole Polytechnique Fédérale de Lausanne.

[7] Dempster, A.P., Laird, N.M. and Rubin, D.B. (1977), "Maximum Likelihood From Incomplete Data via the EM Algorithm", *J. Roy. Statist. Soc. B,* **39**, 1-38.

[8] Escoufier, Y. (1973), "Le traitement des variables vectorielles", *Biometrics,* **29**, 751-760.

[9] Efron, B. (1993), "Missing Data, Imputation and the Bootstrap", *J. Amer. Statist. Assoc.,* **89**, 463-480.

[10] Ford, B.L. (1983), "An Overview of Hot-Deck Procedures" in *Incomplete Data in Sample Surveys, Vol II: Theory and Annotated Bibliography* (Madow, Olkin, and Rubin, Eds), Academic Press, New York.

[11] Frane, J.W. (1976), "Some Simple Procedures for Handling Missing Data in Multivariate Analysis", *Psychometrika,* **41**, 409-415.

[12] Gleason, T.L. and Staelin, R. (1975), "A Proposal for Handling Missing Data", *Psychometrika,* **40**, 229-252.

[13] Little, R.J.A. and Rubin, D.B. (1987), *Statistical Analysis With Missing Data*, John Wiley, New York.

[14] Meng, X.L. and Rubin, D.B. (1991), "Using the EM Algorithm to Obtain Asymptotic Variance-Covariance Matrices: the SEM Algorithm", *J. Amer. Statist. Assoc.,* **86**, 899-909.

[15] Rubin, D.B. (1976), "Inference and Missing Data", *Biometrika,* **63**, 581-592.

[16] Wilks, S.S. (1932), "Moments and Distribution of Estimates of Population Parameters From Fragmentary Samples", *Ann. Math. Stat.*, **3**, 163-195.

Student (1996), Vol.1, No.4, 241-250

Construction de Groupes Stratégiques

Olivier Furrer

Groupe de Gestion d'Entreprise, Université de Neuchâtel, Suisse

Résumé: Le but de l'étude présentée dans cet article est de mettre au point un appareil méthodologique permettant de construire des groupes stratégiques. Trois groupes stratégiques ont été définis *a priori* sur la base de la littérature: l'orientation-coûts, l'orientation-client et l'orientation-produit. Une analyse de variance à un facteur été utilisée pour sélectionner les variables stratégiques mesurant le mieux ces trois orientations. Ces dernières ont ensuite été représentées dans un plan à l'aide d'une analyse en composante principale. Différentes règles d'allocation permettant de classer des entreprises à l'intérieur de ces orientations sont enfin présentées.

Mots clés: Analyse de variance, analyse discriminante, analyse en composantes principales, groupes stratégiques.

1 Introduction

Cette étude vise à mettre en place un appareil méthodologique permettant d'opérationaliser une typologie des orientations stratégiques et de déterminer un système d'allocation des entreprises à l'intérieur des différentes catégories. Il s'agit d'une étape préliminaire à une recherche sur les stratégies de services des entreprises informatiques nécessitant la construction de groupes stratégiques (pour une revue de la littérature sur les groupes stratégiques, voir McGee et Thomas, 1986; Thomas et Venkatraman, 1988). Dans cette étude, nous avons testé un certain nombre de variables stratégiques susceptibles de classer les entreprises à l'intérieur de différents groupes stratégiques définis *a priori* sur la base de l'orientation stratégique des entreprises. Notre recherche consiste à déterminer quelles sont les variables stratégiques qui permettent le mieux de mesurer ces orientations, puis d'utiliser ces variables pour classer les entreprises à l'intérieur

des groupes.

L'objectif de cette recherche n'est pas de mettre en évidence l'existence ou la non existence de groupes à l'intérieur d'une industrie, mais d'avoir un outil qui permette d'étudier les stratégies de services des entreprises informatiques. C'est pourquoi, les groupes ont été définis *a priori* sur des bases théoriques plutôt que construits de manière empirique. Cette typologie doit permettre de vérifier certaines hypothèses sur le comportement stratégique des entreprises, ce qui est plus délicat lorsque les groupes sont formés sans base théorique stable. L'approche *a posteriori* permet la description de groupes stratégiques, mais rend difficile et subjective leur interprétation. Les caractéristiques des groupes identifiés par cette approche varient largement en fonction du nombre et du type de variables utilisées (pour une évaluation et une classification des différentes méthodes de formation des groupes stratégiques, voir Thomas et Venkatraman, 1988; Ketchen, Thomas et Snow, 1993). Sur la base de la littérature (Miles et Snow, 1978; Hofer et Schendel, 1978; Miller et Friesen, 1978; Porter, 1980; etc.) et des pratiques du secteur informatique, nous avons donc décrit trois orientations stratégiques: l'orientation-coûts, l'orientation-client et l'orientation-produit.

Définition 1.1: Une entreprise orientée-coûts cherche à dominer son secteur grâce à des coûts extrêmement bas. Il est important pour elle que les processus de production soient les plus efficaces possible. Une entreprise de ce type cherche à optimiser ses capacités de production et à réaliser des économies d'échelle. Elle se positionne sur des marchés stables avec une clientèle de masse sensible au prix. Les produits sont standardisés, la gamme est réduite afin d'augmenter les quantités et de diminuer les coûts de variantes. La R&D (recherche et dévelopement) est orientée vers le développement de nouveaux processus de production et vers l'amélioration des processus existants. Ce sont les ingénieurs de production et les financiers qui ont le plus d'influence au sein de l'entreprise. Le personnel est motivé au contrôle des coûts.

Définition 1.2: Une entreprise orientée-client cherche à dominer son secteur par la pleine satisfaction de ses clients. Pour ce faire, plus que des produits, elle offre des prestations globales qui prennent en compte l'ensemble des besoins des clients. Cette entreprise se concentre sur un type de clientèle dont elle connaît bien les besoins. Elle se positionne sur des segments de marché où la clientèle est exigeante, connaît bien les produits qu'elle achète et désire des services de haute qualité. Pour ce type de clientèle le service est plus important que le prix. Les produits comme les services sont personnalisés. L'entreprise cherche à conserver ses clients actuels plutôt que

d'en gagner de nouveaux. Elle cherche à établir des liens de partenariat avec ses clients. La R&D est orientée vers la résolution des problèmes des clients. Ce sont les hommes du marketing et du service à la clientèle qui sont le plus influents dans cette entreprise. Le personnel est formé et motivé pour le service au client.

Définition 1.3: Une entreprise orientée-produit cherche à dominer son secteur grâce à la qualité et à l'innovation de ses produits. Elle maîtrise les technologies de pointe, elle innove constamment et lance régulièrement de nouveaux produits sur le marché au risque de cannibaliser les anciens. Son avantage concurrentiel est de toujours être en avance sur ses concurrents. Elle se situe sur des segments de marché à forte croissance et où les clients sont des adopteurs précoces sensibles à la technologie et aux nouveautés autant qu'à la qualité des produits. La R&D est orientée vers la recherche pure et l'innovation technologique. Ce sont les chercheurs et ingénieurs de la R&D qui sont les plus influents dans ce type d'entreprise. Le personnel est formé et motivé à la qualité des produits.

Il s'agit de sélectionner les variables qui mesurent le mieux ces trois orientations stratégiques et qui permettent de les différencier. La littérature sur les groupes stratégiques est riche en variables stratégiques; Ketchen, Thomas et Snow (1993) en recensent 91. C'est parmi ces 91 variables que nous avons fait notre choix. De cette liste, nous avons retiré toutes les variables qui ne pouvaient être évaluées de manière perceptuelle ou mesurées sur une échelle de Likert. Nous avons ensuite retiré les variables qui n'étaient pas pertinentes pour notre recherche, ne correspondant pas au secteur informatique ou à l'étude des services ou qui étaient redondantes. Quelques variables ont ensuite été ajoutées pour équilibrer et compléter certaines catégories. Nous avons abouti à une liste de 37 variables (voir table 1).

Il est difficile de savoir *a priori* quelles sont, parmi ces variables, celles qui mesurent plus particulièrement l'une ou l'autre des orientations stratégiques. Une variable comme la *qualité des produits et services* peut être importante pour l'orientation-produit comme pour l'orientation-client. Il est donc nécessaire de faire une analyse qui tienne compte des liens existant entre les orientations, celles-ci n'étant pas totalement indépendantes, et qui tienne également compte des corrélations entre les variables. Notre analyse se déroule en deux étapes. Dans un premier temps, il s'agit de réduire le nombre de variables et de ne conserver que celles qui permettent le mieux de différencier les trois orientations stratégiques; puis dans un deuxième temps, de construire, à l'aide des variables retenues, un système d'allocation des entreprises à l'intérieur des différents groupes.

1.	Minimisation des coûts
2.	Capacités à faire des produits et services spéciaux
3.	Économie d'échelle dans la production
4.	Service à la clientèle
5.	Forte image de marque
6.	Service après-vente
7.	Développement de nouveaux produits et services
8.	Minimisation des stocks
9.	R&D centrée sur les processus de production
10.	R&D centrée sur l'innovation produit
11.	Concentration sur certains besoins de la clientèle
12.	Qualité des produits et services
13.	Différenciation des produits et services
14.	Concentration sur certains types de clientèle
15.	Produits et services pour des segments de marché à prix élevés
16.	Produits et services pour des segments de marché à bas prix
17.	Accès préférentiel aux matières premières
18.	Compétences technologiques supérieures à la moyenne du secteur
19.	Développement et amélioration des produits et services existants
20.	Innovation dans les processus de production
21.	Prix compétitifs
22.	Stricts contrôles de la qualité des produits et services
23.	Bonne réputation au sein de l'industrie
24.	Niveau des stocks élevé
25.	Différenciation par la communication marketing
26.	Concentration sur un marché géographique
27.	Personnel expérimenté et bien formé
28.	Contrôle des canaux de distribution
29.	Capacité financière élevée
30.	Innovation dans les techniques et méthodes marketing
31.	Extension des canaux de distribution
32.	Autofinancement
33.	Intégration verticale
34.	Large gamme de produits et services
35.	Gamme de produits et services étroite
36.	Dépenses de publicité et promotion supérieures à la moyenne du secteur
37.	Diversification

Table 1: Variables stratégiques utilisées

2 Méthodologie

Afin de mesurer l'importance des variables pour chacune des orientations stratégiques, un questionnaire a été envoyé à 42 experts académiques. Ces experts ont été choisis parmi les professeurs de management et de marketing des universités et hautes écoles suisses. Il leur a été demandé d'évaluer le degré d'importance de chacune des 37 variables stratégiques de la table 1 sur une échelle de Likert à 5 positions (1 = pas du tout important; 5 = très important) pour chacune des trois orientations (orientation-coûts, orientation-client et orientation-produit), celles-ci étant brièvement décrites.

Nous avons reçu en retour 25 questionnaires, soit un taux de réponse de 64.1%. Parmi ces 25 réponses, 3 ont été éliminées parce quelles étaient trop incomplètes pour être utiles. Les réponses comportant des valeurs man-

quantes en nombre limité ont été conservées. L'analyse factorielle utilisée nécessitant qu'il n'y ait pas de données manquantes, celles-ci ont été estimées par la valeur entière de l'échelle de mesure qui minimise la variance intra-groupe (Parreins, 1974). Le taux de réponses utiles est donc d'un peu plus de 56% (22 réponses). Pour un échantillon d'experts, ce taux est relativement faible. Cette faiblesse peut s'expliquer en partie par la longueur et par la complexité du questionnaire. Le nombre de réponses utiles est cependant suffisant pour les traitements statistiques utilisés.

2.1 Sélection des variables stratégiques

La description des orientations stratégiques (moyennes $\overline{x}_j$ et écarts-type s_j) de la table 2 permet d'évaluer quelles sont les variables stratégiques que les experts considèrent comme les plus importantes pour chacune des orientations et clarifie les relations entre variables et orientations stratégiques. Cette description ne permet cependant pas de déterminer les variables qui discriminent le mieux les orientations. Pour mesurer le pouvoir discriminant de chacune des variables stratégiques nous avons utilisé l'analyse de variance à un facteur, ce pouvoir discriminant étant mesuré par le rapport de corrélation η^2 et le rapport F de Fisher[1]. La table 2 donne le classement des variables selon leur pouvoir discriminant. Un F plus grand que 3.11 est significatif à 95%. Il n'y a pas de règle établie pour savoir quelles sont les variables qu'il faut conserver. Nous avons décidé de garder les variables dont le rapport de corrélation est supérieur à 50%. Il s'agit des 13 première variables stratégiques de la table 2. Ce sont ces variables que nous utiliserons pour l'allocation des entreprises à l'intérieur des groupes.

2.2 Représentation des orientations stratégiques

Afin de représenter graphiquement les relations entre ces 13 variables, nous avons effectué une analyse en composantes principales (ACP) . Nous n'avons pas effectué l'ACP directement sur les notes des réponses au questionnaire, mais sur des valeurs centrées et réduites par individu. Cette opération a été rendue nécessaire pas le fait qu'une note 5 (très important) sur l'échelle de Likert, n'a pas la même signification pour l'ensemble des experts intérogés. Le fait d'utiliser une valeur centrée et de réduite par individu permet de conserver le profil des réponses de chacun des individus sans tenir compte de la valeur absolue de la note. De cette ACP, nous avons retenu les deux premiers facteurs restituant près de 63% de la variance (facteur 1 = 48%, facteur 2 = 15%). Le premier facteur oppose

[1] Ces rapports sont calculés comme suit (avec ici $n = 66$ et $k = 3$):
$\eta^2 = \frac{\text{somme des carrés inter-classes}}{\text{somme des carrés totale}}$ et $F = \frac{\text{somme des carrés inter-classes}}{\text{somme des carrés intra-classes}} \cdot \frac{n-k}{k-1}$.

l'orientation-coûts aux deux autres orientations, il est à rapprocher de la dimension avantage concurrentiel de Porter (1980) qui oppose la domination par les coûts à la différenciation. Le deuxième oppose l'orientation-client à l'orientation-produit, et est à rapprocher de la différence que fait Miller (1987) entre deux types de différenciation: différenciation par l'innovation des produits et services et différenciation par le marketing.

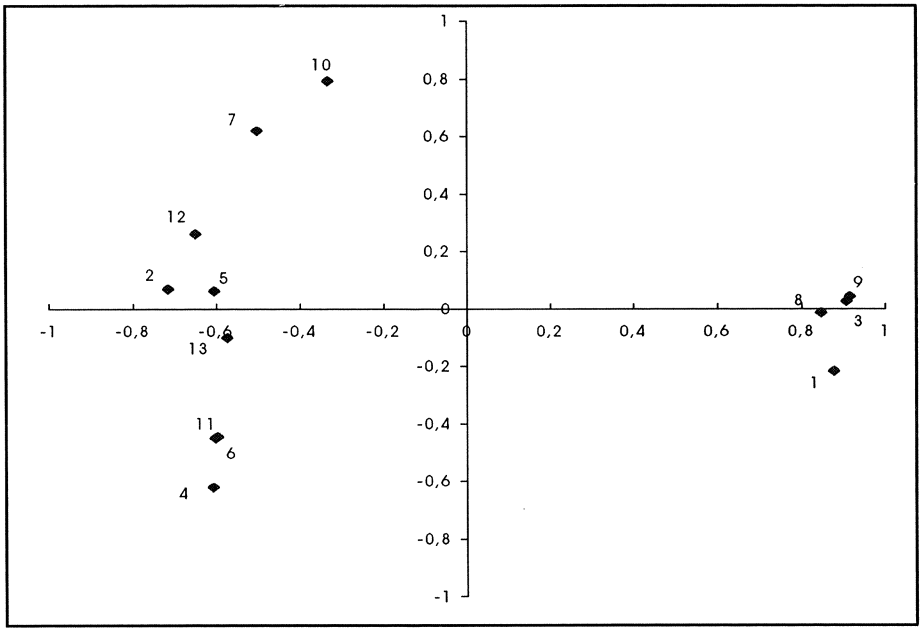


Figure 1: Positionnement des variables stratégiques dans le premier plan factoriel

La représentation des 13 variables dans le plan factoriel formé des deux premiers facteurs (voir figure 1) montre sur le premier axe une nette séparation entre les variables, *minimisation des coûts*, *économie d'échelle dans la production*, *minimisation des stocks* et *R&D centrée sur les processus de production*, qui représentent l'orientation-coûts, et les variables qui représentent les deux autres orientations. Sur le deuxième axe, on remarque l'opposition entre les variables *service à la clientèle*, *service après-vente* et *concentration sur certains besoins de la clientèle*, qui représentent l'orientation-client, et les variables *développement de nouveaux produits et services* et *R&D centrée sur l'innovation produit*, qui représentent l'orientation-produit. Les variables *capacité à faire des produits et services spéciaux*, *forte image de marque*, *qualité des produits et services* et *différenciation*

des produits et services sont importants pour l'orientation-client et pour l'orientation-produit.

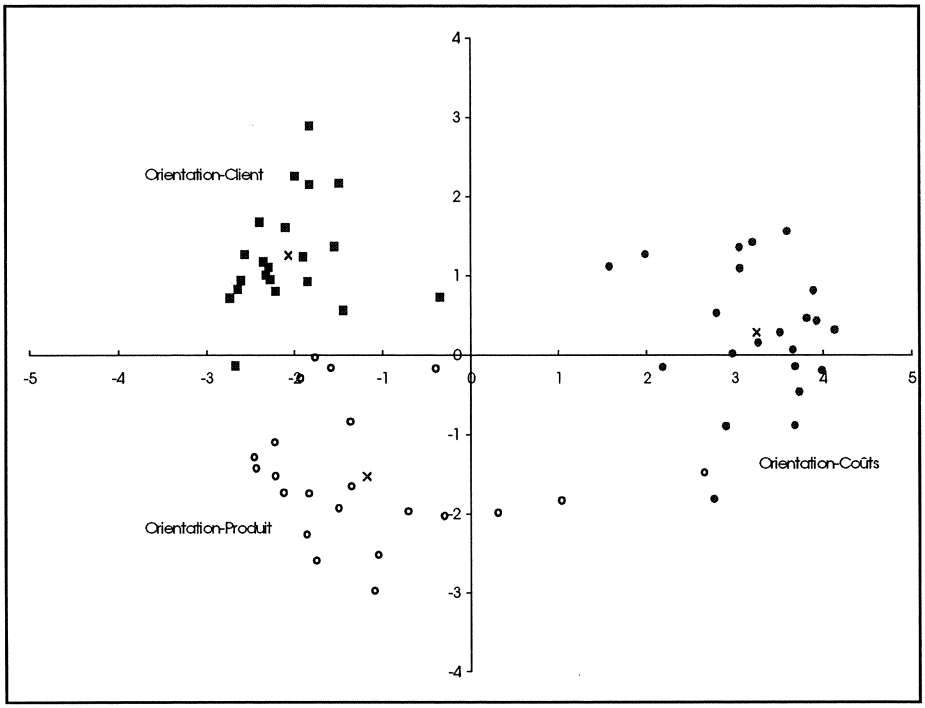


Figure 2: Représentation des orientations stratégiques sur le premier plan factoriel (les centres de gravité des orientations sont marqués par des croix)

De cette représentation graphique, on remarque que nous n'avons pas un axe factoriel par orientation stratégique, mais que les deux premiers axes suffisent pour représenter les trois orientations. Cela est dû au fait que ces orientations ne sont pas indépendantes. En plaçant les individus dans le premier plan factoriel issu de l'ACP, on identifie facilement les trois orientations stratégiques (voir figure 2). Chacune d'elles peut être décrite par son centre de gravité et sa dispersion autour de ce centre. Les coordonnées de ces centres sont mesurée par la moyenne des coordonnées du nuage des points de chacune des orientations sur les deux axes factoriels et la dispersion autour du centre est mesurée par la distance euclidienne moyenne. L'orientation-client est peu dispersée par rapport à l'orientation-coûts et l'orientation-produit (0.69 contre respectivement 0.95 et 1.26). La dispersion de l'orientation-produit est essentiellement due à quelques valeurs extrêmes. Si l'on calcule les distances euclidiennes entre

les centres de gravité, on remarque qu'elles sont nettement supérieures aux distances euclidiennes moyennes de chacune des orientations. La distance entre le centre de l'orientation-coûts et celui de l'orientation-client est de 5.41, entre celui de l'orientation-coûts et celui de l'orientation-produit de 4.78 et entre celui de l'orientation-client et celui de l'orientation-produit de 2.92. Cela indique une bonne séparation entre les groupes et une bonne homogénéité de chacun des groupes.

2.3 Allocation des entreprises dans les groupes stratégiques

Afin de classer les entreprises dans les différents groupes, on procéde comme suit: il s'agit de positionner une entreprise dans le premier plan factoriel, puis de l'allouer à une des trois orientations. Cette allocation peut se faire de plusieures manières: une première règle extrêmement simple consiste à allouer les entreprises se trouvant dans le deuxième cadrant à l'orientation-client; celles se trouvant dans le troisième à l'orientation-produit; et celles se trouvant dans le premier ou dans le quatrième cadrant à l'orientation-coût. Cette règle qui a l'avantage de la simplicité n'est pas cependant pas la plus précise. On peut par exemple décider d'allouer une entreprise à l'orientation dont le centre de gravité est le plus proche en distance euclidienne, ou alors en distance euclidienne pondérée par la dispersion de cette orientation.

3 Conclusion

Cette étude nous a permis d'opérationaliser une typologie des orientations stratégiques définies *a priori*. Tout en donnant un contenu aux trois orientations (orientation-coûts, orientation-client et orientation-produit), nous avons pu déterminer quelles étaient les variables qui pouvaient le mieux les mesurer. Ces trois orientations ont pu être représentées dans un plan d'une manière particulièrement claire. Nous avons ensuite déterminé un système d'allocation des entreprises à l'intérieur des trois orientations basé sur la représentation graphique de ces trois orientations. Cette étude est une étape préliminaire à une recherche sur les stratégies de services des entreprises informatiques. Il s'agira maintenant de tester la fiabilité de ce système d'allocation en l'utilisant pour classer les entreprises informatiques et d'étudier leur stratégies de services.

Abstract: The aim of the study presented in this article is to perfect a methodological instrument to build strategic groups. Three strategic groups were defined *a priori* on the basis of literature: the cost-orientation, the client-orientation and the product-orientation. A one-way analysis of

variance allowed to select the strategic variables that measure these three orientations at best. These three orientations were represeted graphically by means of a Principal Componant Analysis. Several rules were developed to allocate firms to the different groups.

Key words: Analysis of variance, discriminant analysis, principal component analysis, strategic groups.

Références

[1] Hofer, C.W. and Schendel, D. (1978), *Strategy Formulation: Analytical Concepts*, West Publishing Company, St. Paul.

[2] Ketchen, D.J., Thomas, J.B. and Snow, C.C., "Organizational Configurations and Performance: a Comparison of Theorical Approches", *Academy of Management Journal*, **36**, 1278-1313.

[3] McGee, J. and Thomas, H. (1986), "Strategic Groups: Theory, Research and Taxonomy", *Strategic Management Journal*, **7**, 141-160.

[4] Miles, R.E. and Snow, C.C., *Organizational Strategy, Structure, and Process*, McGraw-Hill, New York.

[5] Miller, D. (1978), "The Structural and Environmental Correlates of Business Strategy", *Strategic Management Journal*, **8**, 55-76.

[6] Miller, D. and Friesen, P.H. (1978), "Archetypes of Strategy Formulation", *Management Science,* **24**, 921-933.

[7] Parreins, G. (1974), *Techniques statistiques: moyens rationnels de choix et de décision*, Dunod, Paris.

[8] Porter, M.E. (1980), *Competitive Strategy: Techniques for Analysing Industries*, Free Press, New York.

[9] Thomas, H. and Venkatraman, N. (1988). "Research on Strategic Groups: Progress and Prognosis", *Journal of Management Studies*, **25**, 537-555.

var.	orient.-coûts		orient.-client		orient.-produit				
strat.	$\overline{x}_j$	s_j	$\overline{x}_j$	s_j	$\overline{x}_j$	s_j	η^2	F	p
1.	4.95	0.21	2.82	1.10	2.36	0.90	0.66	61.24	<0.0001
2.	4.64	0.73	2.64	0.85	2.73	0.98	0.64	54.99	<0.0001
3.	1.95	1.05	3.18	0.91	2.77	0.81	0.63	54.00	<0.0001
4.	4.77	0.53	2.32	0.72	2.64	1.18	0.63	53.68	<0.0001
5.	3.64	1.14	2.68	0.84	3.00	1.07	0.57	42.27	<0.0001
6.	3.32	1.36	2.36	0.95	2.95	1.05	0.55	38.93	<0.0001
7.	3.41	1.01	4.32	0.72	4.59	0.73	0.55	38.91	<0.0001
8.	4.14	0.89	2.45	1.10	3.55	0.74	0.55	37.98	<0.0001
9.	2.00	1.02	4.64	0.79	4.27	0.88	0.54	36.98	<0.0001
10.	4.64	0.90	3.00	0.98	3.55	1.14	0.53	35.36	<0.0001
11.	3.36	1.36	3.00	0.87	4.68	0.72	0.52	33.62	<0.0001
12.	4.77	0.43	2.73	0.70	3.00	1.23	0.51	32.78	<0.0001
13.	2.50	1.10	3.50	0.96	4.77	0.53	0.51	32.24	<0.0001
14.	4.55	1.18	3.36	0.95	2.91	1.02	0.48	22.61	<0.0001
15.	2.18	1.18	4.09	0.81	3.73	0.94	0.42	23.18	<0.0001
16.	4.32	1.04	2.77	1.15	2.50	0.86	0.39	20.15	<0.0001
17.	2.18	0.80	4.50	0.96	4.23	0.97	0.37	18.86	<0.0001
18.	3.36	1.14	3.50	1.01	3.27	0.94	0.35	16.48	<0.0001
19.	2.59	1.10	3.91	0.68	4.77	0.61	0.32	14.92	<0.0001
20.	2.95	1.17	3.23	0.97	2.95	0.95	0.32	14.91	<0.0001
21.	3.09	1.31	3.05	1.29	3.32	1.04	0.31	14.05	<0.0001
22.	2.73	1.03	4.95	0.21	3.55	0.67	0.28	12.23	<0.0001
23.	2.91	1.31	4.45	0.67	4.14	0.89	0.26	11.02	<0.0001
24.	2.55	1.14	3.59	0.96	2.59	0.91	0.24	10.01	0.0002
25.	3.27	1.28	4.14	0.64	3.32	1.25	0.22	8.94	0.0004
26.	2.95	1.17	4.86	0.35	3.32	0.89	0.19	7.55	0.0011
27.	2.36	1.40	4.86	0.35	3.41	1.01	0.19	7.35	0.0014
28.	2.32	0.89	4.36	0.66	3.68	0.99	0.17	6.68	0.0023
29.	3.00	1.15	4.32	0.65	3.36	1.29	0.14	4.97	0.0099
30.	2.73	1.08	4.86	0.35	3.91	0.81	0.12	4.32	0.0174
31.	3.18	0.96	4.73	0.46	4.59	0.59	0.11	4.02	0.0228
32.	3.00	1.48	2.77	1.27	2.55	1.01	0.11	3.98	0.0236
33.	2.27	1.20	2.50	1.10	2.45	1.18	0.02	0.71	0.4972
34.	3.73	1.03	3.91	0.97	2.86	1.04	0.02	0.51	0.6047
35.	3.55	1.18	3.18	1.18	2.59	1.01	0.01	0.32	0.7300
36.	3.27	1.03	4.41	0.91	3.82	1.01	0.01	0.27	0.7636
37.	2.95	0.95	3.95	0.79	4.00	0.76	0.01	0.24	0.7913

Table 2: Description et mesure du pouvoir discriminant des variables stratégiques

Student (1996), Vol.1, No.4, 251-252

From the Editor's Notebook

The Madness of Minmad

The *mean* of a data set of real numbers $E = \{x_1, ..., x_n\}$ is known to be the point μ that minimizes the function

$$f(\mu) = mean\left\{(x_1 - \mu)^2, ..., (x_n - \mu)^2\right\}.$$

This is why the mean is also called the minimizer of the "mean of the squared deviations". In a similar way, the *median* of E has the property to be the point μ that minimizes the function

$$g(\mu) = mean\{|x_1 - \mu|, ..., |x_n - \mu|\}.$$

Hence, the median is said to be the minimizer of the "mean of the absolute deviations".

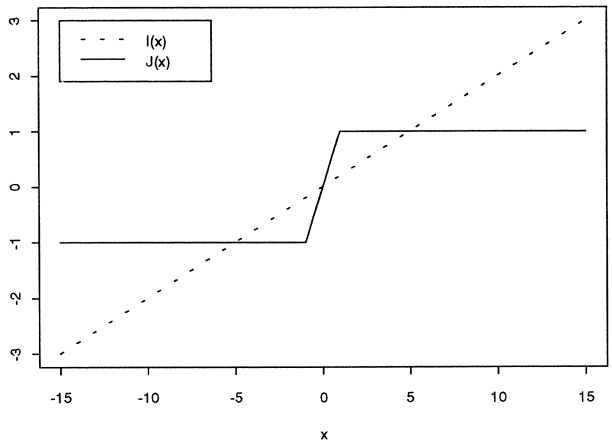
The mean and the median behave very differently with respect to extreme values called *outliers*. To see this, let us define

$$I(x) = mean(E \cup \{x\})$$

and

$$J(x) = median(E \cup \{x\}).$$

These functions measure the influence of a single observation on the mean, respectively on the median of a data set. If we look at their graphs, we see that while $I(x)$ goes to infinity when x becomes large, $J(x)$ remains constant. The figure at the top of page 252 illustrates the above fact with the data set $E = \{-2, -1, 1, 2\}$.



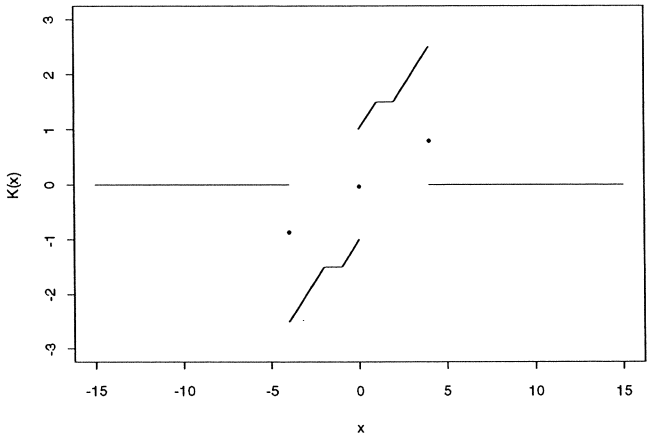
Let us now define the *minmad* of E to be the point that minimizes the function

$$h(\mu) = median\{|x_1 - \mu|, ..., |x_n - \mu|\},$$

that is to be the minimizer of the "median of the absolute deviations". The function

$$K(x) = minmad(E \cup \{x\})$$

has a quite strange behaviour compared to $I(x)$ and $J(x)$. Such a MAD behaviour is illustrated by the following figure with the same data set. Note that since $K(x)$ is not always uniquely defined, the mean of the different solutions of $minmad(E \cup \{x\})$ is plotted as $K(x)$. It appears that for moderate values called *inliers*, $K(x)$ is larger than for extreme values. Thus the minmad is more sensitive to inliers than to outliers!



Student (1996), Vol.1, No.4, 254-256

George A. Barnard Hopes for a Cheerful Consensus

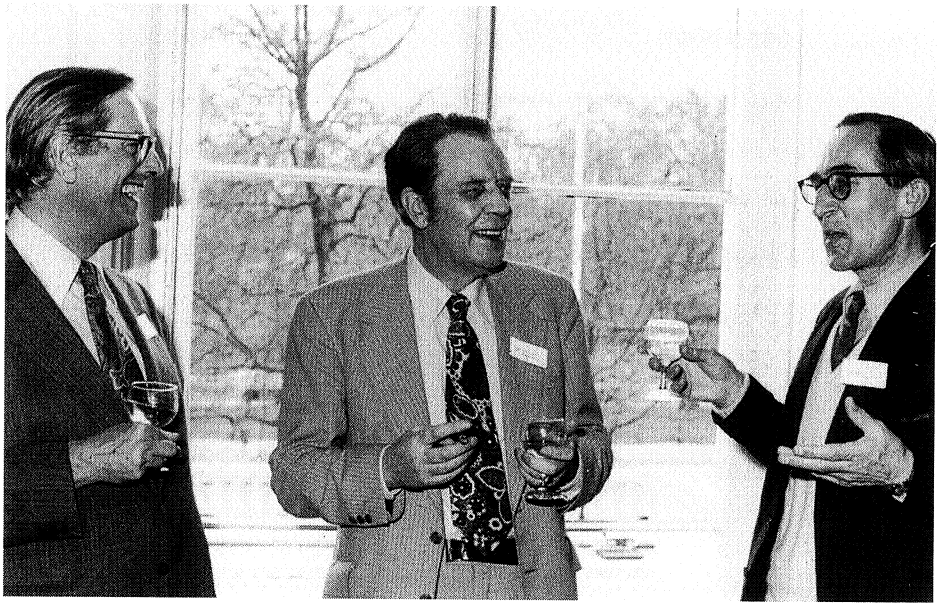
George A. Barnard was born on September 23, 1915, in Walthamstow, Essex, England. He received a B.A. in Mathematics from Cambridge University in 1936, and a D.Sc. from the University of London in 1965. He was a Scientific Officer in the Ministry of Supply Advisory Unit from 1942 to 1945. Professor Barnard became a faculty member in the Mathematics Department of the Imperial College from 1945 to 1966, Professor of Mathematics at the University of Essex from 1966 to 1975 and Professor of Statistics at the University of Waterloo from 1975 until his retirement in 1981. George Barnard has been awarded Gold Medals from the Royal Statistical Society and the Institute of Mathematics and Its Applications. In 1987, he was named an Honorary Fellow of the Institute of Statisticians.

In 1974, at the age of 80, Jerzy Neyman told to the 60 years old George Barnard that "Through disagreement clearly expressed we make progress." In 1995, now 80, George Barnard concludes his paper entitled "Neyman and

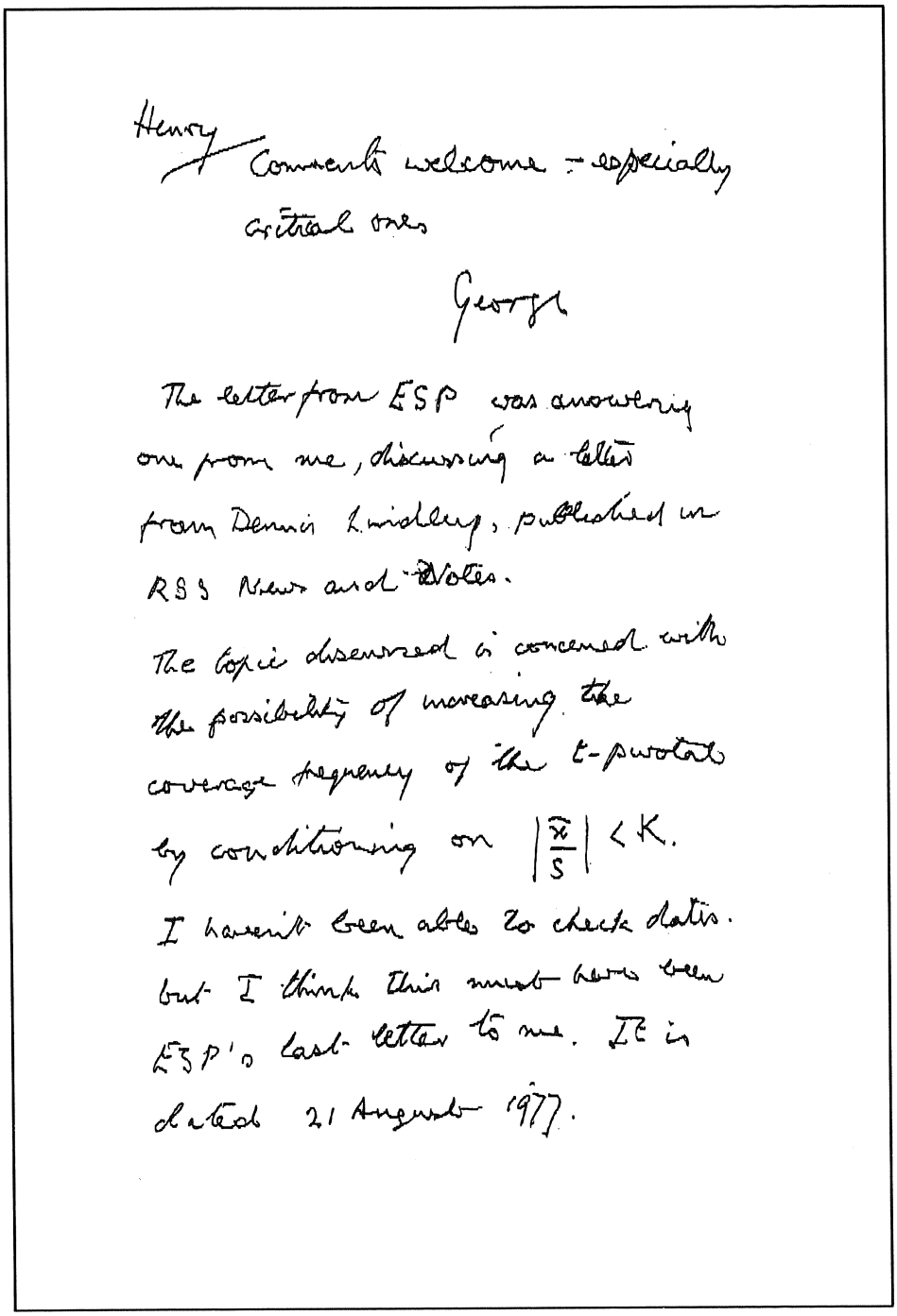
the Behrems-Fisher Problem - An Anecdote" in *Probability and Mathematical Statistics* (Vol.15, 67-71) with: "What Neyman said about differences clearly expressed is true, though clear expression is often hard to achieve."

This is why George Barnard believes that the need for statisticians to be familiar with all four main approaches to the foundations of statistical inference (the Fisherian, the Jeffreys, the Neyman-Pearsonian and the Bayesian), and with their relations to each other is extremely important. In his "Fragments of a Statistical Autobiography" that follow, he once again emphasizes that "In formal logic the advances made mainly in the 1930's have now led to a *cheerful consensus* in which workers in the field explore the consequences of adopting, or not adopting, the various foundational assumptions which are possible, without quarreling about whether these assumptions have or do not have universal validity. We may hope, and indeed expect, that a similar consensus will prevail among statisticians before too long".

When we asked Professor Barnard for a copy of one of his old lecture in handwriting, he replied: "You asked for a specimen of handwriting. I have therefore signed my contribution. But before I went to University in 1933 I bought a portable typewriter; and although I had to write my notes on the lectures when attending them, I have ever since had the habit of composing directly onto type."In spite of that, he was kind enough to search and send us finally a fragment of his handwriting.



George A. Barnard with Maurice Kendall and David Kendall



Henry
Comments welcome – especially
critical ones
George

The letter from ESP was answering
one from me, discussing a letter
from Dennis Lindley, published in
RSS News and Notes.
The topic discussed is concerned with
the possibility of increasing the
coverage frequency of the t-protocol
by conditioning on $\left|\frac{\bar{x}}{s}\right| < K$.
I haven't been able to check details.
but I think this must have been
ESP's last letter to me. It is
dated 21 August 1977.

A note to Henry Daniels about Egon S. Pearson's (ESP) last letter to him.

Student (1996), Vol.1, No.4, 257-268

Fragments of a Statistical Autobiography

George A. Barnard

Department of Mathematics, University of Essex, United Kingdom

Before 1900 we see many scientists of different fields developing and using techniques we now recognise as belonging to modern statistics. After 1900 we begin to see identifiable statisticians developing such techniques into a unified logic of experimental science that goes far beyond its components parts.

(Stephen Stigler, *The History of Statistics*)

Predicting the future is always a risky business, and my feeling, that the century's end will come to be seen as, in some sense, completing a major stage in the development of our subject, may well be seriously in error. But with the four main approaches more or less fully established - the Fisherian, the Jeffreys, the Neyman-Pearsonian, and the Bayesian - it would seem that future developments will consist in further applications of one or other of these approaches, rather than an entirely new basic theory. I am particularly hesitant in making such a guess because if a long life has taught me anything it is, that we are constantly deceived by thinking we have contemplated every possibility. Indeed my principal objection to those who think that the Bayesian approach can serve to deal with every type of statistical problem is, that such an approach requires that all measures of uncertainty must add up to 1.

Karl Pearson's Department of Statistics at University College London was for many years the only place where it was possible to graduate in the subject. The rate of development in the field accelerated markedly in the 1920's. 1925 saw the first publication of "Statistical Methods for Research Workers" and in the next 30 years our field developed explosively, with the first three of the above approaches achieving more or less complete development. It may be significant that the fourth approach is arguably the first, rather than the last, going back as it does to 1763. Any professional statistician now needs to be familiar with all four approaches, and their relations

to each other, so that an appropriate choice of approach, or combination of approaches, can be made to problems as they arise. And excellent courses in statistics now available in many universities throughout the world reflect this fact. I present some fragments of statistical autobiography which may serve to indicate what it could be like to be involved in these developments during that period of explosive growth.

I began to think of myself as a statistician in 1941, nearly the middle of the period of rapid expansion. Before that I was primarily interested in the foundations of mathematics, though I had read Frank Ramsey's account of the personal approach in his book "Foundations of Mathematics and other Essays". And I carried out an early political poll in 1932, trying to study the variation in the strength of association between the political opinions of my schoolmates and those of the newspaper they read, as it was affected by whether they first read the sports section or the political section of their newspaper.

I was born in 1915 and bred in Walthamstow, a London Borough with an excellent grammar school and an outstanding public library which stocked such books as Karl Pearson's "Grammar of Science", Russell's "Introduction to Mathematical Philosophy", Wittgenstein's "Tractatus", and Frank Ramsey's work mentioned above. Having been fascinated by such books, one of the first things I did on arriving in Cambridge in 1933 was to join the long queue of people hoping to attend Wittgenstein's "Conversation Class". His method of reducing his audience to manageable proportions was to make clear that classes would run for 3 hours 3 times a week, and anyone who missed one class would not be admitted to any later classes. I managed to stay the course and was persuaded by Wittgenstein that philosophical problems such as the "problem" of induction typically arose from asking questions which appear to make sense, but which on analysis can be seen to be empty.

In my second and third undergraduate years I concentrated on the foundations of mathematics. Gödel had recently proved his famous theorem on the incompleteness of any axiomatic account of arithmetic and many cases of the decision problem of mathematical logic were being solved. Then in 1936 Church and Turing proved, in their different ways, the impossibility of a general solution of the decision problem. I vividly remember a visit to my tutor Max Newman in the summer of 1936. He talked about a very puzzling paper he had been asked to referee for the London Mathematical Society. It was by Alan Turing and was entitled "On Computable Numbers, with an Application to the Decision Problem". It is not surprising that Newman found it difficult, since it is arguably the most original mathematical paper ever published.

Towards the end of my fourth year Max Newman arranged for me to follow Turing to Princeton, where Turing had gone to relate his result with the equivalent result which had been obtained by Alonzo Church, using the lambda calculus. This left the status of the axiom of choice as the main outstanding problem in the field. But Gödel came to Princeton while I was there to give the lectures in which he proved that, if the axioms of set theory without the axiom of choice are consistent, they remain so when this axiom is added to them. The Church-Turing and the Gödel results were the last of the series of ground-breaking advances in mathematical logic made in the 1930's. Church suggested that I should try to use the recent results in logic to deal with unsolved problems in other branches of mathematics. But it was to be nearly ten years before a major advance in this direction was made in Abraham Robinson's 1949 doctoral thesis on the Metamathematics of Algebra. I had forgotten that I was one of the examiners for this thesis until I was reminded by Dauben's recent biography of the inventor of non-standard analysis. My own clearest recollection of Robinson had been of walking with him, as fast as dignity permitted, from an informal logical seminar to the nearest London Underground Station under a shower of shrapnel from anti-aircraft guns.

When I returned to England in 1939 it was clear that World War II was about to begin. What could a mathematical logician do to assist in the war effort? Some of us thought this question was solved by the appearance, soon after the declaration of war, of an advertisement for graduate mathematicians to serve as Instructor Lieutenants in the Royal Navy. Their duties were said to be to teach mathematics to the midshipmen, and to assist the ship's commander with tactics in battle. Along with many of my Cambridge contemporaries I applied and was interviewed by an Instructor Rear-Admiral who asked me about the mathematical topics I had covered. When I mentioned topology he said "That's a kind of n-dimensional geometry, isn't it ?" Then he passed me on to an Instructor Commander who made it clear that the principal function of the Instructor Lieutenant was to settle ward-room arguments concerning the appropriate odds in card games. After taking a battery of physical and perceptual tests I heard no more, from the Navy. As the war developed I came across most of my contemporaries who had applied and found that their experience had exactly matched my own.

When I was required to register for army service mathematicians had been informally told to state their occupation as "physicist". "Mathematician" was not then recognised as a possible occupation for anyone other than a teacher. Physicists were needed for developments in radar and were exempt from direct military service. I myself was recruited to the Plessey

Company by their chief designer who had realised that a roving mathematician could be useful to them. My first job involved a gear-cutting machine with which they had found difficulty in finding a gear-sequence which would convert British measurements to metric ones. There are 2.54 centimetres in 1.00 inches, and the prime factors of 254 are 2 and 127. The absence of any gear wheel with 127 teeth was the cause of the difficulty. Among projects with which I later became involved was a high-altitude aircraft fuel system. Although this had been shown experimentally to work in aircraft trials, the Air Force hydrodynamicists refused to accept it because they had "proved mathematically" that it could not work. It turned out that the hydrodynamicists had assumed that the flow in the pipes would be turbulent and so the pressure drop would be proportional to the square of the velocity. In fact the flow was laminar and the pressure drop was proportional to the 1.75 power of the velocity. This experience brought home to me the limitations of a purely mathematical approach to real problems. It also served to confirm Wittgenstein's thesis, that practical activity is necessary if one wishes to gain the intuitions needed in solving problems of applied mathematics.

Over time it became clear that the flow of problems with which I could really assist in the Plessey Company could not continue in its initial volume. I had read Shewhart's classic "Economic Control of Manufactured Product" introducing his statistical control chart. So when I was approached by my friend Frank Smithies to join him and John R. Womersley in the London Headquarters of the Ministry of Supply, to help set up a Statistical Advisory Service I was happy to join them. Among the first problems put to me came from the section of the Ministry which dealt with unexploded bombs. I was told how often the routine used to defuse these had been used, thus far successfully, and was asked for the probability that it would be successful the next time it was used. I am not sure how much confidence the men who used the routine had in my answer.

My section of SR17 was called the Research Section partly because we were required to participate in research on new weapons and partly because we were required to develop new statistical methods. The most important of our methodological innovations came about because all our work was urgent, and our experimental resources were very limited. Examples of the consequences of this pressure are the Plackett-Burman experimental designs, sequential tests and inspection procedures, a Bayesian approach to batch sampling inspection problems, and new methods of sampling inspection for continuous production. Most of these innovations were, of course, not published until the war had ended. One problem referred to us late in the war arose from the pattern of explosions formed by the German V1

bombs falling on London. Until a specimen had been captured intact it was not clear what mechanism made them explode, and the data appeared to show a tendency to cluster near the electrified railway lines in the south of London. A Monte Carlo study showed that such patterns as were observed would frequently arise by chance and they could not be taken as significant evidence that the mechanism was electrical in character.

Towards the end of the war in Europe it was clear that the research and development of new weapons could not affect the war's final outcome and the pressure for statistical work eased. So I was glad to be invited to lecture part time at Imperial College, London. At about the same time as I made my move, John Womersley began arranging to set up the Mathematics Division of the National Physics Laboratory, one of the first objects of which was to enable Turing to build a working version of the general purpose computer which he had imagined in his 1936 paper.

In my first full-time year at Imperial College I was blessed with a brilliant group of students to whom I lectured on computing and statistics. Sidney Michaelson went on with Keith Tocher to build one of the earliest programmable computers with interchangeable subassemblies; and after Tony Brooker had gone to Manchester to work with Turing he developed Mercury Autocode, one of the earliest high-level computer languages. The outstanding statistical member of the group was Joseph Gani.

The work of the four sections of SR17 led to a greatly increased demand for statisticians in industry, and I was lucky soon after the war to get to know George Box who had returned from statistical work in the Army to take the degree course given at University College London. George proceeded to work for Imperial Chemical Industries in their dyestuffs division, next to where Owen Davies, an earlier student at University College, had been working for some years in the pharmaceutical division. ICI's recognition of the value of the work they did led the Company to set up a scheme under which mathematics graduates could choose to be employed by the Company and spend their first year taking the course for the Imperial College Diploma in Statistics. Winston Churchill's wartime recognition of the fact that statistics is not a branch of economics led to the establishment of the Statistician Grade in the Civil Service and this brought about further expansion of the demand for specifically trained statisticians. Cambridge University set up its own Diploma Course soon after the war, and before long many other Universities followed suit. Nowadays most Universities possess departments teaching statistics, and attempts are being made to teach statistics in elementary schools. These last attempts are justified by the threat to good democratic government arising from mass media, driven by competition to sensationalise extremely rare unimportant events.

At Imperial College I was able to devote time to the foundations of statistical theory. I had been introduced to Fisher in 1933 by W.L. Stevens who helped me analyse the poll referred to above. When I told Fisher of the difficulties I had in trying to find satisfactory books on mathematical statistics, he reached for a copy of "Statistical Methods for Research Workers" and told me that I would find in it a great many unproven assertions; but as a mathematician I should be able to prove these for myself. When I had done that, he said, I could reckon myself qualified in statistics. The next time I met Fisher was more than twenty years later, when he appointed me as one of his vice-Presidents of the Royal Statistical Society. I was then able to tell him I had just completed the task he had set me.

My long correspondence with Fisher began near the end of the war, when, taking for granted the Neyman-Pearson approach I recognised that with small samples Fisher's "exact" test for 2×2 tables rejected the hypothesis of independence with a frequency in repeated applications far smaller than the stated level of 5%. I outlined in a letter to "Nature" a "CSM" test, the initials of which referred not only to the Convexity, Symmetry, and Maximality properties of the test, but also to the Company Sergeant Major of my Home Guard unit, with whom I was having frequent "discussion". In my letter I claimed that my test was "more powerful than Fisher's". Fisher countered with a critical letter and I replied suggesting that Fisher's countered with a critical letter and I replied suggesting that Fisher's test differed from mine in its sphere of application. We continued with private correspondence, now mostly published, in which I came to see that the Neyman-Pearson theory of tests is essentially a theory of experimental design, involving only probabilities calculable in advance of obtaining the data. Querying the Neyman-Pearson reference set in which the sample size is held fixed, Fisher put to me the question: what if one of the 10 plants you had planted to check for colour was trodden on by someone running for a bus? I published in Biometrika a paper in which I attempted to allow for the unknown probabilities of such accidents. But eventually I came to see that our inferences from experiments must be based on what we know *after* the experiment has been performed. I became persuaded that Fisher's concept of likelihood was better suited than "confidence" to describe inferences from experimental results. The numerical measure of odds for various possible *outcomes* of the experiment before it was performed I called "forward odds", or, in their logarithmic form, "f-lods". The numerical measure for alternative possible *hypotheses* after the outcome was known was "backward odds", or b-lods. The word "lods" is still in use by geneticists, but not by anyone else so far as I know.

The work we had done in SR17 on sequential tests was limited to spe-

cific problems in the design of trials and in sampling inspection. It was put on record, along with some indications of possible extensions, in a letter to "Nature" in 1945. We learned of the general theory developed by Abraham Wald from a restricted report which arrived from the U.S.A. Wald developed this work into his book on the subject, and in 1947 I was invited to review it. In the review I noted that the difference between a "non-sequential" test and a Wald probability ratio test lay in the reference set with which the result was considered. The former involved repetitions in which the sample size was held constant, while the latter referred to repetitions in which the likelihood ratio was held constant. Both the sample size and the likelihood ratio for any pair of simple hypotheses would be known when the experiment had been performed, and I could find no compelling reason for insisting on fixing the sample size rather than the likelihood ratio. I therefore suggested that the "odds of error" in accepting H_0 or the alternative H_1, measured directly by the likelihood ratio, gave a better account of what was in the data than did the tail areas relative to constant sample sizes. Jimmie Savage later told me that the shock he felt in reading such an heretical suggestion was what later led him towards his Bayesian approach.

In my 1949 paper "Statistical Inference" I developed a theory of uncertain inference in which "probability" and "likelihood" appeared as concepts dual to each other, the one applying to "atomic" - that is, fully specified - results yet to be observed, the other to atomic or "simple" hypotheses after the results are known. The duality is not complete because any experiment with atomic results A, B, C, D, ... etc. logically implies the possibility of a less precise experiment with atomic results (A or B), (C or D), ... etc. But the disjunction of two simple hypotheses is , in general, not simple. In the introduction to the paper I explained that reflection on the 2×2 table problem had led me to see that Fisher was right and I was wrong. This led Fisher to remark to my former student Harold Ruben, then working in his laboratory, that "Barnard is the only statistician who has ever admitted that he was wrong". I have often wondered whether that was why, as President of the Royal Statistical Society, he appointed me one of his four vice-Presidents.

At this point I should say that after 1933 my first meeting with an established mathematical statistician was with Maurice Bartelett who had been invited to SR17 by Womersley to comment on our work on 2×2 tables. He was then, and still continues to be, most helpful and encouraging. And in spite of expressed disagreements, in almost all my dealing with Fisher, Neyman, and E.S. Pearson I was treated with the greatest kindness. Pearson was always ready to accept the possibility of different points of view

- he used to say that his theory was just describing his own way of getting his thinking "into gear" with practice. Both Neyman and Fisher were more self-assertive, Fisher deriving his confidence from his great success in natural science, Neyman from his powerful Polish mathematical background. My only serious confrontation with any of them was with Fisher when, at an ISI party in Brussels, our conversation turned to what, exactly, Thomas Bayes had said in his 1763 paper. When I did not accept Fisher's account he stormed at me, saying I should never say such things because they could cause great damage. I replied that I thought my view was correct, but I would confirm this by consulting Bayes's original paper in the National Library of Belgium the very next morning. Next morning I discovered that since Belgium had come into existence as an independent state only in 1830, their run of the Royal Society's Philosophical Transactions did not go back to 1763. I had to wait till I got back home to check with Deming's edition of Bayes's original paper. I then found that, as usual, Fisher had been right. Statisticians have in some quarters an unjustified reputation for constant quarrelling. This comes more from their having very interesting questions to discuss than from widespread belligerence. I should also add to these personal recollections that in SR17 I had the help of Peter Armitage, Peter Burman, Harold Godwin, Dennis Lindley and Robin Plackett - all of them fresh from mathematics at Cambridge. I guess others will envy my good fortune!

I first met Neyman and de Finetti in Paris in 1950, at a conference on the history and philosophy of science. The statistical part of the programme was organised by Maurice Frechet whom I had met soon after the liberation of Paris, when in the bitter winter of 1944-45 I flew to Paris with Szolem Mandelbrojt to give some lectures outlining some of the theoretical and practical developments which we had made during the war. As a guest of the French and a civilian I was not eligible for access to the British Army quarters, but was accommodated in the Hotel Bristol, where I "enjoyed" the height of luxury, which consisted of a thermos flask of hot water which served to take part of the icy chill from my bed. Then, next morning, the same water enabled me to shave with liquid water although the piped water system in my room was frozen solid. After a week of such conditions the British Government set up a hotel for other civilians who were needed to work in Paris. For the remainder of my stay I had the whole hotel and its food to myself. Sadly I could not share this food with my French acquaintances; but when I went with Mandelbrojt among the French mathematicians we were fed most royally by all of them from their very meagre rations.

At the 1949 conference Neyman expounded his view, that there is no

such thing as statistical *inference*, only "inductive behaviour", while de Finetti expounded his views on personal probability. I had been asked by Frechet to answer the question: can there be a theory of statistics independent of the calculus of probability? I argued that there could be many such theories. At that time the group of French mathematicians who published under the name of N. Bourbaki was very influential, and I proposed that statistics could be "Bourbachified" by regarding the statistical version of any mathematical structure S as obtained by taking the product structure $S \times S \times S...$, and then defining the statistical map of such a structure back into S itself as statistical when it was invariant under the symmetric group of transformations. I further pointed out that most of the theory of least squares could be regarded as a theory of approximations using the squared norm. Finally another non-probabilistic account of statistics could be based on the distinction first formally clarified by Alonzo Church between an undetermined value, $F(x)$, of a function F and the function F itself. Church denoted the latter by $\lambda x.F(x)$. In this approach if x denotes a free variable, and a denotes a constant, the observation that the observable F takes the value a can be written $F(x) = a$. Later, in my 1949 paper "Statistical Inference" read to the Royal Statistical Society, I expressed the fact that the "=" sign here does not denote identity by using $X \circ\!\!= x$ to denote that the observable X had been observed to take the value x. More recently I have written $x = x_0$ for the same thing. This distinction between an observable and its observed value is a major source of confusion in theoretical statistics. For example it has led one distinguished statistician to claim a logical flaw in the theory of conditional inference. He points to a cases where the distributions of X and of $Y = 1/X$ belong to the same location-scale family, the (complex) parameter value θ for X corresponding to the parameter value $1/\theta$ for $1/X$. He then shows that given a sample of observations x_i, $i = 1, 2..., n$, of X, the likelihood functions for θ based on the joint conditional distribution of $(\overline{x}, s_x)$ (with the usual notation) does not agree with the likelihood function for $1/\theta$ based on the conditional distribution of $(\overline{y}, s_y)$. The fact is, that the statements $x_i = x_{i0}$ and $y_i = y_{i0}$ can never be logically equivalent because the right hand side in the first expression denotes a small interval for x, while the right hand side in the second expression denotes a small interval for y. But in converting x to $y = 1/x$ small intervals are transformed into large intervals, and vice versa, unless both x and $1/x$ are near to 1.

"Statistical Inference" was the second paper I read to the Royal Statistical Society. The first, on "Sequential Tests in Industrial Statistics", described the work on this subject which our team had done during the war. It introduced the concept of the Process Curve, in effect a partially

observable Bayesian prior distribution which needs to be taken into account in assessing any proposed sampling inspection scheme. Several of my students continued this line of work later, examining various ways of describing process curves; but it was Francis Anscombe, who had for a period worked in SR17, who first perceived that special attention needs to be paid to the behaviour of the process curve in the neighbourhood of the "break even" quality - the point at which the costs involved in acceptance and refection are evenly balanced. The corresponding logical point arising with Bayesian approaches to testing hypotheses is not always given the attention it deserves.

I spent the decade of the 1950's preaching the gospel of likelihood. Computers had made it easy to plot likelihood functions with two parameters, and nowadays, with appropriate software, using colour we can deal with up to four. Such plots enable us to take advantage of serendipity - the occurrence of happy accidents. Results in standard likelihood theory will not lead us far astray provided the likelihood plots from the data we actually have are near enough to quadrics. But they need careful examination when this is not so. The first time this insight into likelihood plots was used in practice was with some data arising from a study on the rate of a chemical reaction. Chemical considerations suggested that the reaction was a double stage one; but the maximum likelihood estimate of the parameters corresponded to a single stage. On plotting the likelihood function it appeared that this had a ridge with a maximum near the single stage line, but there was a secondary peak near the double stage line. Further experimentation confirmed that the reaction was indeed double stage. this work was confidential, and the first time likelihood plots were used in public was in a consumer report on life tests of ladies' nylon stockings.

In 1956 Fisher published his "Statistical Methods and Scientific Inference". In it he vigorously defends his concept of "fiducial probability" as arising when the structure of the problem gives rise to a "fully efficient pivotal" - a function $p(t, \theta)$ of a statistic t and a parameter θ whose density $\varphi(p)$ is known, and which in some sense contains all the information about θ. For every θ_0 the mapping from $\{p\}$ to $\{t\}$ given by inverting the map $p_0 \rightarrow p(t, \theta_0)$ must be $1-1$, and the same must be true, for every t_0, of the map from $\{p\}$ to $\{\theta\}$ given by $p_0 \rightarrow p(t_0, \theta)$. We can obtain the fiducial distribution for the parameter θ by projecting the distribution $\varphi(p)$ onto $\{\theta\}$ by the inverse of mapping from $\{\theta\}$ to $\{p\}$, for given $t = t_0$. The sharp controversy which followed Fisher's publication led David Sprott and me to take time off from preaching likelihood in order to devote time to consider the many objections that had been raised to Fisher's argument. To use a phrase of which Fisher was fond, especially when he had said something

that was not quite right, "it now appears that" pure likelihood theory requires an "exactness assumption" that, given the values of the parameters θ, the distribution of the observables x is precisely determined. In real life, of course, we follow Peter Huber in his robustness studies, but a strict interpretation of the *logic* of any of the four approaches requires that an "exactness assumption" holds. When, as often, such an assumption cannot be made, it is sometimes possible to express what we know of the connection between x and θ by means of a set of "basic pivotals" $p_i(x_i, \theta)$ whose joint distribution is *approximately* known. In many cases we can find a function of the basic pivotals with the properties of $p(t, \theta)$ specified above. We can then derive, as above, an approximate "pivotal", or "fiducial", or "inverse" distribution for θ which can serve to express in ordinary language what we know about θ.

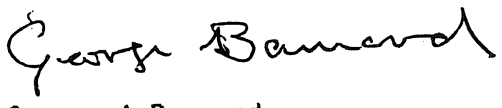
Fisher said that the main purpose of his 1925 "Statistical Methods" was to encourage use of Student's t, whose "exact" distribution Student had conjectured and Fisher had proved. In 1915, by a remarkable tour de force, Fisher derived the exact distribution of the correlation coefficient r in samples of n from a bivariate population known to be exactly normal; and he followed this up with further exact results for multivariate cases. But the fact is, that we never know, in practice, that our distributions are exactly normal. So "exactness" cannot be rigorously insisted upon. For example we often know that x_i, $i = 1, 2, ..., n$ are independent and approximately normal, and that, for given x_i, y_i may be taken as having a fixed approximately normal distribution. When this is so we can deduce that all that we know about the correlation coefficient ρ is contained in the pivotal

$$p(r, \rho) = \frac{1}{2} ln(\frac{1+r}{1-r}) - \frac{1}{2} ln\left(\frac{1+\rho}{1-\rho}\right),$$

(with the usual notation) which is approximately $N(0, 1/(n-3))$. The approximate pivotal distribution for ρ derived from the observation $r = r_0$ will then express what we know about ρ better than any "exact" distribution such as that suggested in Fisher's 1931 paper where he introduced his fiducial argument. Of course Fisher knew that our knowledge of the shapes of real distributions is only approximate; but he thought it would confuse the natural scientists to whom he addressed most of his writings if he made too much of this. Now that the understanding of statistics by experimental scientists is much better than it was in 1925, his hesitations on this point are no longer justified.

One reason why this autobiographical fragment devotes space to logic as well as to statistics is, that subject to a time shift of about 50 years, their historical developments are to a considerable extent parallel. Cantor's

account of set theory and the foundations of classical analysis used the axiom of choice in proving the basic theorems. These "proofs", and his other "proof" using the diagonal argument, that the continuum contains a transfinite "number" of points, were both called into question. In the first third of the 20th century this led to controversy, especially between David Hilbert and L.E.J. Brouwer, which was at times every bit as sharp as any of the controversies between Fisher and Neyman. In formal logic the advances made mainly in the 1930's have now led to a "cheerful consensus" in which workers in the field explore the consequences of adopting, or not adopting, the various foundational assumptions which are possible, without quarreling about whether these assumptions have or do not have universal validity. We may hope, and indeed expect, that a similar consensus will prevail among statisticians before too long.



George A Barnard

Data & Statistics in *Student (1996), Vol.1, No.4, 271-288*

Data Analysis: Strategies and Tactics

A.S.C. Ehrenberg

South Bank University, London, United Kingdom

Abstract: Seven precepts for data analysis are given and two illustrations.

Key words: Data analysis, data reduction.

1 Introduction

Having successfully analysed data for many years, I have consistently found our classical statistical techniques to be unhelpful and even counter-productive. The textbook procedures have seldom allowed me to understand or even *describe* the data. In particular, the output (e.g. correlation coefficients) could seldom be usefully compared with other similar studies. Nor could the results be readily communicated to others.

I therefore briefly set out seven do's and don'ts for successful data analysis, as I see it. I also discuss two numerical illustrations, using data sets published in this journal.

2 Seven precepts

2.1 Understand your data before you model it

With new data, it helps to establish the main descriptive patterns in the data before one tries to model them. Use simple summary measures before complex analysis techniques.

For a two-way summary table, for instance, I first calculate and look at the row and column averages, and summarise (in words) what, if anything, I see. This is before any formal analysis of variance, say, or other significance-testing or "modelling" procedure.

Again, with a new correlation matrix or contingency table, I first look for clusters of "high" or "low" values, perhaps reordering the rows and columns with a spreadsheet to help bring out any patterns visually. In this I would calculate a few summary sub-averages even "by hand" (which can be quicker for an informal "look-see"): *these* correlations are all about 0.8, and *those* are all roughly 0.2.

This is much less opaque and no more subjective than choosing

(i) some more or less arbitrary version of factor, correspondence, or cluster analysis,

(ii) with some more or less arbitrary stopping rule.

(iii) and then somehow verbally "labelling" the factors, etc.

2.2 Make the data user-friendly

To see patterns in the data, I use "Data Reduction" procedures (see *Further Reading*) such as

- drastic rounding, to facilitate mental arithmetic and help my short-term memory,
- order rows and columns by some measure of size (as one would on a graph),
- use row and column averages and sub-averages (to provide a potential summary as well as a visual and/or mental focus for scanning the full data).

Some of this will be illustrated by the examples in Section 3. My aim is visually helpful layout (i.e. "graphic tables"). This is to help me as analyst to see the data patterns in the first place, and then to communicate such patterns to others.

I occasionally want to plot new-new data initially to see its shape and outliers. Later I would occasionally also use graphics to communicate broad qualitative patterns to others, once I myself know what the patterns are. *But I would never leave the data merely as graphs.* They give no quantitative summaries. And different graphs are difficult to compare quantitatively.

2.3 Prior knowledge

Much data analysis is (or should be) more or less repetitive: I or other workers in the area will nearly always have analysed somewhat similar data before, and/or will also analyse other similar data subsequently. New-new data can strictly arise for the very first time at most once!

In any analysis there is therefore almost always prior knowledge of previous results, such as the clusters found in an earlier correlation matrix, or what row or column differences were larger than others (and "significant"),

and by how much. In developing a model I would typically see whether the equation $y \simeq 5x + 12$, say, which had held for previous data holds again (or something like it), and within something like the previous limits of scatter. That is instead of estimating a *new* (and different) best-fit regression-type equation every time, with no hope of developing any generalisable results.

Before most analyses I therefore have some positive prior hypotheses, or at least conjectures, of what the new data might be like. I organise my analysis of the new data according to these prior empirical results. The new results will then also highlight any discrepancies from my priors, major or minor.

2.4 Generalisable results and MSoD

Scientific results have to be generalisable. They have to be known to hold for a variety of different circumstances. The aim here is "significant sameness". If the generalisation is successful, the result will then be routinely predictable within the range of conditions that has been covered. If not, the result will probably not be usable *at all* in future - just an isolated finding.

This requires looking across many sets of data (MSoD) to see how far the same result does hold again (within its stated limits of approximation). If it does not, there is no single simple generalisation. The first result becomes questionable (i.e. zilch) if it does not hold next time.

To find generalisable results means actively looking for them. This is a matter of *design*, deliberately replicating studies, but under very different circumstances. It can be done even within a given study, by comparing differentiated sub-samples (e.g. men and women, morning and the afternoon, severely or less severely affected patients, etc.). The main aim remains to see if one can establish significant sameness - generalisations - not significant differences.

In contrast, most statistical analysis procedures in our textbooks are geared to analysing just a single set of data (SSoD). Examples are t-tests, correlations, and regressions. The process of analysing MSoD is not covered in most of our teaching. If it were, there would be far more success stories.

2.5 Major effects

The emphasis in our analyses has to be to see first of all if the *major* effects will generalise. There is little point in attempting to establish small differences, or to model minor variations and interactions in any great detail, before the wider generalisability of the main effects has been established. This encourages parsimonious modelling.

2.6 Samples and inferences

It helps to remember that statistical inferences from sample data are in fact merely statements about the population parameters. These are what is being estimated. The sample data and the process of statistical inference are therefore only a means to that end.

With a large sample (or with many cumulative MSoD samples) one therefore does not need much - if any - formal statistical inference. "Everything" is significant, virtually as in the whole population. We know that the standard error of a sample mean based on 2,000 or 20,000 readings will be "very small".

But is that result *generalisable*, i.e. will it also hold for (samples from) *other* populations? That is the first thing to establish. If it does *not* generalise, what is the point of analysing and modelling it further?

If it *does* generalise, that is where real analysis will, in my experience, begin. For example, how do I then describe and model the findings, i.e. the generalisable conclusions about all the different population parameters? How do they fit in with other such studies? How do I extend or test them further? What do the results *mean*?

2.7 The choice of model

Ways of choosing the form of one's model are hardly discussed in the statistical literature, let alone are they *formalised.* (The traditional quantitative emphasis is instead just on parameter estimation for a chosen model form, with more or less automatic reliance on least squares or maximum likelihood.)

I find that the choice of a suitable model is in practice determined by seeing what particular form of model holds across many different sets of data (MSoD). A model which holds only for an SSoD - a single set of data in isolation, but for no others - is generally of little or no practical use or interest to me, however close or significant the fit in that single instance. The change to judging model in terms of it's generalisability for MSoD rather than its degree of "best fit" for an SSoD is a paradigm shift.

3 Two Examples

3.1 A first example: "verbal memory following temporal lobectomy"

Channon and Hand (1996) compared people's ability to recall certain sets of 16 descriptive nouns (e.g. robin, sparrow, duck; cherry, banana, etc)

which were made up of certain categories (birds; fruit; etc).

The nouns were presented in three ways: either uncategorised (UC), grouped conceptually but nonetheless randomised in presentation (RC), or categorised and clustered (CC), e.g. four birds, then four fruits, etc. Comparisons of both "Immediate" and "Delayed" recall were made between a group of 12 lobectomized patients and 12 healthy controls.

Channon and Hand's Table 1 gives their full data, consisting of 72 individual readings in a two-way array.

		Immediate recall			Delayed recall		
Group	subjects	UC	RC	CC	UC	RC	CC
Left temporal	1	9	5	6	1	5	6
lobectomy	2	6	6	10	3	3	2
	3	4	2	8	1	1	2
	4	2	6	1	0	0	1
	5	10	9	12	1	7	8
	6	7	9	14	3	6	14
	7	6	6	10	1	6	10
	8	7	7	8	2	0	5
	9	4	4	7	0	0	1
	10	6	6	5	2	1	4
	11	4	9	8	2	1	6
	12	6	6	8	4	1	6
Healthy controls	13	6	11	13	2	10	10
	14	9	11	16	8	13	15
	15	8	7	14	5	10	13
	16	11	12	16	10	10	12
	17	11	15	16	10	16	15
	18	7	6	10	5	6	6
	19	8	10	12	2	11	9
	20	9	11	16	2	14	14
	21	8	10	13	2	11	8
	22	9	10	14	6	4	12
	23	5	4	8	2	5	6
	24	7	8	7	2	7	6

Table 1: Number of words recalled by the left temporal lobectomy and control subjects on the 3 word lists

It is not, I find, easy to see and summarise what this data says, or to communicate it to others. *Analysis* is therefore needed. The question is what kind of analysis? I opt (as usual for me) for simple data reduction procedures. Channon and Hand opt (as usual?) for F-tests and bar charts.

3.1.1 The main effects: averages

Table 2 gives my initial summary, by way of some averages. It shows how the Controls' average scores were 4 points higher overall than the Lobectomized ones (9 versus 5 on average). With this overall average difference of 4 in mind it is easy to see the difference was also broadly 4 or so for each of the six individual measures (Immediate UC, RC, CC, etc). And the UC seems were on average smaller (by 2 or 3 units) than the RC, and these were in turn lower (by about 2 units) than CC. (Note that a comparable analysis of variance is really as an analysis of *averages*. But it would usually not actually describe the results in this way at all, but only say how *significant* the results were).

One can also see that the three Immediate Recall averages were some 2 points higher than Delayed Recall (about 8 versus 6 overall). None of this is very clear from the original Table 1, even with hindsight. In other words, some "purely descriptive" analysis *is* helpful.

	Immediate recall			Delayed recall			
	UC	RC	CC	UC	RC	CC	Average
Lobectomized	6	6	8	2	3	5	5
Controls	8	10	13	5	10	11	9
Average	7	8	10	3	6	8	7

Table 2: Average verbal recall scores: Lobectomized and Control groups

3.1.2 Saying less by showing more

Table 3 shows the same cell means in a different way: with more precision one more digit but without summaries (the row and column averages). None of the above summary conclusions are affected. But they are a good deal less easy to see.

(One "Data Reduction" rule is "Round to two Effective Digits", i.e. ones which vary in that data. My Table 2 was in fact deliberately overrounded which sometimes helps perception, as here. But it worked. Another rule is: always work out averages first and *then* see if they help).

	Immediate recall			Delayed recall		
	UC	RC	CC	UC	RC	CC
Lobectomized	5.9	6.2	8.0	1.7	2.6	5.4
Controls	8.2	9.6	12.4	4.7	9.8	10.5

Table 3: The average recall scores as in Table 2 (to a second effective digit, without averages)

3.1.3 Individual differences

There is more to the Channon and Hand data in Table 1, namely the 12 individual subjects in each group (Lobectomized and Controls). The variation here needs also to be described, e.g. by calculating mean or standard deviations as summary figures (more "averages"). Thus the mean deviations of the 12 scores for each measure and subgroup are about 2 (with a slight trend in the mean absolute deviations (MADs) with the mean score). This provides a good feel for the data a MAD of about 2 and the basis for descriptive comparisons with other such studies (is the scatter there bigger, smaller, or about the same?)

Next, to see whether the scores of the individual "subjects" on the six different recall measures (Immediate UC, RC, etc) tend to go together, one can look and see a bit, and formalise this by rearranging the dots or calculating simple correlations.

Visually nothing obvious strikes the eye. This is confirmed by the correlations between pairs of measures being weak about 0.5 for the Lobectomized patients, and consistently slightly higher at 0.7 for the Controls, and also fractionally higher for the Immediate Recall measures in both cases.

An alternative is to use "ordering by size": the twelve rows (i.e. portions) in each block can be ordered by their overall now average, any common trend will then show up better (e.g. the higher control values for rows 16, 17 and 19). The details are "left as an exercise for the reader".

3.1.4 The need for generalisability

These results in isolation seem of virtually no interest. To me the question is how they tie in with other similar data (past and/or future) on (i) Verbal Recall, and (ii) Surgical interferences and/or Epilepsy. To do this quantitatively one has to have quantitative summary measures such as those above.

3.1.5 Statistical significance

The question of statistical significance will be raised. In as far as the data were not explicitly sampled (more or less randomly) from two well-specified populations, I am not sure what the techniques of statistical inference would tell us.

The data seem to me to be for two rather isolated and somewhat arbitrary (or "deliberate") local mini-populations of 12 lobectomized patients and 12 controls or the like, with no real "sampling". Nonetheless, I note that most of the systematic differences in Table 2 would turn out "signif-

icant", given the MAD's of 2 and n = 12. (Perhaps I did no *formal* tests because I knew what the answers would be?).

3.1.6 Alternative analyses

Channon and Hand (1996) purposely presented their data set for analysis by others. One example of an anlysis is in the original publication (Channon, Dawn and Polkey, 1989) to which the current authors refer. This consisted of:

(i) A considerable number of in-text F-tests, with the F-values themselves being recorded, plus the relevant degrees of freedom and p-values.

But typically in my experience, there was no *numerical* record of the means themselves, or any explicit description or discussion of the size of the mean differences between any different of the subgroups. We were told whether differences were "significant" (i.e. might well exist in the unstated population). But we were NOT told what it was that differed!

(ii) Half a dozen line or bar charts displaying sub-group means (as in Tables 1 and 2, but purely graphically).

Graphs can show up a clear story-line "graphically" if there is one. But otherwise they seem to me of limited value because they (a) usually do not bring out the story line explicitly, (b) are typically given insufficient "voice-over" verbal summaries in the text, (c) are largely qualitative (unless you get out a ruler *and* read off and then write down the numbers!), (d) allow no simple numerical summaries (overall means, differences, etc), (e) are very difficult to compare (especially if there were six graphs, say, on six separate pages).

3.2 A second example: "growth of bacteriological infection".

Francis and Whittaker (1996) provide a second example of data for analysis. It concerns a conductance measure which was recorded at 316 6-minute intervals. We are told that a (major) increase in the measure would indicate contamination in the biological food sample in question. Francis and Whittaker's Table 1 give the full data for two independent replicates. Half of this, Replicate I-1, is reproduced here as Table 4. The table is not very easy to read (although not impossible): where does a "major increase" occur?

Replicate I-1										
0.0	-2.5	-4.9	-6.7	-9.7	-12.9	3.3	3.4	2.9	3.4	0.5
-0.4	-2.2	-5.0	4.0	4.3	4.1	3.7	4.0	4.4	3.9	4.1
4.0	4.7	4.4	4.6	4.6	5.0	5.0	4.6	4.8	5.4	5.3
5.8	5.8	6.5	6.5	6.8	8.3	11.1	13.6	15.4	18.7	21.3
24.4	28.3	34.8	42.7	50.0	56.8	64.0	73.6	83.8	97.9	107.7
122.9	132.9	142.9	144.7	164.0	183.6	194.9	201.3	211.1	222.0	223.7
237.7	241.4	247.7	261.4	270.6	289.4	292.0	302.6	308.8	314.5	317.4
321.7	323.2	314.7	313.7	334.3	323.5	320.6	339.2	331.0	329.5	320.8
340.4	332.7	331.7	322.3	344.9	339.6	338.9	331.5	325.7	347.3	342.2
332.4	335.1	326.4	350.6	348.0	342.2	343.0	330.0	327.9	340.6	341.0
339.6	340.3	335.5	332.7	356.5	354.9	350.8	348.4	348.2	343.9	366.1
365.9	362.6	360.3	354.9	351.8	371.3	371.3	369.5	367.1	362.7	359.8
355.5	376.2	376.0	372.5	369.9	365.7	362.7	358.4	352.2	375.0	369.5
360.5	368.5	355.8	354.8	352.5	349.6	373.4	373.9	373.2	370.8	369.9
367.3	363.4	362.4	358.2	378.7	379.4	379.4	377.8	375.5	374.3	370.9
367.6	364.3	361.4	357.7	381.1	381.3	381.1	380.4	378.7	375.9	375.0
372.3	369.2	368.3	365.5	361.9	383.8	383.9	384.3	384.6	384.6	385.4
385.7	385.5	386.1	386.8	386.8	387.3	386.6	387.3	387.8	388.7	388.7
388.9	388.7	389.4	389.9	390.3	389.4	390.5	391.4	391.5	391.7	392.2
392.8	393.1	393.0	394.2	394.0	394.5	395.1	395.4	394.9	396.1	396.9
397.2	396.7	389.6	390.5	392.2	393.1	392.8	394.5	396.3	397.9	399.7
401.3	403.3	403.3	402.9	401.3	400.6	400.4	401.5	401.7	403.3	403.4
404.3	405.8	406.6	407.0	409.0	409.5	410.2	411.8	412.9	413.1	414.7
415.8	416.7	418.1	418.8	420.3	420.8	421.9	423.2	423.9	425.0	426.2
426.8	428.6	429.5	430.8	431.8	433.1	434.2	435.3	436.4	437.6	438.6
440.4	441.1	442.9	443.7	445.7	447.1	448.2	449.7	450.6	452.4	453.4
454.8	456.3	458.0	459.6	461.1	461.6	463.5	465.0	466.6	467.2	469.6
471.4	472.6	474.2	475.5	477.2	478.9	479.8	482.0	483.2	485.0	486.7
488.6	489.3	491.6	493.5	494.6	495.5	497.8	499.7			

Table 4: Conductance in μ-Siemens for 316 measurements at 6 min intervals on a biological sample (Serie I-1)

The authors' graph in their Figure 1 shows much more clearly that there is a large increase after about 50 readings in both data sets, levelling off after roughly 80 readings. It also shows some erratic variation after that, but in Replicate I:1 only. This picture of the data is however essentially qualitative (the numbers are not explicit and difficult to read off from their graph). Nor are such graphs easy to use when comparing many different data sets, especially not in routine quality control or the like.

Table 5 represents my initial step of analysis. It simply re-presents Replicate I-1, but with all the numbers rounded to two effective digits and with hourly summary averages. The table also uses time-related labelling of the readings (in hours), as being more transparent and also more flexible (e.g. if other such data were not to use 6-minute intervals). The table shows the main shape of the data clearly, I believe, and quantitatively: there is a

dramatic increase in the hourly means from about 5 before the fifth hour to almost 350 by hour 10, and a slow rise to 490 after that.

To see this broad picture routinely, the individual readings need of course no longer be reported (they can be left in the computer). This is shown in Table 6 for the second replicate. The trend is again clear. It is perhaps even clearer in the smaller Table 7, where the later hourly readings are grouped. (For routine purposes, selecting the hours to be grouped could be suitably computerised.)

Hour	Readings at 6-minute intervals										Aver.
1	0	-2	-5	-7	-10	-13	3	3	3	3	-3
2	1	0	-2	-5	4	4	4	4	4	4	2
3	4	4	4	5	4	5	5	5	5	5	5
4	5	5	5	6	6	7	7	7	8	11	7
5	14	15	19	21	24	28	35	43	50	57	31
6	64	74	84	98	107	120	130	140	140	160	110
7	180	190	200	210	220	220	240	240	250	260	220
8	270	290	290	300	310	310	320	320	320	310	300
9	310	330	330	320	340	330	330	320	340	330	330
10	330	320	320	340	340	330	330	350	340	330	340
11	340	330	330	350	340	340	330	330	340	340	340
12	340	340	340	330	360	350	350	350	350	340	350
13	370	370	370	360	350	350	370	370	370	370	360
14	360	360	360	380	380	370	370	370	360	360	370
15	350	380	380	360	370	360	350	350	350	370	360
16	370	370	370	370	370	360	360	360	380	380	370
17	380	380	380	370	370	370	360	360	360	380	370
18	380	380	380	380	380	380	370	370	370	370	380
19	360	380	380	380	380	380	390	390	390	390	380
20	390	390	390	390	390	380	390	390	390	390	390
21	390	390	390	390	390	390	390	390	390	390	390
22	390	390	390	390	390	400	400	390	400	400	390
23	400	400	390	390	390	390	390	390	400	400	390
24	400	400	400	400	400	400	400	400	400	400	400
25	400	400	400	410	410	410	410	410	410	410	410
26	410	410	410	420	420	420	420	420	420	420	420
27	420	420	430	430	430	430	430	430	430	430	430
28	430	440	440	440	440	440	440	440	440	450	440
29	450	450	450	450	450	450	450	460	460	460	450
30	460	460	460	470	470	470	470	470	470	470	470
31	480	480	480	480	480	480	490	490	490	490	480
32	490	495	490	490	500	500					490

Table 5: Replicate I-1, with better layout, rounded, and averages

There is also rather variable short-term variability in the original readings (as was noted by the authors). To summarise this, two descriptive measures can be used, the average differences between successive pairs of

Hours	1	2	3	4	5	6	7	8	9	10	11
Average	3	3	3	7	28	110	240	320	370	380	390
Hours	12	13	14	15	16	17	18	19	20	21	22
Average	400	420	430	430	430	440	440	440	450	450	460
Hours	23	24	25	26	27	28	29	30	31	32	
Average	460	470	480	490	510	520	530	550	560	580	

Table 6: Replicate I-2: summary

readings within each hour and their MADs. They could be set out as two extra volumes (e.g. on the computer). But for routine reporting they can be summarised further and verbalised. Thus the within-hour MADs are mostly very low, mostly about 2 (showing the steadiness of successive readings). But between Hours 5 to 9 both measures are high, up to just over 10. This simply reflects the big but steady upward trend in readings then!

Hours	1	2	3	4	5	6	7	8	9	10-19	20-29	30-32
Average I-1	-3	2	5	7	31	110	220	300	330	350	420	480
Average I-2	3	3	3	7	28	110	240	320	370	420	480	560

Table 7: Replicate I-1 and I-2: condensed

From Hour 9 to 17 the MADs in Table 5 however still remain quite high, whereas the average differences (i.e. *with* signs) in each hour are again small, at 2 or less. This highlights and also quantifies that during these hours there are *irregular* short-term fluctuations in the readings (as was shown qualitatively in the authors' Figure 1). In contrast, Replicate I-2 there were no such short-term wobbles. All this *could* be printed out more explicitly, if desired.

3.2.1 Alternative analysis

Alternative analysis are possible. For example, the authors refer to an earlier analysis of these and similar data sets (Whittaker and Frühwirth-Schnatter, 1994).

This involved fitting structural models of, basically, random walks with drift from unknown change-points. Assumptions about the normality etc of the residuals were made early on, Kalman filters were used, posterior distributions of the change-points were estimated, and so on, with many pages of dense if elementary mathematics. Apart from 20 graphics broadly like Figure 1 in the Francis and Whittaker (1996) note, the actual *results* in the earlier paper were reflected only in two small tables of estimated probabilities (sic!), shown to 4 decimal places at ten time points ($t = 12, 24, 36$ etc), as reproduced in Table 8. My question is what this complex modelling adds at this (or any) stage to the simple but quantitative data summaries

$Pr(infected\|y^t)$	col-12	ent-4	I-1	A-1	ent-15	lis-152
t=12	0.2968	0.2947	0.2977	0.2976	0.2976	0.2979
t=24	1.0	0.2928	0.2933	0.2932	0.2857	0.2939
t=30	1.0	0.9828	0.2909	0.2911	0.2806	0.2922
t=36	1.0	1.0	0.2889	0.2892	0.2756	0.2890
t=42	1.0	1.0	0.2942	0.2887	0.2704	0.2868
t=51	1.0	1.0	1.0	0.4180	0.2633	0.2845
t=57	1.0	1.0	1.0	1.0	0.7170	0.2821
t=63	1.0	1.0	1.0	1.0	1.0	0.2820
t=102	1.0	1.0	1.0	1.0	1.0	0.6254
t=114	1.0	1.0	1.0	1.0	1.0	0.9996

Table 8: On-Line probabilities of infection corresponding to Figure 1

in my Tables 5 and 6 or 7. (*My answer*: nothing.)

4 Discussion

I believe that *Data & Statistics* has been launched due to widespread dissatisfaction with the current state of data analysis. My own view is that:

(i) Academic statisticians are generally inept (or worse) at the simple tasks of analysing, summarising and reporting real data. This may seem remarkable and hence I am remarking on it and have also illustrated it from examples in this journal.

(ii) The degree of ineptness seems generally correlated negatively with the degree of mathematicisation they adopt.

(iii) There is an obsession with statistical inference, as if this were an end itself, i.e. just to show "that there is something there", rather to describe and understand what it actually is.

(iv) There is a frequent flight into qualitative but nonetheless usually rather opaque graphics.

As a result, teaching (as in our standard text-books) largely concentrates on statistical techniques and on "inference", rather than on analysing and only then perhaps "modelling" the data. There also is usually no mention of looking for any generalisable or "lawlike" results (i.e. across many sets of data), let alone advice and precepts for this.

Does this perhaps explain the widely reiterated dissatisfaction of the statistical profession and its leaders with the low regard in which they and their subject's usefulness are held?

As a contrast, I have attempted here to outline and illustrate a more data-orientated approach.

The reader may find both my seven precepts simplistic, and also my

reanalysis of the illustrative datasets. Almost any reasonably numerate person could have done them. There is no special kudos to be gained, as a somewhat sophisticated statistician, using F-Tests and Kalman filters.

The question is whether the analyst is more interested in results ("understanding data") or in techniques, and in communicating them to others or stroking his ego.

References

[1] Channon, S., Dawn, I., and Polkey, C.E. (1984), "The Effect of Categorization on Verbal Memory After Temperal Lobectomy", *Neuropsychologia*, **27**, 777-785.

[2] Channon, S. and Hand, D. (1996), "Verbal Memory Following Temporal Lobectomy", *Data & Statistics* in *Student*, **1**, 215-219.

[3] Francis, B. and Whittaker, J. (1996), "Gassing During the Growth of Bacteriological Infection", *Data & Statistics* in *Student*, **1**, 211-214.

[4] Whittaker, J. and Frühwirth-Schnatter, S. (1994), "A Dynamic Change-Point Model for Onset of Growth in Bacteriological Infections", *Applied Statistics*, **43**, 625-640.

Further Reading

[1] Ehrenberg, A.S.C. (1982, repr. 1994), *A Primer in Data Reduction*, John Wiley, Chichester and New York.

[2] Ehrenberg, A.S.C. and Bound, J.A. (1993), "Predictability and Prediction", *J. of the Roy. Statist. Soc. A*, **156**, 167-206.

[3] Lindsay, M.R. and Ehrenberg, A.S.C. (1993), "The Design of Replicated Studies", *The American Statistician*, **47**, 217-228.

Comment

David J. Hand

Department of Statistics, The Open University, United Kingdom

Statistical tools are developed and refined so that data and mechanisms generating data can be modelled more accurately and reliably. Unfortunately, it is an unpalatable truth that sometimes (academic) statisticians get swept up in the methodology itself, and in particular the mathematics of the methodology, and become distracted from the real objective of the exercise. As a result of this there has been a tendency for some branches of

statistics to develop in a way which is of little use as far as understanding data goes. Careful thought about the objective of a study and the gain in understanding which is likely to result from application of a sophisticated technique is an important pre-requisite to any analysis. Models which are over-refined, which are based on too many assumptions, or which are unreasonably complex, are unlikely to shed useful light on the structure of a set of data, and can even attract ridicule. On the other hand, sophisticated models allow one to separate out different aspects of relationships, and to identify, for example, that some particular pattern is not an artifact induced by other relationships.

In view of all this, I agree with much of what Ehrenberg has to say. In particular, I agree with the need for some preliminary investigation of the data. This can prevent gross mismodelling and embarrassing mistakes. The modern computer, although permitting ready application of immensely powerful data analytic techniques, also permits gross errors to be made with equal facility (Hand, 1984, 1985). The only way to avoid these is through careful understanding of the data. Elementary techniques play crucial role in this.

I also have a few specific points to make about Ehrenberg's examples. One lesson to be drawn from his first example (his analysis of the temporal lobectomy data), is that the way one summarises data must depend on the objectives - on what sort of patterns or relationships one is seeking or on what hypotheses one is exploring.

Turning to the bacteriological infection data, it is clear that how one presents a table of data must depend on why one is presenting it. Table 4 reproduced from Francis and Whittaker (1994) might not, as Ehrenberg says, be easy to read, but the aim was to present the data so that others could reanalyse it. A similar point applies to graphical displays - one might use them to communicate information or to reveal structure during the course of an analysis. I am an enthusiast for graphical displays - and in the case of these data prefer a diagram to Ehrenberg's Table 5. There is, of course, no harm in supplementing a graph with a table. While on the subject of Table 5, it seems to me that this table would be made even more clear if the terminal zeros in many of the figures were dropped, expressing the measurements in units of 10 μ-Siemens. The same point applies to Table 6.

Ehrenberg also makes some concluding comments. I think it is going too far to assert that 'academic statisticians are generally inept (or worse) at the simple tasks of analysing, summarising and reporting real data'. However, I would agree that there are *some* academic statisticians who fit this characterisation. There are some academic statisticians who have

never actually undertaken any statistical analysis of real data.

One of the positive effects of the computer has been a shift in emphasis away from formal inference to a more data analytic perspective. An example of a book which seeks to present a less formal viewpoint is Lunn and McNeil (1991) and an important book describing strategic issues of data analysis is that of Chatfield (1995).

References

[1] Chatfield C. (1995), *Problem Solving: a Statistician's Guide* (2nd ed.), Chapman and Hall, London.

[2] Hand D.J. (1984), "Statistical Expert Systems: Design", *The Statistician, 33,* 351-369.

[3] Hand D.J. (1985), "Statistical Expert Systems: Necessary Attributes", *J. of Appl. Statist.*, **12**, 19-27.

[4] Lunn A.D. and McNeil A.D. (1991), *Computer Interactive Data Analysis*, John Wiley, Chichester.

Comment

Joe Whittaker

Mathematics Department, Lancaster University, United Kingdom

Ehrenberg states seven precepts, most of which are uncontentious though somewhat poorly thought out. On the positive side he makes a valuable point concerning scientific data analysis by drawing the distinction between data analysis that takes place in the context of a single data set and data analysis that involves several similar sets. It is only the latter that give rise to findings which generalise to comparable but different populations and away from the data currently in the analyst's laboratory. And, as a corollary, he questions the merit of elaborate model fitting to data sets which, for whatever reason, only admit a single realisation.

On the down side I find part of the approach he advocates laborious, uninsightful and out of date. I illustrate by remarking on the comment Ehrenberg makes on the 'Growth of Bacteriological Infection' data set of Francis and Whittaker. First I would point out that this data set is only one of very many: each series is produced by the public health laboratory examination of a suspected food sample, and an unfortunate concomitant of European agri-business growth is an increase in the prevalence of food poisoning.

The 'complex modelling' proposed by Whittaker and Fruhwirth-Schnatter (1994) was designed to provide an early warning system for infected samples. They estimate the posterior probability that, given the sample data up to time t, the sample is a contaminated sample. For the sample I-1, this probability goes from about 0.3 to 1.0 between $t = 42$ and $t = 50$. In principle the laboratory could switch off the apparatus after about 1 hour, here saving some 30 hours of laboratory recording.

The technique generalised well with application to other similar series from different food samples with different infecting bacteria, eg col-12, ent-4 and others. It is hard to see how taking hourly averages, advocated by Ehrenberg as part of his philosophy of data reduction, would efficiently reduce laboratory processing time; in fact, it seems to directly violate Ehrenberg precept 1. Hand summaries of data are not the way forward in the current era of computer aided statistics, an era which allows both the recording and processing of massive data sets (eg stock exchange transactions, satellite image transmissions).

My final remark is to note that one of the classical aims of statistical inference is to avoid the pitfall of 'change capitalisation' inherent in the analysis of a single sample, for instance, by significance testing. More recently bootstrap inference, made possible by computational advances, is a serious endeavour to make the analysis of a single data more like the analysis of several data sets.

Rejoinder

A.S.C. Ehrenberg

South Bank University, London, United Kingdom

Reply to David J. Hand: In principle, David Hand and I agree. In practice, we differ widely. The reason, I think, is that he does not spell out what he means by models which are "over-refined" or which make "too many assumptions". To him, his own procedures are of course not like that, but merely "sophisticated models". It is like Maurice Kendall of old saying "To be used with care" of his more complex analysis procedures. In other words, OK when *he* did, but no good if anybody did.

Hand agrees that "elementary data analysis" is important. But Chatfield, in his *Initial Data Analysis* - a leading exponent - typically does not address the problem of many sets of data, nor of course did Tukey in his *Exploratory Data Analysis.* There is no clue how to bridge the gap with

what I am talking about - the analysis of MSoD.

I strongly disagree with Hand (and others) over "the way one tackles data must depend on the objectives." That looks too much like a licence to select and suppress (and is often treated like it!)

At the risk of a small degree of oversimplification, I think a data summary should first of all *describe* all the "main" features of the data in two ways, as

(i) Systematic effects (e.g.means $x = 10$, $y = 15$; or a relationship $y \simeq 6x + 12$).

(ii) Irregular variation (e.g. residual standard or mean deviations, outliers, with possible systematic subpatterns still lurking).

The criterion of a *good* summary is then whether you could reconstruct all the data from it, *within your stated stochastic limits.* After that you can try to interpret the data, test hypotheses, make predictions, or whatever, according to your chosen objectives. Let's keep testable (and "objective") data description and subjective objectives as separate as possible.

David Hand says that the table from Francis and Whittaker reproduced as my Table 1 was given to let others reanalyse the data. But my version of the table still makes the same data available "for reanalysis" (but is also *readable*). Anyway, journals are not the place for giving extensive data "for reanalysis". In the old days you said "Please write or phone for the data". Now you stick it on the WWW or the like.

Hand (like many) is an enthusiast for graphs. He says he prefers a diagram to my Table 5. But he doesn't actually give the diagram! (Do try drawing or computing one for Table 5, reader!). Nor does Hand say how any such diagram could then conceivably be used by anyone "for reanalysing the data". Graphs are next to no use (or worse) for reading off the individual readings for any kind of numerical (i.e. statistical) analysis.

Finally, a minor communication technicality is Hand's not uncommon suggestion to drop the terminal zeros in numbers such as those in Tables 5 or 6. But the eye readily filters out the zeros at a glance. The tables show that the third digits in three-digit numbers are *always* zero (see p.220 "Variable Rounding" in Ehrenberg 1982). Filtering out the zeros is vastly to be preferred to suddenly changing units to "10 μ-Siemens". Fancy suddenly expressing heights and lengths in 10cm or 10m. Ugh! Typically we are used to length like 1mm, 1cm, 100cm, 1m, 1bm, etc. People avoid the unusual, but use conventional units.

Reply to Joe Whittaker: Whittaker says my precepts are mostly uncontentious. But since he has not practised them, nor says that he *will* practice them in future, we do have contention: there is something to discuss and

to consider.

He notes that his data example is only one data set out of very many. Of course - most data are. But he still *treats* it as a *sample* data set. He does not say what he has learned or knows about the other all data sets, or how that has influenced what he has done with the one set on which he has actually reported.

For example, given that his is a very practical and routine problem, why not report that all the data so far have been of two types (if that *was* the case), namely

Steady: in some sets all the readings stayed below 10, say, for up to the maximum of 30 hours or so covered.

Variable: in the other data sets the readings were initially, for instance, below 10 for up to 5 hours or so; increasing fast up to 300/400 within the next few hours (say); and finally increasing slowly up to 600 for the final 20 hours.

There should also be a summary of the variability of the successive 6-minute readings (e.g. as ± 2 generally, but more for periods of fast growth of course, etc).

If the data sets *were* not like this, one should say so, and describe *how* they differed (e.g. was every data set really totally different from Table 1 and the above description).

Whatever the outcome, one can then review the monitoring system and decide what measures one needs routinely and get the computer to compute them. For example, simply to spot and highlight if and when the readings go well above 50 (say). In such cases one may want to have the readings also in more detail. Otherwise one can scrap that data set. That kind of "insight" is what I expect from having looked at all the data. Is anything more needed? I don't know what.

Three more specific points are:

(i) I do not understand why Whittaker feels that the lay-out in Table 7 does not provide the required early warning system. It shows more clearly than before if and when the data start to increase fast (which can then also easily be automated).

(ii) That although my few colleagues use more mainframe computer time than any other group in my present and last Universities, as well as personal Pentiums, I still believe that pencil-and-paper averages can sometimes not be improved upon (e.g in eye-balling the computer output).

(iii) He very clearly spells out the aim of classical statistical inference, namely to avoid over-interpreting a single data set. But that is my point precisely: if you have many data sets, you do not much need classical inference. Time marches on.

Data & Statistics in *Student (1996), Vol.1, No.4, 289-293*

Total Fat Content Evaluation of Olive Pastes by Nuclear Magnetic Resonance

Luís Gusmão, Maria do Céu Pinheiro-Alves and João Tiago Mexia

Estação Agronómica Nacional, INIA, Oeiras, Portugal

Núcleo de Tecnologia Alimentar de Elvas, INIA, Elvas, Portugal

Faculdade de Ciências e Tecnologia, Universidade Nova de Lisboa, Portugal

Abstract: Data set refers to total fat content evaluation on olive pastes, of four olive varieties, by five analytical methods, with samples collected along five weeks, and analysis performed in triplicate for each method, variety and week. Since the factorial ANOVA allows exclusion of the null hypothesis for systems, varieties and weeks, but not for the complex interactions, an additive conversion factor is not possible. However, correlation analysis of the results for the different analytical systems revealed a very high positive correlation for the Soxhlet system with the Oxford 4000 (based on Nuclear Magnetic Resonance). Therefore, although these last two analytical systems yield different results, we can convert one into another using linear regression of the respective results.

Key words: Correlation analysis, factorial ANOVA, regression analysis.

1 The scientific context

The Soxhlet system, for total fat content evaluation on olive paste and olive pomace (Portuguese official methods), requires a long period for extraction and solvent elimination, therefore, making it unsuitable for the purposes of routine labor control of olive pasts. There are, however, other analytical systems for total fat content evaluation, such as the Soxtec, Rafatec, Fosslet (using fat extraction by a solvent) and the Oxford 4000 (using Nuclear

Magnetic Resonance). Analytical systems using solvents differ according to the sample amount, the solvent used and the conditions of fat extraction. Apart from the solvent type, amount and temperature, the extraction duration, as referred to by Bernardini (1986), is paramount to the extraction effectiveness. In the Oxford 4000 system (described in Ruiz et al., 1991), a Nuclear Magnetic Resonance (NMR) spectroscope makes a quick measure of the hidrogen content of the sample, allowing total fat content evaluation. Although it requires a drying period for the sample (particularly long for olive pastes with higher water content), the measuring time is conveniently short (16 seconds).

Present data were used to show that the NMR technique, besides more convenient, can bred more reliable results than the other alternatives (in reference to the official method, Soxhlet) and to establish the adequate conversion procedure (see Pinheiro-Alves and Gusmão, 1994).

Since the factorial ANOVA allows exclusion of the null hypothesis for systems, varieties and weeks, but also for the complex interactions, an additive conversion factor was not possible. However, correlation analysis of the results for the different analytical systems revealed a very high positive correlation for the Soxhlet system with the Oxford 4000 (based on Nuclear Magnetic Resonance). The linear regression of the respective results showed a high determination coefficient ($R^2 = 0.72$) for the conversion equation, $y = -11.58 + 1.16x$, where y is the expected result from the Soxhlet system and x is the result obtained from the Oxford 4000 system.

2 The data

Experimental material reports to the 1992/93 campaign, with the varieties Picual, Cobrançosa, Galega and Blanqueta, sampled weekly (from 23rd November to 29th December), with three replications for each factor combination (analytical system, variety and week).

In Tables 1 to 4 total fat content in the dry matter of olives pastes assessed by Soxhlet, Rafatec, Soxtec, Oxford 4000 and Foss-let systems are presented, respectively for de varieties Picual, Cobrançosa, Galega and Blanqueta.

	Total fat content in the dry matter (%)				
Week	Soxhlet	Rafatec	Soxtec	Oxford 4000	Foss-let
1st	33,29	27,79	28,29	40,32	46,28
	32,59	30,27	24,44	39,78	45,65
	33,60	29,05	25,31	40,64	45,65
2nd	32,54	29,47	23,61	39,59	42,97
	33,09	34,04	27,59	39,73	42,88
	31,72	32,04	26,19	39,43	44,88
3rd	32,95	36,07	30,33	42,16	49,83
	34,53	38,26	32,55	41,80	49,83
	32,90	36,47	32,04	42,37	50,01
4th	33,33	34,16	27,86	38,81	46,40
	32,38	34,79	28,76	39,21	46,58
	32,25	34,76	32,00	39,33	47,03
5th	42,21	33,35	25,57	46,05	45,41
	42,68	31,30	25,84	45,11	45,49
	43,65	34,76	29,80	45,21	47,07

Table 1: Variety Picual

	Total fat content in the dry matter (%)				
Week	Soxhlet	Rafatec	Soxtec	Oxford 4000	Foss-let
1st	32,65	24,11	24,63	34,51	36,49
	31,82	28,73	24,42	35,57	38,52
	32,03	26,36	24,93	34,23	38,43
2nd	36,32	24,90	26,10	37,57	40,87
	34,60	26,68	29,47	37,40	40,67
	34,98	27,52	24,28	37,45	40,18
3rd	35,95	29,00	30,51	39,74	43,35
	35,12	30,66	28,51	39,19	42,88
	36,35	31,89	29,03	38,63	42,41
4th	27,21	27,35	23,66	34,11	47,15
	28,60	28,15	23,83	33,32	46,25
	29,14	28,54	24,06	34,22	46,70
5th	32,75	33,08	27,03	37,14	40,28
	33,69	34,52	28,41	38,00	41,25
	30,50	35,05	24,55	38,40	40,45

Table 2: Variety Cobrançosa

	Total fat content in the dry matter (%)				
Week	Soxhlet	Rafatec	Soxtec	Oxford 4000	Foss-let
1st	25,42	26,50	20,03	33,99	35,34
	22,62	29,38	23,07	34,79	35,34
	23,46	27,64	25,76	34,15	35,43
2nd	31,47	25,60	28,72	37,85	37,87
	31,27	26,09	27,93	38,49	39,46
	29,83	26,54	23,83	37,40	38,62
3rd	34,35	31,27	26,14	39,88	48,31
	36,79	30,51	25,66	38,96	47,47
	34,63	29,90	22,68	38,92	46,46
4th	35,41	34,32	20,08	39,00	44,50
	34,03	34,22	23,54	40,02	44,67
	34,16	34,86	23,94	39,42	43,98
5th	34,18	30,93	18,54	38,12	44,30
	32,46	30,13	15,64	37,38	43,51
	32,48	30,84	12,80	36,77	45,45

Table 3: Variety Galega

	Total fat content in the dry matter (%)				
Week	Soxhlet	Rafatec	Soxtec	Oxford 4000	Foss-let
1st	35,68	33,46	30,43	41,72	47,08
	36,29	34,73	34,66	42,24	46,49
	35,81	35,71	29,22	41,82	47,08
2nd	34,12	36,29	29,77	39,32	41,77
	33,66	34,80	32,85	38,59	41,69
	33,24	37,11	30,96	38,95	42,74
3rd	32,99	36,04	34,08	37,28	41,85
	33,18	36,37	33,79	37,54	42,34
	33,92	35,94	33,95	37,93	42,09
4th	34,74	33,37	24,69	40,44	45,96
	36,65	34,23	24,62	40,70	45,80
	36,24	32,78	21,91	40,06	46,54
5th	41,59	34,48	25,97	43,63	45,98
	42,33	34,86	28,20	43,85	46,22
	40,63	35,62	28,28	43,93	46,70

Table 4: Variety Blanqueta

Address for correspondence: João Tiago Mexia, Faculdade de Ciências e Tecnologia, Universidade Nova de Lisboa, 2825 Monte da Caparica, Portugal. Phone: 351 1 2954464, Fax: 351 1 2955641.

References

[1] Bernardini, E. (1986), *Tecnologia de aceites y grasas*, Spanish version of *The New Oil and Fat Technology,* Alhambra S.A., Madrid.

[2] Pinheiro-Alves, M.C. and Gusmão, L. (1994), "Use Of Nuclear Magnetic Resonance for the Determination of Total Fat Content of Olive Pastes" in *The Second International Conference on Applications of Magnetic Resonance in Food Science, Abstracts,* Universidade de Aveiro.

[3] Ruiz, L.F., Rodriguez, A.G.-O, Fernandez, M.H., Marquez, A.J., Pozo, M.P.L.Del, Bernardino, J.M., Ayuso, M.T.R. and Ojeda, M.U. (1991), "Analistas de laboratório de almazara", Dirección General de Investigación y Extensión Agrárias, Centro de Información y Documentación Agrária, Apuntes, Sevilha, 37-39.