IN THIS ISSUE

BAYESIAN SPATIAL MODELS FOR MAPPING CHILD GROWTH AND NUTRITION

Malnutrition and impaired child growth is unfortunately still a reality in many parts of the world. It has received much attention in the last 30 years from the general public, governement agencies as well as from medical researchers. Data obtained on child growth shows distinct geographical pattern in many countries. Ivo Müller and Penelope Vounatsou demonstrate the use of hierarchical Bayesian models for spatial, normally distributed data, in the area of nutritional epidemiology. The emphasis lay on the investigation of different ways of modelling the spatial information available in nutritional indicators, in order to study the nature of spatial patterns in child growth in Papua New Guinea.

Ivo Müller was born on April 6, 1967 in Luzern, Switzerland. He studied biology and mathematics at the university of Basel, Switzerland. He obtained his Master of Science in Plant Population Biology in 1995 and his Master of Science in Statistics in 1998 from the University of Neuchâtel. Scince 1996 works on his Ph.D thesis on the Applications of spatials statistics in the field of tropical health at the Swiss Tropical Institue, Basel.

SAMPLE SIZE DETERMINATION IN CLINICAL TRIALS

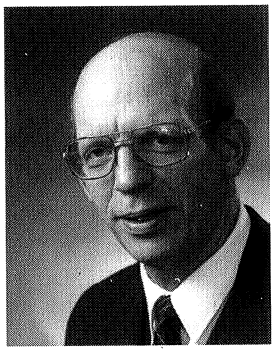
In the second article of this issue, Hamid Pezeshk and John Gittins provide an exact solution to a Bayesian version of the sample size problem in clinical trials. They show that the number of users of the new treatment is a function of the posterior mean of the performance parameter.

Hamid Pezeshk was born on July 5, 1961 in Tehran, Iran. He obtained his Bachelor of Science in Statistics in 1987 from Shiraz University, Iran. Since 1997 he is a Dphil student of the Department of Statistics at the University of Oxford, working under direction of Professor John C. Gittins on *Decision Theoretic Methods for the Determination of Sample Sizes for Clinical Trials*.

ESTIMATING A PROBABILITY DISTRIBUTION WITH GIVEN MOMENTS FROM EMPIRICAL DATA

In the third article of this issue, Daniel Costa and Edward A. Silver propose to find the distribution which best matches an empirical distribution. Considering a sample of observations of a discrete variable and its first m moments, they propose two objective functions, namely the least squares and the maximum likelihood, to measure the resemblance between the empirical distribution and the distribution to be estimated. Two algorithms are then used to solve the resulting convex programming problems.

Daniel Costa is currently working as a logistics manager at Nestlé. He was visiting researcher at the Department of Computer Science, Simon Fraser University, in 1991-92. After obtaining his doctoral degree in Operations Research from the Swiss Federal Institute of Technology in 1995, he held a postdoctoral position in the Faculty of Management at the University of Calgary and was senior lecturer in the Group of Statistics at the University of Neuchâtel. His research interests are in the fields of scheduling, inventory management, graph colouring and heuristics for deterministic and stochastic combinatorial optimization problems.



Dr. Edward A. Silver holds the Carma Chair at the University of Calgary. He received a Science Doctorate in Operations Research at M.I.T. Professor Silver has consulted and provided executive training and other workshops for a wide range of organizations. These activities have covered a broad range of topics including inventory management, business process improvement, production planning, and the use of quantitative modelling to aid in decision making. He has co-authored three editions of the textbook *Inventory Management and Production Planning* and Scheduling and has had over 120 articles published in professional journals. Dr. Silver is a past president of the Canadian Operational Research Society and received the 1990 Award of Merit from the Society. He was elected as a Fellow of the Institute of Industrial Engineering in 1995. Professor Silver is a past president of the International Society for Inventory Research.

PROOF WITHOUT WORDS : SUM OF SQUARED INTEGERS

When ranking methods are used to juge the effect of a treatment, the variance of the signed rank statistic is the sum of squared integers, if there is no treatment effect. Nanny Wermuth and Hans-Jürgen Schuh provide the reader with a graphical method to explain the sum of squared integers.

JOHN W. TUKEY, FOUNDER OF EXPLANATORY DATA ANALYSIS

In the *Capturing History* section of this issue, homage is paid to one of the living legend statistician of our time, John W. Tukey, the founder of Exploratory Data Analysis. Since one of the major goal of this part of the journal is to preserve the precious hand-writing of great statisticians, we took the opportunity to furnish the readers of this issue with an extract of his course *Statistics 411* developped at Princeton University with his kind permission.

DATA & STATISTICS

Dr. Philippe Mojon provides the readers of Data & Statistics with a data set from clinical trials. In his article entitled *Reliability of clinical indices* Dr. Mojon shows how to assess the reliability of 9 clinical dental indices. The data comes from 15 subjects who were examined by 2 trained examiners and re-examined to weeks later. He then evaluated the reliability of these 9 indices by computing three indexes, the Russel and Rao Index, the Lerman Index and the overall non weighted and weighted Kappa Index.

Student (1999), Vol.3, No.1, 1-18

Bayesian Spatial Models for mapping Child Growth and Nutrition

Ivo Müller and Penelope Vounatsou

Swiss Tropical Institute, Department of Public Health and Epidemiology, Basel, Switzerland

Abstract: Studies on child nutrition in Papua New Guinea have indicated geographical differences in nutritional status, but the nature of and the factors influencing the spatial heterogeneity have not been adequately investigated. Nutritional status is characterised by indicators such as stunting (height-for-weight) and wasting (weight-for-height), which are normally distributed. In this paper we demonstrate the use of hierarchical Bayesian models for spatial normally distributed data in the area of nutritional epidemiology. The focus of the work is on investigating the different ways of modelling the spatial information available in nutritional indicators. Various models exploring different between area correlation structures are proposed. MCMC methods are used for inference and prediction. The models were validated on simulated data and applied to a nutritional data set from Papua New Guinea. The results reveal strong spatial associations in child malnutrition, which are best described with a neighbour-based correlation structure.

Key words: Bayesian inference, conditional autoregressive priors, distance priors, exchangeable priors, hierarchical models, Markov Chain Monte Carlo, nutritional indicators; spatial modelling.

1 Introduction

Malnutrition and impaired child growth is unfortunately still a reality in many parts of the world. It has received much attention in the last 30 years from the general public, government agencies as well as from medical researchers. Most of the effort was spent investigating causes of malnutrition and its effects on child growth (Waterlow, 1992) as well as identifying areas of greatest need. Such studies of child nutrition normally focus on children

under 5 years, the period when growth potential of a child is determined and the influence of nutritional deficiencies is therefore strongest and long lasting. The roles of different risk factors for malnutrition and the geographical distribution of high risk areas are well known by now, but no studies have so far tried to explicitly model the spatial variation of child growth, although many authors noticed strong regional differences. In Papua New Guinea (PNG), for example, Heywood et al. (1988) found striking differences in child growth among different population and regions. Though the geographical patterns of child nutrition were easily discernible and differences were related to differences in diet and to differences in physical environment (Smith et al., 1993), a statistical investigation of the nature of these spatial patterns has not been done yet. A better knowledge of the nature of these spatial patterns in child growth and especially their relationship with different environmental, socio-economic and demographic factors would help to identify what might cause these regional differences and help in formulating health policies and in designing interventions.

For a long time spatial analysis of health data was restricted to the mapping of individual cases or rates of particular diseases and other health relevant parameters. More of these 'health' maps are becoming available as the use of geographical information systems (GIS) in health related contexts increases. Although disease maps may provide clues to the aetiology of diseases and health maps may facilitate decisions concerning planning of health systems, sound analytical methods are needed to assess associations between health events, aetiological and environmental factors that vary gradually over geographical regions.

The last two decades saw a vast and still ongoing development of statistical methods for the analysis of spatial data (Bailey and Gatrell, 1995, Cressie, 1993, Haining, 1990). Exploratory techniques, such as spatial moving average, median polish, spline and kernel smoothing have been used to investigate unstructured variation, while (co)variograms have been used to test for spatial correlations (Bailey and Gatrell, 1995). Different models based on the concept of complete spatial randomness (CSR) were developed to cope with spatial point patterns and space-time clustering (Diggle, 1993). In the context of ecological analysis various modelling approaches, such as spatial regression and autoregression models, have been suggested for continuous normal responses as well as for Poisson data (see Cressie, 1993 and Haining, 1990, for an overview). Many of these methods involve very highly parameterised models, and model fitting can pose serious computational problems.

Computational advances in Bayesian statistics with the development of Markov Chain Monte Carlo (MCMC) techniques led to new advances in spatial modelling. In the last few years hierarchical Bayesian models have been

developed to deal with spatial problems. They provide a more flexible and accurate way of modelling spatial data since prior specifications allow a broad type of spatial structures to be studied. In addition, the propagation of uncertainty through the different levels of hierarchical structures allows more accurate estimation. Hierarchical Bayesian models have been first successfully applied to model extra-Poisson variation in disease risk and mortality data, e.g. breast cancer and Hodgkin's disease mortality (Bernardinelli and Montomoli, 1992). Recently, models for normal data in an economic context (Gelfand et al., 1998) and for multinomial data in the context of geographical mapping of gene and haplotype frequencies (Vounatsou et al, 1998) have been developed.

The purpose of this work was to demonstrate the use of hierarchical Bayesian models for spatial, normally distributed data, in the area of nutritional epidemiology. The emphasis lay on the investigation of different ways of modelling the spatial information available in nutritional indicators, in order to study the nature of spatial patterns in child growth in PNG. The work is closely related to approaches outlined by Bernardinelli and Montomoli (1992), Gelfand et al. (1998) and Vounatsou et al. (1998). The models considered in this study are appropriate for area spatial data (Cressie, 1993, Haining, 1990), that is data associated with spatial zones or areas. The entire surface considered is partitioned into a distinct set of such areas, which are often irregular but may also have a regular lattice structure. The spatial information on individual data points is restricted to the areas where they were measured. Therefore, the areas are the primary spatial units on which the analyses are based. In many analyses these areas are given by administrative units such as municipalities, provinces, states. This type of data is frequently found in the context of health, economics (e.g. distribution of household income among villages within a state) or education (e.g. test scores from children attending different schools within a town).

A description of the data is given in Section 2. Section 3 sets out the formal modelling. Section 4 deals with appropriate model choice criteria, while implementation details of the MCMC simulation are given in Section 5. In Section 6 we validate the models on simulated data and we apply them on a nutritional data set. Section 7 discusses the results and suggests possible extensions.

2 Data description

The data for this study come from the 1982/83 Papua New Guinea National Nutrition Survey (NNS) (Heywood et al., 1988). This database consists of the results of a cluster sample survey of child anthropometry carried out in 1279 villages, distributed over 18 out of 19 provinces of Papua New Guinea

in 1982/3. The survey involved 27'588 children aged under 5 years. The data in the survey include anthropometric measures (age, height, weight) of a child as well as dietary, socio-economic factors and demographic data about the child and its family. A detailed description of the data is given by Smith et al. (1993).

The nutritional status of a child is usually described by investigating indicators based on anthropometric measurements. The most frequently used indicators are height adjusted for age (HAZ), weight adjusted for height (WHZ) and weight adjusted for age (WAZ) scores, each describing a different aspect, i.e. stunting, wasting and underweight, of child growth and nutrition (Waterlow, 1992). Such scores can be obtained by the LMS method developed by Cole (1992). This procedure yields age-independent standard normal deviate Z-scores.

As no reliable village co-ordinates were available, different spatial units of analyses had to be used. For planning and resource management purposes, Papua New Guinea was divided into geographical units called resource mapping units (RMU) by the Division of Land Use Research of the Commonwealth Scientific and Industrial Research Organisation (CSIRO). These units could be used in the analyses as each village can be assigned to an RMU, whose exact geographical position is known. This choice of the RMU as unit of spatial analyses also allows to link the NNS database to databases containing environmental (Bellamy, 1986) and agricultural data (Allen et al., 1994).

The neighbourhood structure of spatial units needed for the analyses as well as areas for mapping the results were obtain through tessellation (Cressie, 1993, algorithm used by Lee & Schachter, 1980). Distances between units were calculated as Euclidean distance between the centroids of the units.

3 Hierarchical modelling of nutritional scores

Several models to analyse the data are considered here, which differ by the way the spatial correlations among the different areas are modelled.

The models used are built in a hierarchical way. Let z_{iq} denote the z-score of the individual child q within the spatial area i and let θ_{iq} be $\theta_{iq} \equiv E(z_{iq})$. At the first stage of hierarchy we assume that the z_{iq} are conditionally independent given the θ_{iq}, i.e. $z_{iq} \sim N\left(\theta_{iq}, \sigma_z^2\right)$. The θ_{iq} are defined using customary linear models and include any covariate information available at the level of individuals as well as effects at the level of the areas. For notational convenience we denote the area effects by ϕ_i, in case these effects incorporate the spatial structure of the units and by α_i in case the units are considered to be independent. In the Bayesian context, all parameters introduced will be random. Modelling these parameters constitutes the second

level of the hierarchy. The following second stage models are considered:

$$\begin{array}{ll} \text{Model (1)} & \theta_{iq} = \mu + \alpha_i \\ \text{Model (2)} & \theta_{iq} = \mu + \phi_i \\ \text{Model (3)} & \theta_{iq} = \mu + \alpha_i + \phi_i \end{array} \qquad i = 1, \ldots, R \text{ and } q = 1, \ldots, n$$

The first model assumes that areas are independent of each other. The parameter μ represents a global mean, which expresses the average nutritional score throughout the map. The parameter α_i represents the area effect which expresses the differences between the global average nutritional score and the average score of the area i. For the global mean μ we adopt a non-informative uniform prior $\mu \sim U[-\infty, \infty]$, while for the area effects α_i we adopt normal independent priors with mean 0 and variance σ_α^2. It is convenient to choose standard conjugate $IG(a_0, b_0)$ and $IG(a_1, b_1)$ hyper-priors for the parameters σ_z^2 and σ_α^2 respectively with chosen a_0, b_0 , a_1, b_1 to give vague priors. It should be noted that it is not possible to assume a non-informative uniform prior for the variances σ_z^2 and σ_α^2, as this would lead to an improper posterior.

This parameterisation leads to the following joint posterior distribution:

$$\begin{aligned} p(\mu, \alpha_1, \ldots, \alpha_R, \sigma_z, \sigma_\alpha, \sigma_\phi \mid z) \propto & \prod_{i=1}^{R} \prod_{q=1}^{n_i} p(z \mid \mu, \alpha_i, \sigma_z) \\ & \times \prod_{i=1}^{R} p(\alpha_i \mid \sigma_\alpha) \\ & \times p(\mu) \times p(\sigma_z) \times p(\sigma_\alpha) \end{aligned} \qquad (1)$$

Model (1) will be referred to as the independent model (IND) since it assumes no structure in in the α_i and therefore in areal units. This model represents a kind of baseline model against which the other models can be compared

In Model (2) introduces spatially structured area effects ϕ_i. These spatial effects ϕ_i are modelled using the conditional autoregressive model suggested by Besag (1974) and Clayton and Kaldor (1989), i.e. we assume spatial a Gaussian prior for the conditional distribution of each ϕ_i given ϕ_j, $j \neq i$, which can be written:

$$p(\phi_i \mid \phi_{j, j \neq i}, \sigma_\phi^2) \propto \frac{1}{\sigma_\phi} \exp\left(-\frac{w_{i+}}{2\sigma_\phi^2} \left(\phi_i - \frac{1}{w_{i+}} \sum_{j=1}^{R} w_{ij} \phi_j \right)^2 \right) \qquad (2)$$

with $w_{ii} = 0$, $w_{ij} = w_{ji}$ for $i \neq j$ and $w_{i+} = \sum_j w_{ij}$ (Bernadinelli & Montomoli, 1992).

The weights w_{ij} reflect the degree of influence of the area unit i to the unit j, while the σ_ϕ^2 specifies the variability in the ϕ_i's conditional on all the other

area effects ϕ_j, $j \neq i$. As for the independent model it is computationally convenient to choose an inverse gamma prior $IG(a_2, b_2)$ for σ_ϕ^2.

The joint prior for the $\phi = (\phi_1, \ldots, \phi_r)^T$ is a multivariate normal with mean 0 and an inverse variance-covariance matrix B with diagonal elements w_{i+}/σ_ϕ^2 and off-diagonal elements $-w_{ij}/\sigma_\phi^2$(Besag, 1974). This prior distribution is improper, since the matrix B^{-1} is not of full rank. Nevertheless the model can be estimated, as the data contain information on the location of the ϕ_i's.

The above prior constitutes a general family of prior distributions. Different priors can be derived from it depending on our prior beliefs on the nature of spatial structure or on the way we assume that each area influences the others. This is specified in the weights w_{ij}. If we assume spatial dependence only among neighbours, we define the w_{ij}'s to be $w_{ij} = 1$, $i \neq j$, if i is contiguous to jth area and $w_{ij} = 0$ otherwise. This results in a conditional autoregressive (CAR) type of prior (Besag et al., 1995). If $w_{ij} = 1$ $\forall i$, j, $i \neq j$ an exchangeable prior (EX) is obtained, which captures heterogeneity but is not truly spatial as it assumes each area to be a neighbour of all other areas. Alternatively we can assume that spatial interaction decreases with distance, by setting

$$w_{ij} = \frac{1}{d_{ij}}, \; i \neq j,$$

where d_{ij} is a measure of distance between the centroid of area i and the centroid of area j. This choice of w_{ij} defines a DIST prior.

We extend this prior to model spatial dependence not only in terms of neighbours or distance but as a mixture of both. Accordingly, we define the w_{ij} to be $1/d_{ij}$, if i and j are neighbours and $w_{ij} = 0$ otherwise. This will be referred to as CARDIST prior.

For the parameters μ and σ_z^2, a uniform prior is adopted for μ and an inverse gamma prior $IG(a_0, b_0)$ for σ_z^2 as in Model (1).

The above parameterisations lead to the following joint posterior distribution for Model (2):

$$\begin{aligned} p(\mu, \phi_1, \ldots, \phi_R, \sigma_z, \sigma_\mu, \sigma_\phi \mid \mathbf{z}) \;\propto\; & \prod_{i=1}^{R} \prod_{q=1}^{n_i} p(z_{iq} \mid \mu, \phi_i, \sigma_z) \times p(\mu) \\ & \times p(\phi \mid \sigma_\phi) \times p(\sigma_z) \times p(\sigma_\phi) \end{aligned} \quad (3)$$

In Model (3) includes spatially structured (ϕ_i's) as wells as independent (α_i) area effects. This model with the CAR prior for the ϕ_i's and the independent prior for the α_i's corresponds to the convolution model (Besag, 1989, Mollié, 1996). The full conditional hierarchical specification of the model together with the Bayesian learning assure that the parameters α_i's and ϕ_i's are identifiable in the posterior. An EX prior for the ϕ_i's makes lit-

tle sense for this model, as it would lead to a model with two factors ϕ_i and α_i which capture similar types of variation.

Model (3) allows a direct comparison of spatial vs. non-spatial variance components, as the conditional variance of the area means $\psi_i = \alpha_i + \phi_i$ given all other ψ_j, $j \neq i$ is the sum of the variances of the independent components ϕ_i and α_i:

$$\text{var}\left[\theta_i \mid \theta_{-1}, \upsilon\right] = \frac{\sigma_\phi^2}{w_{i+}} + \sigma_\alpha^2 \tag{4}$$

with

$$w_{i+} = \sum_j w_{ij}.$$

If $\dfrac{\sigma_\phi^2}{w_{i+}}$ is bigger than σ_α^2 then spatially structured variation dominates; otherwise unstructured heterogeneity dominates (Molliè, 1996).

The joint posterior distribution for Model (3) is given by:

$$\begin{aligned} p\left(\mu, \alpha_1, \ldots, \alpha_R, \phi_1, \ldots, \phi_r, \sigma_z, \sigma_\alpha, \sigma_\phi \mid \mathbf{z}\right) \propto & \prod_{i=1}^{R} \prod_{q=1}^{n_i} p\left(z_{iq} \mid \mu, \alpha_i, \phi_i, \sigma_z\right) \\ & \times p\left(\mu\right) \times \prod_{i=1}^{R} p\left(\alpha_i \mid \sigma_\alpha\right) \\ & \times p\left(\phi \mid \sigma_\phi\right) \times p\left(\sigma_z\right) \times p\left(\sigma_\alpha\right) \\ & \times p\left(c\right) \end{aligned} \tag{5}$$

In the following, we name Model (3) with the CAR prior for ϕ_i's CONCAR and with the DIST prior for the ϕ_i's CONDIST (Table 1).

To complete model specification we should define the scale and shape hyperparameters of the inverse gamma priors. In absence of any prior information it is reasonable to choose a vague IG prior with mean based on the observed area means and large variance. A reasonable guess for σ_α^2 might be to chose a_1, b_1 in such a way that the mean a_1/b_1 of the hyperprior is $s_{\overline{z}_i}^2$, where $s_{\overline{z}_i}^2$ is the variance in observed area means and b can be set small in order to obtain a sufficiently vague prior. Similarly, for the hyperprior of σ_ϕ^2, parameters a_2, b_2 may be taken to be such as $a_2/b_2 = \overline{w} s_{\overline{z}_i}^2$, where $\overline{w}$ is the average value of w_{i+}. For the convolution model, Mollié (1996) advocates giving equal importance to σ_α^2 and σ_ϕ^2, that is $a_1/b_1 = s_{\overline{z}_i}^2/2$ for σ_α^2 and $a_2/b_2 = \overline{w} s_{\overline{z}_i}^2/2$ for σ_ϕ^2. Bernadinelli et al. (1995), working in the context of Poisson regression, suggest $\sigma_\alpha^2 \approx 0.7\sigma_\phi^2$ based on a simulation investigation. Since the priors are chosen to be vague, the two approaches are not likely to differ much and priors giving equal importance to σ_α^2 and σ_ϕ^2 were used. In the case of σ_z^2 a reasonable guess for a_0/b_0 is the average observed within area variance in all models.

4 Model Implementation

The posterior densities of the above hierarchical models are very high dimensional and complex. Markov Chain Monte Carlo simulation is needed to obtain estimates of the posterior and predictive quantities of interest. To simulate draws from the respective posterior densities a Gibbs sampling scheme (Gelfand & Smith, 1990) was used.

In the spatial models discussed previously, the parameters μ and ϕ_i are not identifiable, since the models are invariant under the transformation $\mu' = \mu + c$ and $\phi_i' = \phi_i - c$ for any constant c. To overcome this problem, the constraint $\sum_{i=1}^{R} \phi_i = 0$ was imposed at the end of each Gibbs iteration, by adding the mean of ϕ_i's to the samples of μ and subtracting it from each ϕ_i. A similar constraint has to be imposed on the α_i.

A multiple chain Gibbs sampler with 5 independent chains run in parallel was used. In order to assure independent sampling from the posterior densities a sampling scheme was designed which involved a 'burn-in' period of 1'000 iterations. After convergence was reached a final sample of size 1500 was collected by taking every 30th value of each chain. Convergence was assessed using the method of Gelman & Rubin (1992).

The programs were written from scratch in C++ and use IMSL statistical libraries functions. A single run of programs for the spatial models (2) with the nutritional data sets (i.e. 575 different areas) took approximately 1 hours on a Pentium II 333 MHz PC.

5 Model comparison

The spatial models are not regular since the number of their parameters does depend on the number of data points, therefore model choice criteria using degrees of freedom are not applicable and other ways of comparing the models must be used. Also, the use of improper priors for the spatial effects ϕ_i precludes traditional Bayesian model comparison via Bayes factors, therefore an alternative approach developed by Gelfand and Gosh (1998) was used. The above authors propose penalised likelihood criteria arising under the predictive posterior distribution. Their method asses the predictive preformance of a model in terms of its ability to predict a replicate of the data. Let $\mathbf{z}_{new}$ be a replicate of the observed data vector $\mathbf{z}_{obs}$ arising under the model, say M. Then the posterior predictive distribution of $\mathbf{z}_{new}$ under the model M is

$$p^{(M)}\left(\mathbf{z}_{new} \mid \mathbf{z}_{obs}\right) = \int p\left(\mathbf{z}_{new} \mid \eta^{(M)}\right) p^{(M)}\left(\eta^{(M)} \mid \mathbf{z}_{obs}\right) \qquad (6)$$

where $\eta^{(M)}$ is the vector of model parameters.

Following Gelfand et al. (1998) we choose the model M which provides the smallest value of

$$D_k(M) = P(M) + \frac{k}{k+1} G(M) \tag{7}$$

with respect to (6), where

$$G(M) = \sum_{i=1}^{R} \sum_{q=1}^{n_i} \left(z_{iq,obs} - E_p \left(z_{iq,new}^{(M)} \mid \mathbf{z}_{obs} \right) \right)^2$$

and

$$P(M) = \sum_{i=1}^{R} \sum_{q=1}^{n_i} \mathrm{var}_p \left(z_{iq,new}^{(M)} \mid \mathbf{z}_{obs} \right),$$

$G(M)$ is interpreted as a goodness of fit term, which is 0 when

$$E_p \left(z_{iq,new}^{(M)} \mid \mathbf{z}_{obs} \right) = z_{iq}$$

and increases as $E_p \left(z_{iq,new}^{(M)} \mid \mathbf{z}_{obs} \right)$ moves away from z_{iq}. $P(M)$ is interpreted as a penalty term and can be shown to be approximately a weighted sum of predictive variances. It is inflated for both underfitted and overfitted models. A poor model will make both $G(M)$, $P(M)$ large and therefore $D_k(M)$, while a good fit is characterised be both small $G(M)$ and $P(M)$. As models become more complex, $G(M)$ will decrease while $P(M)$ will increase.

The posterior predictive expectations and variances are computed by Monte Carlo integration using samples from the required predictive distributions $p(z_{iq,new} \mid \mathbf{z}_{obs})$. Replicate data $z_{iq,new}$ under the model M are generated using the composition method by noting that $p\left(\mathbf{z}_{new} \mid \eta^{(M)}\right) \sim N\left(\underline{\theta}^{(M)}, \sigma_z^{2(M)} \mathbf{I}\right)$.

In particular for each sample $\left\{\underline{\theta}^{(M)}, \sigma_z^{2(M)}\right\}$ drawn from the posterior distribution via the Gibbs sampler a $\mathbf{z}_{new}$ is generated from the above normal distribution.

6 Results

6.1 Simulated data set

Besides the nutritional data described above, 3 simulated data sets, which resemble the nutritional data in their structure, were used in order to study the general behaviour of the models.

A first data set was simulated based on the assumption of independence of areal units in order to test whether the models rightly differentiate the inde-

pendent model against the spatial ones. To test for the ability of the models to properly identify spatially structured data sets another 2 data sets were simulated, which resemble a conditional autoregressive spatial structure. Unfortunately it is impossible to simulate such data, as the joint distribution of ϕ_i's is improper under the assumption of a conditional autoregressive model. The inverse variance-covariance matrix $\mathbf{B}$ of the joint distribution has diagonal elements w_{i+}/σ_ϕ^2 and off-diagonal elements $-w_{ij}/\sigma_\phi^2$. This matrix has all row sums equal to 0 and is therefore not invertable. In order to obtain a matrix $\mathbf{B}^*$ which 'mimics' a conditional autoregressive structure, but nevertherless is invertable a constant of 1 was added to diagonal elements of $\mathbf{B}$. In the case of a neighbour-based spatial structure, adding a constant of 1 is equivalent to defining the area i to be neighbour of itself. This allowed the simulation of spatially structured data sets, but their spatial structure is not CAR and estimates of the model will not exactly reproduce the variances used in the simulation of the data sets.

All three data sets consist of 35 areas in a regular 5×7 lattice. Within each area 3 to 8 individual values $z_{ij} \sim N\left(0, \sigma_z^2\right)$ were calculated. The θ_i's were defined as in Model (1) $\alpha_i \sim N\left(0, \sigma_\alpha\right)$ and Model (2) $\phi_i \sim MVN\left(0, \mathbf{B}^{*-1}\right)$. The global mean μ was taken to be 0 except for the spatial data set 2. It is noted that individual variance was chosen $\sigma_z^2 \gg \sigma_\phi^2, \sigma_\alpha^2$ the variance of area means, which is a typical characteristic of nutritional data sets. The parameters chosen are presented in Table 2.

Taking the limitations of the spatial data sets into account, the analyses of the simulated data sets still provide valuable information on the behaviour of the different models. All models fitted have a tendency to be underfitted for all three data sets as the penalty $P(M)$ tends to be higher than the goodness-of-fit term $G(M)$ and show generally similar values of D_k scores (Table 2). This is due to the fact that within-area variability dominates the between-area variability (see σ_z^2 vs. $\sigma_\alpha^2 + \sigma_\phi^2/\widetilde{w}_+$ in Table 3), as it is usually the case in nutritional data sets, especially when no covariate information at the individual level is taken into account. Also, the model selection is not sensitive to the choice of k, despite the small difference in D_k.

Table 2 confirms that the model choice criterion is able to identify the model which has generated the data. As we can see, data generated under the independent model are best fitted by IND, CONCAR and CONDIST models, since for these models D_k has the smallest value. The variance estimates for the CONCAR and CONDIST models also indicate that the data set is correctly identified as non-spatial. The estimated variance components for σ_ϕ^2 and σ_α^2 are very close to the true ones (Table 3).

Data generated under model (2) are also fitted best by the models CAR and CONCAR. We can also see that the DIST prior does worse than the CAR one. The advantages of the CAR over the DIST prior become even

more pronounced, when the variance estimates are considered. In the case of the CAR prior the introduction of a non-spatial random effect, that is CAR vs. CONCAR model, does not change the estimated of σ_ϕ^2 significantly. In the model with the DIST prior, on the other hand, the estimate of σ_α^2 is reduced drastically from 0.75 to 0.10 (Table 3). Results for the second data set generated under Model (2) are very similar to those for the first data set (Table 2 and 3).

In both data sets the models therefore identify a CAR spatial structure of the data, even thought the variance-covariance matrix in the simulation was not strictly speaking CAR.

6.2 Papua New Guinea National Nutrition Survey

The spatial models discussed above were applied to the nutritional indicators height-for-age (HAZ) and weight-for-height (WHZ) and the results confirmed earlier findings by Heywood et al. (1988) and Smith et al. (1993) of regional differences in child growth and nutrition in PNG (Figures 2).

In the case of the weight-for-height (WHZ) data there is strong evidence for a neighbour-based structure in the data. The CAR model shows the best overall fit, with the CONCAR model being second best (Table 4), while all other models show very similar fits. As the model choice criterion D_k generally agree for different values of k in our simulated data sets, only values of D_1 are displayed in Table 4. Further evidence for the existence of a CAR structure in WHZ is shown in the estimates for the different variance components. In the case of the CAR prior the estimates of the variances show that by including independent area factors into the model the spatial variance $\widehat{\sigma}_\phi^2$ is only marginally reduced: from 0.379 to 0.293, and the non-spatial variance is very small $\left(\widehat{\sigma}_\alpha^2 = 0.018\right)$. Different results are obtained under the DIST prior: the spatial variance in the CONDIST model is much smaller $\left(\widehat{\sigma}_\phi^2 = 0.310\right)$ than in the DIST model $\left(\widehat{\sigma}_\phi^2 = 0.797\right)$ and there is considerable non-spatial variance $\left(\widehat{\sigma}_\alpha^2 = 0.106\right)$ (Table 5). Based on (4), this shows that the spatially structured variation dominates in the CONCAR model $\left(\widehat{\sigma}_\phi^2 = 0.050 \text{ vs. } \widehat{\sigma}_\alpha^2 = 0.018\right)$, while the unstructured variation dominates in the CONDIST model $\left(\widehat{\sigma}_\phi^2 = 0.064 \text{ vs. } \widehat{\sigma}_\alpha^2 = 0.104\right)$.

The results for the height-for-age nutritional score are less clear. The CAR model shows again the lowest D_1 value (Table 4). The fit of the IND model is close to the CAR model. As the CAR and IND models have very similar values of the model choice criterion used, it may not be possible to decide based on that criterion only which of the two models characterises the spatial structure of the data best. Based on the D_1 score the CONCAR model is not the best model, however it offers the possibility to separate the two sources of variation as it includes spatially structured as well as independent area

effects. If the CAR prior describes the data better than the IND one, we would expect to find a significant amount of variation σ_ϕ^2 in spatial area effect in the CONCAR model. The estimate of the spatial variance σ_ϕ^2 is not strongly reduced in the CONCAR as compared to the CAR model (0.478 to 0.397), but the non-spatial variance is much smaller in CONCAR than in the IND model (See also Figure 3). It can therefore be concluded that the neighbour-based correlations described by the CAR model most adequately characterise the spatial pattern in linear growth (HAZ) in PNG children as well as in the increase in weight (WHZ) (Figure 4).

7 Discussion and extensions

This study shows that Bayesian hierarchical models can be successfully applied to model spatially structured nutritional indicators. The models employed constitute powerful tools to investigate the amount and the nature of spatial variation in nutritional data sets, even though model choice is difficult due to the fact that the variation in nutritional scores is strongly dominated by individual variation. The possibility to model different forms of spatial heterogeneity through the choice of different spatial priors together with a combination of spatial and independent area effects nevertheless allows an in depth examination of spatial structure in data.

While the models presented here can accurately describe spatial variation and separate spatial from non-spatial variation, they do not give any explanation of the origin of that variation. In order to investigate what causes the observed spatial patterns, an extension of the above models to include covariate information is necessary.

A further limitation of the methodology presented above and the conventional ways of analysing nutritional data in general is that the nutritional indicators are only analysed univariately. Child growth is a very complex phenomenon influenced by many factors, with children growing in height and weight simultaneously as they develop. The above analyses show that complexity clearly; regions display distinct patterns of growth, ranging from areas where children are small and light, others where they are light but tall or small but heavy, to areas where they are both tall and heavy (Figure 4). These complex patterns of child growth and nutrition can only be properly analysed if two or more nutritional indicators are considered together. A further extension of the models is therefore needed in order to analyse the complex bi/multivariate patterns present in child growth data.

Acknowledgements: We would like to thank Tom Smith for supervising this work and discussing many aspects of the NNS data and the relevance of

models and results. Many thank also to Marcel Tanner for his encouragement and support. This work was supported by Swiss National Science Foundation grant 3200-049809.96.

References

[1] Allen et al. (1994) Agricultural Systems of Papua New Guinea Working Papers 1- 23, Department of Human Geography, RSAPS, Australian National University, Canberra

[2] Bellamy, J.A. (1986). Papua New Guinea Inventory of Natural Resources, Population Distribution and Land Use, Institute of Biological Resources, CSIRO, Canberra.

[3] Bernadinelli, L. and Montomoli, C. (1992). Empirical Bayes versus fully Bayesian analysis of geographical variation in disease risk. Statistics in Medicine, 11, 983-1007.

[4] Bernadinelli, L., Clayton, D.G. and Montomoli, C. (1995). Bayesian estimation of disease maps: how important are priors?. Statistics in Medicine, 21/22, 2411-2432.

[5] Besag, J. (1974). Spatial interaction and the statistical analysis of lattice systems (with discussion), Journal of the Royal Statistical Society B, 36, 192-236.

[6] Besag, J. and Mollié, A. (1989).Bayesian Mapping of mortality rates. Bulletin of the International Statistical Institute, 53, 127-128.

[7] Besag, J., Green, P., Higdon, D., and Mengersen, K. (1995). Bayesian computation and stochastic systems (with discussion). Statistical Science, 10, 3-66.

[8] Clayton, D.G. and Kaldor, J. (1989). Empirical Bayes estimates of age-standardized relative risks for use in Disease Mapping, Biometrics, 43, 671-681.

[9] Cole, T.J. (1992). Smoothing reference centile curves: the LMS method and penalized likelihood. Statistics in Medicine, 11, 1305-13, 19.

[10] Cressie, N.A.C, (1993). Statistics for spatial data. Wiley and Son, New York.

[11] Gelfand, A.E. and Gosh, S.K. (1998). Model choice: A minimum posterior predictive loss approach. Biometrics, 58, 1-11.

[12] Gelfand, A.E., Gosh, S.K., Knight, J. and Sirmans, C.F. (1998). Spatio-temporal modelling of residential sales markets, Journal of Business and Economic Statistics, (to appear).

[13] Gelfand, A.E. And Smith, A.F.M. (1990). Sampling based approaches to calculating marginal densities. Journal of American Statistical Association, 85, 398-409.

[14] Gelman, A. and Rubin, D.B. (1992). Inference from iterative simulations using multiple sequences (with discussion). Statistical Science, 7, 457-511.

[15] Haining, R. (1990). Spatial Data Analysis in the Social and Environmental Sciences. Cambridge University Press, Cambridge.

[16] Heywood, P.F., Singleton, N. and Ross, J. (1988). Nutritional status of young children - The, 1982/3 National Nutrition Survey. Papua New Guinea Medical Journal, 31, 91-101.

[17] Lee and Schachter (1980), International Journal of Computer and Information Sciences, Vol. 9, No. 3, 1980, 2, 19-242

[18] Mollié, A. (1996). Bayesian mapping of disease. In: Gilks, W.R., Richardson, S. and Spiegelhalter, D.J. (eds.), Markov Chain Monte Carlo in Practice, Chapman & Hall, London, p. 359-379.

[19] Smith, T.A., Earland, J., Bhatia, K., Heywood, P. and Singleton, N. (1993). Linear Growth of Children in Papua New Guinea in Relation to Dietary, Enviromental and Genetic Factors. Ecology of Food and Nutrition, 31, 1-25.

[20] Vounatsou, P., Smith, T. and Gelfand, A.E. (1998). Spatial modelling of multinomial data with latent structure; an application to geographical mapping of human gene and haplotype frequencies. Technical report (TR 9815), Department of Statistics, University of Connecticut

[21] Waterlow, J.A. (1992). Protein Energy Malnutrition. Edward Arnold, London.

Annexes

Table 1: Definitions of all models fitted

		Weights w
Model (1)		
IND	$\theta_{iq} = \mu + \alpha_i$	
Model (2)		
EX	$\theta_{iq} = \mu + \phi_i$	$w_{ij} = 1\ \forall i, j$
CAR	$\theta_{iq} = \mu + \phi_i$	$w_{ij} = 1$ if i, j neighbours, $w_{ij} = 0$ otherwise
DIST	$\theta_{iq} = \mu + \phi_i$	$w_{ij} = 1/d_{ij}\ \forall i \neq j$, $w_{ij} = 0\ \forall i = j$
CARDIST	$\theta_{iq} = \mu + \phi_i$	$w_{ij} = 1/d_{ij}$ if i, j neighbours, $w_{ij} = 0$ otherwise
Model (3)		
CONCAR	$\theta_{iq} = \mu + \alpha_i + \phi_i$	$w_{ij} = 1$ if i, j neighbours, $w_{ij} = 0$ otherwise
CONDIST	$\theta_{iq} = \mu + \alpha_i + \phi_i$	$w_{ij} = 1/d_{ij}\ \forall i \neq j$, $w_{ij} = 0\ \forall i = j$

Table 2: Penalty term (P), Goodness-of-fit measures (G) and overall model choice criterion (D_k) with $k = 1, 3, \infty$. Simulated data sets.

Simulated data set	Model M	P	G	D_1	D_3	D_∞
Independent						
Model (1)	IND	404.2	306.1	557.3	633.8	710.4
$z_{ij} \sim N(\alpha_\iota, 2)$	EX	405.8	307.7	559.7	636.7	713.6
$\alpha_\iota \sim N(0,1)$	CAR	412.4	311.5	568.2	646.1	724.0
	DIST	407.5	307.7	561.3	638.2	715.2
	CONCAR	401.9	304.8	554.3	630.5	706.8
	CONDIST	401.6	306.9	555.0	631.8	708.5
Spatial 1						
Model (2)	IND	424.8	325.2	587.4	668.7	750.0
$z_{ij} \sim N(\phi_\iota, 2)$	EX	426.7	329.1	591.3	673.6	755.8
$\phi_\iota \sim MVN(\mathbf{0}, \mathbf{1})$	CAR	407.5	336.7	575.9	660.1	744.3
$\mu = 0$	DIST	419.2	325.7	582.1	663.5	744.9
	CONCAR	405.9	331.3	571.6	626.8	737.2
	CONDIST	422.7	325.0	585.3	639.4	747.8
Spatial 2						
Model (2)	IND	426.7	338.9	596.2	680.9	765.6
$z_{ij} \sim N(\mu + \phi_\iota, 2)$	EX	466.0	396.4	664.2	763.3	862.4
$\phi_\iota \sim MVN(\mathbf{0}, \mathbf{1})$	CAR	414.4	351.9	590.3	678.3	766.3
$\mu = 0.25$	DIST	423.7	341.5	594.5	679.9	765.3
	CONCAR	412.4	343.3	584.1	669.9	755.7
	CONDIST	422.6	338.4	591.8	676.4	761.0

Table 3: Posterior medians and 95% confidence intervals for parameters under Model (1) - (3) and different spatial priors (see Table 1 for description). $\widetilde{w}_+$: average weight per area (see (4)). Simulated data.

Simulated data set	Model M	σ^2_ϕ	$\sigma^2_\phi / \widetilde{w}_+$	σ^2_α	σ^2_z	μ
Independent						
Model (1)	IND			1.09 (0.36,3.07)	1.75 (1.23,2.46)	0.07 (-0.68,0.81)
$z_{ij} \sim N(\alpha_\iota, 2)$	EX	34.91 (7.66,110.51)	1.00 (0.22,3.16)		1.77 (1.30,2.70)	0.07 (-0.27,0.36)
$\alpha_\iota \sim N(0,1)$	CAR	4.62 (0.82,15.84)	0.98 (0.17,3.38)		1.79 (1.31,2.75)	0.06 (-0.28,0.36)
	DIST	14.93 (3.26,47.61)	1.07 (0.23,3.40)		1.77 (1.30,2.71)	0.07 (-0.27,0.36)
	CONCAR	0.03 (0.00,7.74)	0.01 (0.00,1.65)	1.03 (0.01,3.48)	1.75 (1.25,2.66)	0.07 (-0.55,0.79)
	CONDIST	0.05 (0.00,28.20)	0.00 (0.00,2.01)	1.00 (0.01,3.42)	1.75 (1.25,2.64)	0.07 (-0.59,0.80)
Spatial 1						
Model (2)	IND			0.91 (0.19,2.68)	1.85 (1.28, 256)	0.09 (-0.57,0.77)
$z_{ij} \sim N(\phi_\iota, 2)$	EX	29.41 (0.74,96.03)	0.84 (0.02,2.74)		1.87 (1.35,2.90)	0.09 (-0.27,0.42)
$\phi_\iota \sim MVN(\mathbf{0}, \mathbf{1})$	CAR	1.38 (0.26,4.95)	0.29 (0.02,0.35)		1.84 (1.34,2.73)	0.06 (-0.29,0.38)
	DIST	10.54 (2.55,34.13)	0.75(0.18,2.44)		1.84 (1.33,2.64)	0.08 (-0.27,0.41)
	CONCAR	1.21 (0.14,5.49)	0.26 (0.03,1.17)	0.06 (0.00,1.38)	1.82 (1.24,2.53)	0.05 (-0.34,0.51)
	CONDIST	1.33 (0.00,28.59)	0.10 (0.00,2.04)	0.56 (0.00,2.67)	1.84 (1.26,2.60)	0.08 (-0.55,0.65)
Spatial 2						
Model (2)	IND			0.57 (0.03,1.83)	1.89 (1.24,2.69)	0.24 (-0.30,0.85)
$z_{ij} \sim N(\mu + \phi_\iota, 2)$	EX	8.55 (0.00,60.28)	0.24 (0.00,1.72)		2.10 (1.32,3.28)	0.23 (-0.15,0.52)
$\phi_\iota \sim MVN(\mathbf{0}, \mathbf{1})$	CAR	0.91 (0.10,4.13)	0.19 (0.02,0.88)		1.89 (1.35,2.89)	0.26 (-0.09,0.59)
$\mu = 0.25$	DIST	6.50 (0.14,23.90)	0.46 (0.01,1.71)		1.89 (1.31,2.90)	0.24 (-0.10,0.56)
	CONCAR	0.73 (0.00,4.66)	0.16 (0.00,0.99)	0.07 (0.00,1.72)	1.86 (1.28,2.78)	0.25 (-0.15,0.72)
	CONDIST	0.11 (0.00,20.28)	0.01 (0.00,1.45)	0.45 (0.00,2.12)	1.88 (1.28,2.73)	0.24 (-0.31,0.88)

Table 4: penalty term (P), Goodness-of-fit measures (G) and overall model choice criterion (D_k) with $k = 1$. Spatial models for nutritional data sets.

Data set	Model M	P	G	D_1
WHZ	IND	16'930	16'183	25'022
	EX	16'930	16'175	25'017
	CAR	16'865	16'243	24'897
	DIST	16'929	16'178	25'018
	CARDIST	16'871	16'297	25'020
	CONCAR	16'873	16'238	24'992
	CONDIST	16'935	16'177	25'024
HAZ	IND	17'553	16'790	25'947
	EX	17'558	16'796	25'956
	CAR	17'521	16'845	25'944
	DIST	17'554	16'796	25'952
	CARDIST	17'517	16'882	25'958
	CONCAR	17'532	16'843	25'954
	CONDIST	17'559	16'811	25'965

Table 5: Posterior medians and 95% confidence intervals for parameter under Model (1) - (3) and different spatial priors (see Table 1 for description). $\widetilde{w}_+$: average weight per area (see (4)). Spatial models for nutritionnal data.

Data set	σ^2_ϕ	$\sigma^2_\phi/\tilde{w}_+$	σ^2_α	σ^2_z
WHZ				
IND			0.191 (0.148, 0.240)	0.815 (0.788, 0.846)
EX	109.94 (83.92, 138.64)	0.192 (0.146, 0.242)		0.815 (0.780, 0.837)
CAR	0.379 (0.285, 0.502)	0.065 (0.049, 0.086)		0.815 (0.780, 0.838)
DIST	0.797 (0.614, 1.016)	0.164 (0.127, 0.209)		0.815 (0.780, 0.837)
CARDIST	0.016 (0.012, 0.022)	0.043 (0.032, 0.059)		0.816 (0.782, 0.839)
CONCAR	0.293 (0.172, 0.442)	0.050 (0.029, 0.075)	0.018 (0.002, 0.050)	0.815 (0.789, 0.841)
CONDIST	0.310 (0.044, 0.683)	0.064 (0.009, 0.141)	0.106 (0.031, 0.199)	0.815 (0.789, 0.841)
WAZ				
IND			0.167 (0.121, 0.217)	0.845 (0.817, 0.872)
EX	95.65 (73.19, 120.05)	0.167 (0.128, 0.209)		0.845 (0.808, 0.868)
CAR	0.478 (0.373, 0.621)	0.082 (0.064, 0.106)		0.846 (0.808, 0.869)
DIST	0.701 (0.539, 0.890)	0.144 (0.111, 0.183)		0.845 (0.809, 0.869)
CARDIST	0.020 (0.016, 0.028)	0.055 (0.043, 0.075)		0.847 (0.809, 0.870)
CONCAR	0.397 (0.244, 0.572)	0.068 (0.042, 0.098)	0.016 (0.002, 0.052)	0.846 (0.820, 0.872)
CONDIST	0.260 (0.022, 0.739)	0.054 (0.005, 0.152)	0.098 (0.006, 0.183)	0.846 (0.820, 0.874)

Figures

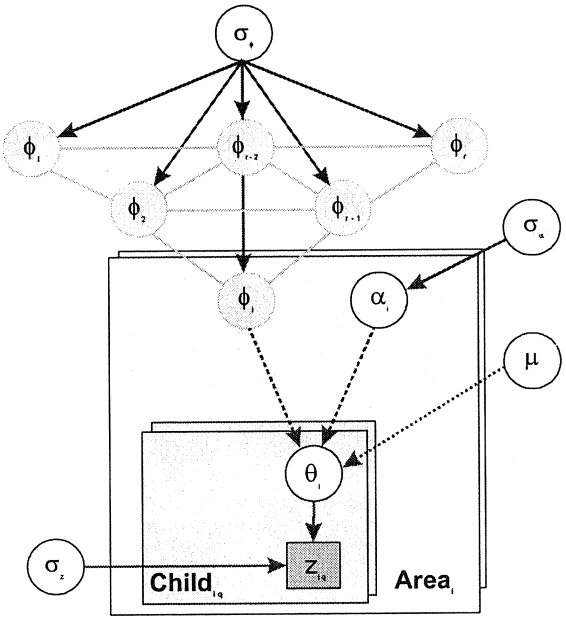


Figure 1: graphical representation of Model (3). Solid lines indicate stochastic dependence, dashed lines deterministic dependence.

FIGURE 2: Z-scores for 575 areas included in nutritional analyses

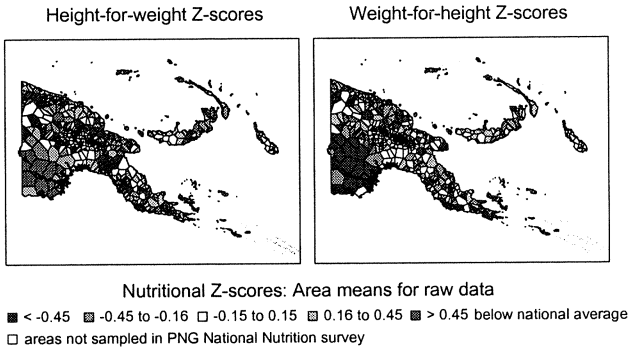


FIGURE 3: Estimated area effects for weight-for-height Z-scores

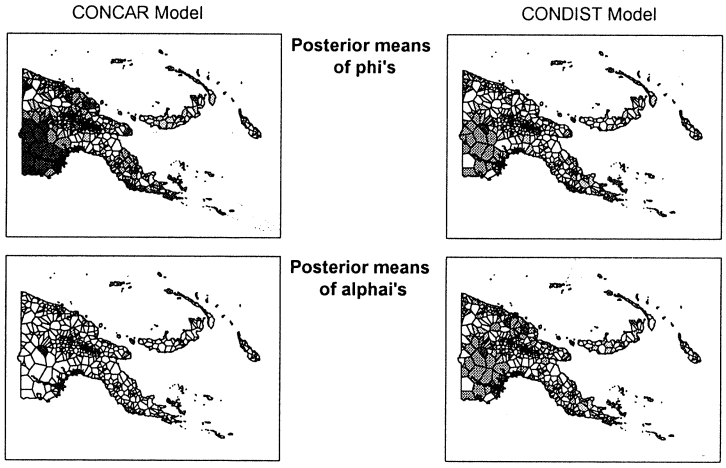
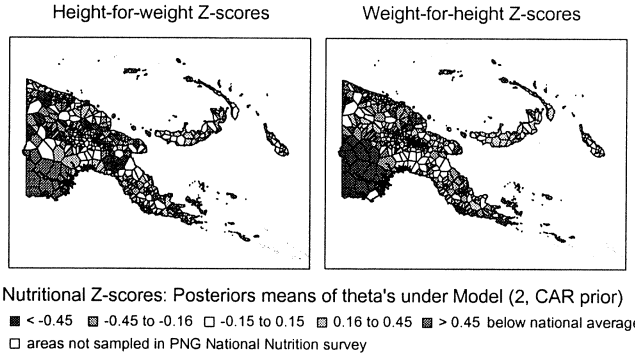


FIGURE 4: Spatial patterns in nutritional indicators



Student (1998), Vol.3. No.1, 19-26

Sample Size Determination in Clinical Trials

Hamid Pezeshk and John Gittins

Department of Statistics, Oxford University, England

Abstract: A decision theoretic framework for sample size determination for a clinical trial is presented in which the final decision whether to use the new treatment is taken by potential users and their medical advisers on the basis of the strength of the evidence provided by the trial. This is more realistic than the usual decision theoretic assumption that both the sample size and terminal decisions are taken by a single rational decision maker.

Key words: Clinical trials, sample size determination, Bayesian approach.

1 Introduction and Executive Summary

The problem of sample size determination is a well known problem both in quality control, and in clinical trials. In the frequentist approach sample sizes are usually determined either from power calculations or from formulae based on confidence interval widths (see, for example Singer (1997)). In the decision theoretic approach the sample size is determined by balancing expected costs and benefits, using a Bayesian prior distribution for the unknown parameters.

The first paper along these lines was by Grundy *et al* (1956); the methodology was set out in detail by Raiffa and Schlaifer (1961), and has been described recently by Lindley (1997). A major problem in following this approach is that the ultimate decision on whether or not to use the new treatment is taken by a large number of patients and their advisers, and does not depend on the outcome of the trial in any clear cut way. To find a utility function which would make the Raiffa and Schlaifer paradigm of a single rational decision maker convincing is not easy, and its use is not

common. Bayesian methods without utility functions are, in contrast, much more firmly established, and are reviewed by Adcock (1997).

In this paper we return to the full decision theoretic analysis of Raiffa and Schlaifer, with the modification that instead of a utility-maximising terminal decision a plausible model is assumed for the way patients and their medical advisers respond to the evidence from a trial. Our assumptions are somewhat oversimplified, the aim being to show that the approach is computationally tractable and gives sensible answers. Further work with more realistic assumptions is in progress.

The data are assumed to come from a normal distribution, $N(\theta, \sigma^2)$, with unknown parameter, θ (this might be the true cure rate of a new treatment), and known variance σ^2. A normal prior distribution is assumed for θ. The objective function is based on the number of subsequent users of the new treatment, the expected benefit of using the new treatment, and the cost of conducting the trial.

In section 2 we state our sample size problem and introduce the notation. In this section we give a brief background and establish our objective function.

In section 3 we provide an exact solution for the problem in the case where the number of subsequent users is proportional to the number of standard deviations by which the posterior mean of θ exceeds the performance of the current treatment. An explicit expression is obtained for the net expected benefit of the trial, and for the minimizing sample size. We show that the optimal sample size is an increasing function of the prior uncertainty of θ, and a decreasing function of cost/benefit ratio.

A discussion about the limitation of the proposed model along with a brief description of the way forward are presented in section 4.

2 The Sample Size Problem

Suppose X_i (i=1,2,...) is the clinical outcome on some appropriate scale for the ith patient using the new treatment. Assume that X_i has a normal density which is $N(\theta, \sigma^2)$, where σ^2 is known from previous experience, and θ is unknown and has a prior density which is $N(\mu, \tau^2)$. We can think of θ as the cure rate of a new treatment and θ_0 (known) as the cure rate of the treatment in current use. For every patient treated by the new treatment, there is a benefit $b(\theta - \theta_0)$, where b is known. Let: cn be the cost of doing an experiment on a sample of size n; μ' be the expectation of the posterior density of θ after doing the experiment; $m(\overline{x}_n)$ be the number of subsequent users of the new treatment. The question is what sample size maximizes the net expected benefit?

The objective function is the total benefit from the patients using the treatment minus the cost of the trial. It can be written as follows

$$r(n) = \int_{\overline{x}_n} \int_{\theta} m(\overline{x}_n) b(\theta - \theta_0) d\pi^n(\theta|\overline{x}_n) f(\overline{x}_n) d\overline{x}_n - cn, \tag{1}$$

where $\pi^n(\theta|\overline{x}_n)$ is the posterior distribution of θ, given the sufficient statistic $\overline{x}_n$. We are looking for n^* which maximizes $r(n)$.

It is well known that if $x_i|\theta \sim N(\theta, \sigma^2)$ and $\theta \sim N(\mu, \tau^2)$ then $\theta|\overline{x}_n \sim N(\mu', \tau'^2)$ where

$$\mu' = \frac{\sigma^2\mu + n\tau^2\overline{x}_n}{\sigma^2 + n\tau^2}, \qquad \tau'^2 = \frac{\sigma^2\tau^2}{\sigma^2 + n\tau^2}.$$

By calculating the inner integral in (1) we arrive at the following representation of $r(n)$

$$r(n) = \int_{\overline{x}_n} m(\overline{x}_n) b(\mu' - \theta_0) f(\overline{x}_n) d\overline{x}_n - cn \tag{2}$$

Putting the density function of $\overline{X}_n$, which is $N(\mu, \frac{n\tau^2+\sigma^2}{n})$, in (2) we can see that

$$r(n) = \int_{\overline{x}_n} m(\overline{x}_n) b(\mu' - \theta_0) \sqrt{\frac{n}{2\pi(n\tau^2 + \sigma^2)}} e^{\frac{-n}{2(n\tau^2+\sigma^2)}(\overline{x}_n - \mu)^2} d\overline{x}_n - cn,$$

In section 3 we consider the number of subsequent users of the new treatment as a linear increasing function of μ' and calculate the optimal sample size by maximizing the expected benefit minus the expected cost.

3 Sample Size Determination

We shall suppose that

$$m(\overline{x}_n) = k_1(\frac{\mu' - \theta_0}{\tau'}) + k_2,$$

where k_1 and k_2 are constants and $k_1 > 0$. The term in k_1 means that m increases with the strength of the posterior evidence that the new treatment is better than the current treatment. A high value of k_1 corresponds to a large number of potential users. The constant k_2 has no influence on the size of the trial, as it represents a fixed benefit, independent of the outcome. This expression for $m(\overline{x}_n)$ is unrealistic in that $m \to \pm\infty$ as $\mu' \to \pm\infty$. Its main advantage is the simplicity of the calculations to which it leads for the illustrative purposes of this paper.

The objective function can be written

$$r(n) = \int_{\overline{x}_n} [k_1(\frac{\mu' - \theta_0}{\tau'}) + k_2] b(\frac{\mu' - \theta_0}{\sigma}) f(\overline{x}_n) d\overline{x}_n - cn. \tag{3}$$

Substituting for μ' and τ' in (3), and since $E(\overline{X}_n) = \mu$, and $Var(\overline{X}_n) = \frac{\sigma^2}{n} + \tau^2$, it follows from (3) that

$$r(n) = \frac{k_1 b}{\sigma\sqrt{\frac{\sigma^2\tau^2}{\sigma^2+n\tau^2}}}[(\mu-\theta_0)^2 + (\frac{\tau^2}{\frac{\sigma^2}{n}+\tau^2})^2(\frac{\sigma^2}{n}+\tau^2)] + \frac{k_2 b}{\sigma}(\mu-\theta_0) - cn.$$

Let us now define

$$\Theta = \frac{\mu-\theta_0}{\sigma}, \qquad T = \frac{\tau}{\sigma},$$

so that

$$r(n) = k_1 b(T^{-2}+n)^{1/2}[\Theta^2 + nT^2(T^{-2}+n)^{-1}] + k_2 b\Theta - cn. \tag{4}$$

Since the aim of this paper is illustrative, rather than to set out a detailed practical procedure, we suppose n is a continuous variable. Thus to find the maximum of $r(n)$ let $\frac{dr(n)}{dn} = 0$. We have

$$\begin{aligned}\frac{dr(n)}{n} = \; & k_1 b[1/2(T^{-2}+n)^{-1/2}(\Theta^2 + nT^2(T^{-2}+n)^{-1}) \\ & +(T^{-2}+n)^{1/2}(T^2(T^{-2}+n)^{-1} - nT^2(T^{-2}+n)^{-2}) - \frac{c}{k_1 b}] = 0.\end{aligned}$$

If we define $N = (T^{-2}+n)$, $H = N^{-\frac{1}{2}}$, and $C = \frac{c}{k_1 b}$ it follows that

$$\frac{dr(n)}{dn} = 0 \quad \textit{if and only if} \quad H^3 + (\Theta^2 + T^2)H - 2C = 0. \tag{5}$$

We now use the standard method for solving the cubic equation $x^3 + bx^2 + cx + d = 0$. Define:

$$f = \frac{b}{3} = 0, \quad g = \frac{c}{3} - f^2 = \frac{(\Theta^2+T^2)}{3}, \quad h = \frac{d}{2} + f^3 - \frac{fc}{2} = -C,$$

$$j = g^3 + h^2 = \frac{(\Theta^2+T^2)^3}{27} + C^2.$$

Since $j > 0$ equation (5) has one real root. This can be found as follows.

Let $k = \sqrt{j}$, $l_1 = (-h+k)^{1/3}$, and $l_2 = (-h-k)^{1/3}$. Then the real root H^* of the equation is equal to $l_1 + l_2$, so

$$H^* = \left\{C + \sqrt{\frac{(\Theta^2+T^2)^3}{27} + C^2}\right\}^{1/3} + \left\{C - \sqrt{\frac{(\Theta^2+T^2)^3}{27} + C^2}\right\}^{1/3}.$$

We have

$$\begin{aligned}\frac{d^2 r(n)}{dn^2} &= \left(3H^2 H' + \left(\Theta^2+T^2\right)H'\right)\left(\frac{k_1 b}{2}\right) \\ &= \left\{3H^2\left(-1/2N^{-3/2}\right) + \left(\Theta^2+T^2\right)\left(-1/2N^{-3/2}\right)\right\}\left(\frac{k_1 b}{2}\right) \\ &< 0.\end{aligned}$$

It follows that H^* gives the maximum value of $r(n)$. Note that n^* can be found from the equation $n^* = H^{*-2} - T^{-2}$.

Figure 1 illustrates the variation of the scaled objective function $R(n) = \frac{r(n)}{k_1 b} - \frac{k_2}{k_1}\Theta$. The optimal size of trial n^* for this case is 23.

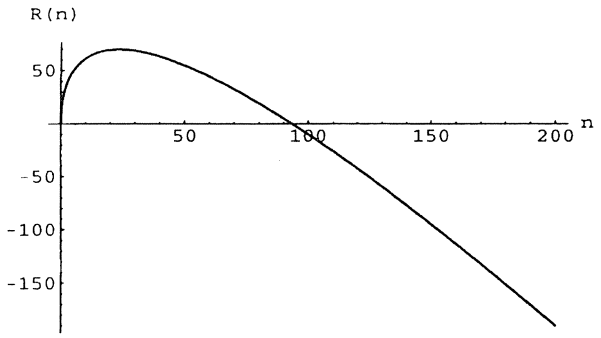


Figure 1: Net expected benefit when $\Theta = 2, T = 5$ and $C = 3$.

3.1 The Parameters Determining Optimal Sample Size

From equation (5) it is clear that the optimal value H^* of H depends only on the parameters C, T, Θ, which may be interpreted, respectively, as follows:

C is a cost/benefit ratio,

T is a measure of our prior ignorance of θ,

Θ is a measure of our prior expectation of the superiority of the new treatment.

Calculations show that H^* is an increasing function of C and a decreasing function of both T and Θ, from which it follows that n^* is a decreasing function of C and an increasing function of both T and Θ.

In figure 2 the logarithm of the optimal sample size n^* is plotted as a function of T and C. The optimal sample size increases when prior uncertainty about θ increases, and decreases when the cost/benefit ratio of using the new treatment increases.

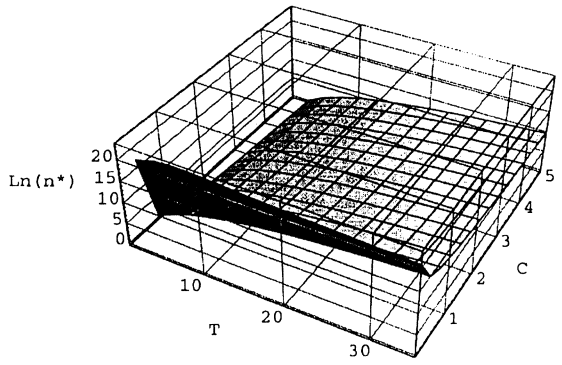


Figure 2: Logarithm of the optimal sample size n^*, as a function of cost/benefit ratio C, and the parameter of uncertainty T.

4 Discussion

That n^* should decrease as the cost/benefit ratio increases, and that it should increase as prior ignorance increases, is just what common sense suggests. That it should also be large when we expect the new treatment to be much better than the current one at first looks odd. In that case why do we need a trial at all ? The reason for this is the factor $\frac{\mu'-\theta_0}{\tau'}$ in the objective function (3). For instance, if it turns out that $\frac{\mu-\theta_0}{\tau} = 6$ and we choose the sample size big enough to ensure that $\tau' = \frac{1}{2}\tau$, then $E(\frac{\mu'-\theta_0}{\tau'})$ becomes approximately 12, with a consequent increase in the objective function, $r(n)$. On the other hand a smaller value of $\frac{\mu-\theta_0}{\tau}$ leads to a smaller increase in $r(n)$, and consequently a smaller value of n^*. A change in the function $m(\overline{x}_n)$, so that it reaches a maximum value as μ' increases, rather than tending to infinity, should correct this anomaly.

While it is realistic to suppose that the decision to use the new treatment depends on the certainty with which it is believed to improve on the current treatment, our assumption of the nature of this relationship is too crude. Moreover the usage of the new treatment depends on the magnitude of the estimated difference, as well as on our confidence that it is positive. A model which takes these effects into account is now being worked on. It might also be desirable to include the possibility of a follow-up trial.

In due course it is hoped to model the situation when several new compounds are available for testing the same indication, so as to be able to

formulate a sensible policy for testing them sequentially, or possibly to some extent in parallel.

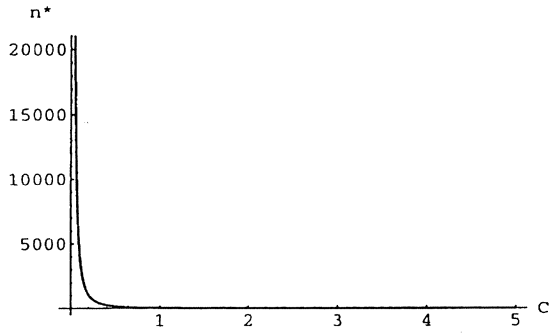


Figure 3: Optimal size, n^*, as a function of cost/benefit ratio, C: clearly n^* is a decreasing function of C; $T = 3$.

As shown by the graphs, for small values of C the optimal sample size is greater than zero unless the measure of uncertainty, T, is small. On the other hand if the cost/benefit ratio is high and the prior evidence does not show a significant difference between the cure rates of the new treatment and the existing one it is not worth doing any trial.

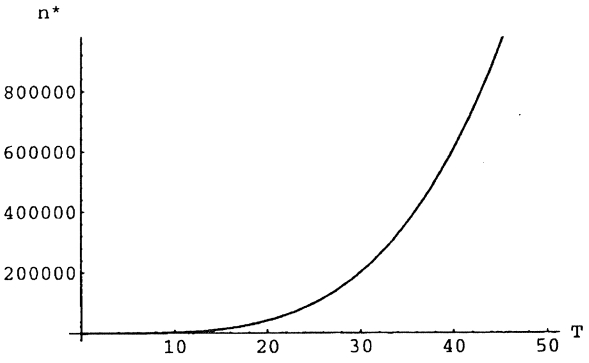


Figure 4: Optimal sample size, n^*, as a function of the parameter of uncertainty, T; $C = 1$.

Acknowledgements: This project was supported by the Ministry of Culture and Higher Education of the Islamic Republic of Iran under a scholarship for studying abroad. The authors would like to thank an anonymous referee for her/his insightful comments which led to a much improved paper.

References

[1] Adcock, C. J. (1997). Sample Size Determination: A Review. *Statistician* **46,** 261–283.

[2] Grundy, P. M., Healy, M. J. R. and Rees, D. H. (1956). Economic Choice of the Amount of Experimentation. *J. R. Statis. Soc. A* **18,** 32–48.

[3] Lindley, D. V. (1997). The Choice of Sample Size. *The Statistician* **46,** 129–138.

[4] Raiffa, H. and Schlaifer, R. (1961). *Applied Statistical Decision theory.* Division of Research Harvard Busines.

[5] Singer, J. (1997). Estimating Sample Size for Continuous Outcomes, Comparing more than two Parallel groups with Unequal Sizes. *Statistics in Medicine* **16,** 2805–2811.

Student (1998), Vol.3. No.1, 27-40

Estimating a probability distribution with given moments from empirical data

Daniel Costa

Nestec Ltd, Management Services, Vevey, Switzerland

Edward A. Silver

Faculty of Management, University of Calgary, Canada

Abstract: We consider the problem of estimating the probability distribution of a discrete variable knowing its first m moments. Given a sample of observations the problem is to find the distribution which best matches the empirical distribution while meeting the known moments. Two objective functions, namely the least square and the maximum likelihood functions, are proposed to measure the resemblance between the empirical distribution and the distribution to be estimated. The resulting convex programming problems are solved using, respectively, a modified simplex algorithm and an adaptation of the Franck and Wolfe algorithm.

Key words: Density estimation, moments, least square, maximum likelihood, convex programming.

1 Introduction

We are interested in estimating a probability distribution from empirical data when one knows one or more exact moments. There are several applications where such a situation occurs. An example can be found in Silver et al. (1998) where the mean values of order statistics, conditional upon knowing their sum, can be found easily, contrary to their conditional distribution function which cannot be computed analytically in the general case. Another example is in Strijbosch and Heuts (1994) where in an inventory context the lead time and the demand per unit time are independent random variables so that it is easy to compute the mean and variance of the total demand

in a lead time, but it is difficult to find its distribution analytically. More generally, consider any situation where one is interested in the sum of a number of independent random variables with known distributions. It is easy to compute the moments of the sum from the moments of the individual variables. However, the complete distribution is in general very difficult to obtain analytically and one may resort to simulation, thus generating empirical data.

In this study the objective is to find the probability distribution of a discrete variable which is the closest to the observed empirical distribution, in a sense to be defined, and which meets the known moments. Note that this way of proceeding is different from the standard statistical approach where the goal is to find estimates of the moments given a sample of observations and possibly knowing the type of the underlying probability distribution.

This paper is organized as follows. In the next section we lay out the notation and the assumptions to be used and then we formulate the problem of interest. Two different objective functions are chosen, namely the sum of the squares and the likelihood function. Section 4 is devoted to the presentation of two solution methods. These methods are illustrated in section 5 through a simple example. The paper concludes with some summary comments.

2 Notation, assumptions and problem definition

The results presented hereafter are based on the following notation and assumptions.

- X : discrete variable taking on n different values x_i.
- p_i : probability of X taking on the value x_i ($p_i = Prob(X = x_i)$).
- μ_r : moment of order r of the distribution ($\mu_r = \sum_{i=1}^{n} p_i \cdot x_i^r$).
- H : size of the sample data of X.
- h_i : number of times the value x_i is observed ($h_1 + \ldots + h_n = H$).

We assume that the first m moments of X are known and we observe a sample from X of size H. The probability distribution $P = (p_1, \ldots, p_n)$ is unknown and must be estimated taking account of the constraints below:

- The first m moments of P must be equal to the exact moments μ_r ($r = 1, \ldots, m$).

- P must reflect the observed sample and thus be as close as possible to the empirical probability distribution $\widehat{P}_h = (\frac{h_1}{H}, \ldots, \frac{h_n}{H})$.

This amounts to solving the following constrained optimization problem:

$$\begin{array}{ll} \underset{p_1,\ldots,p_n}{\text{Optimize}} & D(p_1,\ldots,p_n) \\ \text{s.t.} & g_0(p_1,\ldots,p_n) = \sum_{i=1}^{n} p_i - 1 = 0 \\ & g_1(p_1,\ldots,p_n) = \sum_{i=1}^{n} p_i \cdot x_i - \mu_1 = 0 \\ & g_2(p_1,\ldots,p_n) = \sum_{i=1}^{n} p_i \cdot x_i^2 - \mu_2 = 0 \\ & \vdots \\ & g_m(p_1,\ldots,p_n) = \sum_{i=1}^{n} p_i \cdot x_i^m - \mu_m = 0 \\ & g_{m+1}(p_1,\ldots,p_n) = p_1 \geq 0 \\ & \vdots \\ & g_{m+n}(p_1,\ldots,p_n) = p_n \geq 0 \end{array} \tag{1}$$

where D is a function measuring either the distance or the similarity between the two distributions P and $\widehat{P}_h$. D is to be minimized if it represents a distance and maximised otherwise. Let S be the set of $\underline{p} = (p_1,\ldots,p_n)$ meeting constraints $g_j = 0$ $(j = 0,\ldots,m)$. To make sure that S is non empty, we will make the additional assumption that $n \geq m+1$ (see lemma 1).

Lemma 1: If the x_i's are all different and $n \geq m+1$ then there exists at least one vector $\underline{p} = (p_1,\ldots,p_n)$ meeting constraints $g_j = 0$ $(j = 0,\ldots,m)$.

Proof: Constraints $g_j = 0$ $(j = 0,\ldots,m)$ are linear and thus the $\underline{p}$'s in S must be such that:

$$A \cdot \underline{p} = \underline{b} \tag{2}$$

where A is a $(m+1) \times n$ matrix:

$$A = \begin{pmatrix} 1 & 1 & \cdots & 1 \\ x_1 & x_2 & & x_n \\ \vdots & \vdots & & \vdots \\ x_1^m & x_2^m & \cdots & x_n^m \end{pmatrix}$$

and $\underline{b}$ is a vector of dimension $m+1$:

$\underline{b} = (1, \mu_1, \mu_2, \ldots, \mu_m)^T$

Since $m+1 \leq n$, let A_{m+1} be the squared submatrix made up of the first $m+1$ columns of A. The Vandermonde formula (see Hofmann and Kunze, 1971):

$$det(A_{m+1}) = (-1)^{m+1} \prod_{i<j} (x_i - x_j)$$

reveals that the determinant of A_{m+1} is not nil provided that the x_i's are all different. Hence A_{m+1} is regular and $m+1 = rank(A_{m+1}) = rank(A)$.

It follows that for any value of $(p_{m+2}, \ldots, p_n)$, there exists a solution $\underline{p}_{m+1} = (p_1, \ldots, p_{m+1})$ to the system below:

$$A_{m+1} \cdot \underline{p}_{m+1} = \underline{\tilde{b}}$$
$$\text{where } \underline{\tilde{b}} = (1 - \sum_{i=m+2}^{n} p_i, \mu_1 - \sum_{i=m+2}^{n} p_i \cdot x_i, \ldots, \mu_m - \sum_{i=m+2}^{n} p_i \cdot x_i^m)^T$$

This means that S is a vector space of dimension $n - m - 1 \geq 0$ in $\Re^n$. ■

In the remainder we focus on two objective functions, namely the least square distance:

$$\text{Minimize } D_{LS}(p_1, \ldots, p_n) = \sum_{i=1}^{n} (p_i - \frac{h_i}{H})^2 \tag{3}$$

and the maximum likelihood function:

$$\text{Maximize } D(p_1, \ldots, p_n) = \prod_{i=1}^{n} p_i^{h_i}$$

which is equivalent to:

$$\text{Maximize } D_{ML}(p_1, \ldots, p_n) = \log D(p_1, \ldots, p_n) = \sum_{i=1}^{n} h_i \cdot \log p_i \tag{4}$$

Let $\widehat{P}_{LS}$ and $\widehat{P}_{ML}$ denote the distributions achieved by solving respectively the least square and the maximum likelihood versions of problem (1). It is important to point out that both functions define a convex programming problem. Indeed, D_{LS} and $-D_{ML}$ are two convex functions to be minimized (D_{ML} is concave) over a set of feasible distributions. This set is clearly convex since the problem is linearly constrained. Many efficient algorithms have been proposed in the literature to solve convex programming problems. For an overview, see Wismer and Chattergy (1978) for example. In this paper, we propose two different approaches for optimizing D_{LS} and D_{ML} respectively.

In a recent study, Ferson (1997) used a different approach to cope with the lack of information regarding the distribution function of a random variable. Instead of looking for the "closest" distribution function to the observed empirical distribution with the same moments, Ferson focused on displaying the full spectrum of all distributions meeting some underlying restrictions (e.g., mean, mode, quantile, variance).

A variety of techniques are proposed in Silverman (1986) to estimate a distribution function starting from sample data. However these techniques do not impose any restriction on the moments of the distribution to be estimated.

3 Two solution methods

3.1 The least square function

In this section we propose a modified simplex algorithm for finding a solution to the Karush-Kuhn-Tucker conditions for constrained optimization (Wismer and Chattergy,1978; Minoux, 1983; Rardin, 1998):

(a) $\frac{\partial D}{\partial p_i} - \sum_{j=0}^{m} \lambda_j \cdot \frac{\partial g_j}{\partial p_i} - \sum_{k=1}^{n} \eta_k \cdot \frac{\partial g_{m+k}}{\partial p_i} = 0 \quad \forall i = 1, \ldots, n$
(b) $g_j = 0 \quad and \quad g_{m+k} \geq 0 \quad \forall j = 0, \ldots, m \quad \forall k = 1, \ldots, n$
(c) $\eta_k \cdot g_{m+k} = 0 \quad \forall k = 1, \ldots, n$
(d) $\eta_k \geq 0 \quad \forall k = 1, \ldots, n$

In the field of convex programming, these conditions determine uniquely the optimal solution of a problem. Unfortunately the solution of the Karush-Kuhn-Tucker conditions is often as difficult as solving the original problem. Here this approach is applicable because the objective function

$$D_{LS}(p_1, \ldots, p_n) = \sum_{i=1}^{n} (p_i - \frac{h_i}{H})^2$$

is quadratic. The distribution of interest $\widehat{P}_{LS} = (p_1, \ldots, p_n)$ is an optimal solution of problem (1) if and only if there exist $m + n + 1$ numbers $\lambda_0, \ldots, \lambda_m, \eta_1, \ldots, \eta_n$ such that conditions (a)-(d) are satisfied.

Contrary to the η_k's, the λ_j's are not restricted to be nonnegative. Note that equations (b) simply regenerates the constraints in (1). Taking into account the definitions of D_{LS} and $g_0, \ldots, g_{m+n}$, the above conditions transform into:

(a) $2 \cdot (p_i - \frac{h_i}{H}) - \lambda_0 - \sum_{j=1}^{m} \lambda_j \cdot x_i^j - \eta_i = 0 \quad \forall i = 1, \ldots, n$
(b_1) $\sum_{i=1}^{n} p_i - 1 = 0; \quad \sum_{i=1}^{n} p_i \cdot x_i^j - \mu_j = 0 \quad \forall j = 1, \ldots, m$
(b_2) $p_i \geq 0 \quad \forall i = 1, \ldots, n$
(c) $\eta_i \cdot p_i = 0 \quad \forall i = 1, \ldots, n$
(d) $\eta_i \geq 0 \quad \forall i = 1, \ldots, n$

For each of the n pairs (p_i, η_i), the two variables p_i and η_i are referred to as complementary variables, because only one of the two can be nonzero. Since

all $2n$ variables are nonnegative, constraints of type (c) can be combined into a single constraint:

$$(c_1) \sum_{i=1}^{n} \eta_i \cdot p_i = 0$$

Therefore the original problem reduces to the equivalent problem of finding a feasible solution $(p_1, \ldots, p_n, \lambda_0, \ldots, \lambda_m, \eta_1, \ldots, \eta_n)$ to the conditions (a), (b_1), (b_2), (c_1),(d). Note that these are linear programming constraints except for the complementary constraint (c_1). Let us consider the following linear programming model obtained by setting $\lambda_j = \lambda_{1,j} - \lambda_{2,j}$ with $\lambda_{1,i}, \lambda_{2,i} \geq 0$ and by introducing $n + m + 1$ artificial variables z_i:

Minimize $z_1 + \ldots + z_{n+m+1}$
subject to

$$2 \cdot (p_i - \tfrac{h_i}{H}) - (\lambda_{1,0} - \lambda_{2,0}) - \sum_{j=1}^{m} (\lambda_{1,j} - \lambda_{2,j}) \cdot x_i^j - \eta_i + z_i = 0 \quad \forall i = 1, \ldots, n$$

$$\sum_{i=1}^{n} p_i - 1 + z_{n+1} = 0$$

$$\sum_{i=1}^{n} p_i \cdot x_i^j - \mu_j + z_{n+j+1} = 0 \quad \forall j = 1, \ldots, m$$

$$p_i \geq 0 \quad \lambda_{1,0}, \lambda_{2,0}, \lambda_{1,j}, \lambda_{2,j} \geq 0 \quad \eta_i \geq 0 \quad z_k \geq 0$$

$$\forall i = 1, \ldots, n \quad \forall j = 1, \ldots, m \quad \forall k = 1, \ldots, n + m + 1$$

Thanks to the artificial variables z_i, an initial basic solution for the simplex method can be obtained as follows ($p_i = \eta_i = \lambda_{1,0} = \lambda_{2,0} = \lambda_{1,j} = \lambda_{2,j} = 0$):

$z_i = \frac{2h_i}{H} \qquad i = 1, \ldots, n$
$z_{n+1} = 1$
$z_{n+j+1} = \mu_j \qquad j = 1, \ldots, m$ (if $\mu_j < 0$ then z_{n+j+1} is replaced by $-z_{n+j+1}$ in the above constraints and thus $z_{n+j+1} = -\mu_j$)

The complementary constraint (c_1) is not included in the model and thus the ordinary simplex method provides an optimal solution which does not necessarily meet all of the Karush-Kuhn-Tucker conditions. Constraint (c_1) implies that it is not permissible for both complementary variables (p_i, η_i) to be (non degenerate) basic variables. To make sure that constraint (c_1) is satisfied at each step of the algorithm, we consider a modified simplex method using a restricted entry rule. For each of the n pairs (p_i, η_i), whenever one of the two variables is a basic variable, the other variable cannot be an entering basic variable. With this restricted entry rule the final solution meets all conditions (a)-(d) provided that the sum $z_1 + \ldots + z_{n+m+1}$ is nil. The resulting p_i's make up the distribution $\widehat{P}_{LS} = (p_1, \ldots, p_n)$ which minimizes the least square function. If the optimal z_k's are not all equal to zero

$(k = 1, \ldots, n+m+1)$, then the set $S \cap \{p_i \geq 0, i = 1, \ldots, n\}$ is empty and there is no feasible solution to problem (1).

At this stage, it is worth distinguishing a particular case which might permit us to avoid solving the above linear programming problem. If the moments of the empirical distribution $\widehat{P}_h = (\frac{h_1}{H}, \ldots, \frac{h_n}{H})$ do not differ much from the known theoretical moments $(g_j(\frac{h_1}{H}, \ldots, \frac{h_n}{H}) \approx 0 \ \forall j = 1, \ldots, m)$, then the desired p_i's will be close to the values $\frac{h_i}{H}$, with the latter all being nonnegative. Thus it is likely to end up with $p_i \geq 0$ without imposing it explicitly through conditions (b_2). By setting $\eta_i = 0$, complementary constraint (c) is satisfied automatically. With condition (a), the p_i's can be written as follows:

$$p_i = \frac{h_i}{H} + \frac{1}{2} \cdot (\lambda_0 + \lambda_1 \cdot x_i + \lambda_2 \cdot x_i^2 + \ldots + \lambda_m \cdot x_i^m). \tag{5}$$

Using (5) in $g_0 = 0, \ldots, g_m = 0$, we obtain the following linear system of $m+1$ equations in $m+1$ unknowns $\lambda_0, \ldots, \lambda_m$:

$$\begin{pmatrix} n & \sum_{i=1}^n x_i & \cdots & \sum_{i=1}^n x_i^m \\ \sum_{i=1}^n x_i & \sum_{i=1}^n x_i^2 & \cdots & \sum_{i=1}^n x_i^{m+1} \\ \vdots & \vdots & & \vdots \\ \sum_{i=1}^n x_i^m & \sum_{i=1}^n x_i^{m+1} & \cdots & \sum_{i=1}^n x_i^{2m} \end{pmatrix} \cdot \begin{pmatrix} \lambda_0 \\ \lambda_1 \\ \lambda_2 \\ \vdots \\ \lambda_m \end{pmatrix} = \begin{pmatrix} 0 \\ 2 \cdot (\mu_1 - \mu_1^S) \\ 2 \cdot (\mu_2 - \mu_2^S) \\ \vdots \\ 2 \cdot (\mu_m - \mu_m^S) \end{pmatrix} \tag{6}$$

where $\mu_r^S = \sum \frac{h_i}{H} \cdot x_i^r$ is the sample moment of order r $(r = 1, \ldots, m)$. The above system has a solution $(\lambda_0, \ldots, \lambda_m)$ if and only if its matrix B, of dimension $(m+1) \times (m+1)$, is regular. This property is verified in accordance with the assumptions made in section 2 (see lemma 2). Thus the solution of (6), combined with relation (5), enables us to compute the p_i's easily. If the resulting p_i's are nonnegative, we are done. Otherwise there is need to go through the modified simplex method presented above. The above check may save us time and is worth performing when the empirical moments are close to the theoretical ones.

Lemma 2: Matrix B in (6) is regular (i.e. B has rank $m+1$) if and only if
(i) $m + 1 \leq n$
(ii) the x_i's are all different

Proof : Matrix B is symmetric and can be rewritten as $B = C^T C$ where C is a $n \times (m+1)$ matrix equal to the transpose of the matrix A introduced in the proof of lemma 1:

$$C = \begin{pmatrix} 1 & x_1 & \cdots & x_1^m \\ 1 & x_2 & & x_2^m \\ \vdots & \vdots & & \vdots \\ 1 & x_n & \cdots & x_n^m \end{pmatrix}$$

Let v be a nonzero vector of size $m+1$. The relation $Bv = 0$ implies $0 = v^tBv = v^tC^tCv = \|Cv\|^2$ and thus $Cv = 0$. This means that B is regular if and only if the kernel of the linear mapping $C : \Re^{m+1} \rightarrow \Re^n$ is reduced to $\{0\}$.

The fundamental relationship

$$dim(Kernel(C)) = dim(\Re^{m+1}) - dim(Image(C)) = m + 1 - rank(C)$$

shows that the kernel of C is nil if and only if C is of rank $m+1$. If $m+1 > n$, the latter is clearly not possible and thus B is singular independently of the values taken by the x_i's. If $m+1 \leq n$, we can show, using the Vandermonde determinant as in the proof of lemma 1, that C is of rank $m+1$ if the x_i's are all different. Hence B is regular if and only if conditions (i) and (ii) are met. ■

3.2 The maximum likelihood function

When D_{LS} is replaced by D_{ML} there does not exist an easy way to solve the Karush-Kuhn-Tucker conditions. To tackle problem (1) a gradient search algorithm known as the Franck and Wolfe algorithm (Minoux, 1983; Hillier and Lieberman, 1990) is proposed in this section. The idea of the algorithm is particularly simple. It combines linear approximations of the function D_{ML} (enabling us to use the simplex method) with a standard one dimensional search procedure. Details of our adaptation of the Franck and Wolfe algorithm are given in Table 1.

At each stage k of the algorithm, $P^{(k)}$denotes the current probability distribution, C is just a constant and $\nabla D_{ML}^T(P^{(k)})$ is the transpose of the gradient vector of D_{ML} evaluated at $P^{(k)}$. The linear function $f(Q)$ is the first-order Taylor's series expansion of D_{ML} around $P^{(k)}$. Because $D_{ML}(P^{(k)})$ and $\nabla D_{ML}^T(P^{(k)}) \cdot P^{(k)}$ have fixed values, they can be dropped in the definition of f. The simplex method is applied to produce the optimal solution Q^* of LP($P^{(k)}$). Note that the value of f increases along the line segment from $P^{(k)}$ and Q^*, but that is not necessarily the case for the original objective function D_{ML}. Therefore, rather than just accepting Q^* as the next trial distribution, we choose the point that maximizes D_{ML} along this line segment. The resulting function $g(\alpha)$, $0 \leq \alpha \leq 1$, is concave and thus the optimal value α^* can be found easily by implementing any one dimensional search technique (see Rardin 1998, for example).

Table 1: The Franck and Wolfe algorithm

(1) $k := 0$;
Let $P^{(0)} \in S$ be a feasible distribution such that $p_i > 0 \quad (i = 0, \ldots, n)$;

(2) $k := k + 1$;

(3) Use the simplex algorithm to solve the LP($P^{(k)}$) problem (C is a constant):
Max $f(Q) = D_{ML}(P^{(k)}) + \nabla D_{ML}^T(P^{(k)}) \cdot (Q - P^{(k)}) = C + \sum_{i=1}^{n} \frac{h_i}{p_i^{(k)}} \cdot q_i$
s.t. $Q = (q_1, \ldots, q_n) \in S$, $q_i \geq 0 \quad (i = 0, \ldots, n)$.
Let Q^* be the optimal solution achieved.

(4) Solve the following one-dimensional problem:
Max $g(\alpha) = D_{ML}\left(P^{(k)} + \alpha \cdot (Q^* - P^{(k)})\right) = \sum_{i=1}^{n} h_i \cdot \log(p_i^{(k)} + \alpha \cdot (q_i^* - p_i^{(k)}))$
s.t. $\alpha \in [0, 1]$
Let α^* be the optimal value achieved.

(5) Set $P^{(k+1)} = P^{(k)} + \alpha^* \cdot (Q^* - P^{(k)})$ and return to step 2.

It is important to point out that in step 1 we force the components of $P^{(0)}$ to be strictly positive. If such a solution exists, it can be found using the Fourier-Motzkin reduction method. The latter is a step by step procedure which transforms a given polyhedron U_m in $\Re^m$ into a polyhedron U_{m-1} in $\Re^{m-1}$ such that $U_m \neq \emptyset \Leftrightarrow U_{m-1} \neq \emptyset$. Details of the Fourier-Motzkin reduction method can be found in Chvatal (1983) or Duffin (1974).

Since $0 < D_{ML}(P^{(0)}) \leq D_{ML}(P^{(1)}) \leq \ldots \leq D_{ML}(\widehat{P}_{ML})$ (the values of the trial solutions generated are non decreasing), all intermediate distributions $P^{(k)}$ will have strictly positive components. This is essential when computing the function $f(Q)$ because of the ratios $\frac{h_i}{p_i^{(k)}}$. The Fourier-Motzkin method is also helpful in determining whether the set $S \cap \{0 < p_i < 1, i = 1, \ldots, n\}$ is empty. If at least one component of $p \in S$ is required to be negative then there is no feasible solution to problem (1). If at least one component has to be nil, then $D_{ML}(\widehat{P}_{ML}) = -\infty$ and the problem reduces to finding any solution in $S \cap \{0 \leq p_i \leq 1, i = 1, \ldots, n\}$.

The Franck and Wolfe algorithm is shown to converge to a local optimum in Minoux (1983). Since a local optimum is necessarily a global optimum in convex programming, it converges to the distribution of interest $\widehat{P}_{ML}$. In practice, the algorithm is stopped as soon as two consecutive trial distributions $P^{(k)}$ and $P^{(k+1)}$ are close enough together (in terms of the Euclidian distance for example). Note that this algorithm can be used to solve the problem studied in section 3.1 but it is not as efficient since the optimal solution is found through a convergence process only.

Before concluding this section, let us point out that problem (1), with the ordinary maximum likelihood function instead of D_{ML}, can be seen as

a geometric programming problem. Geometric programming is useful for modelling engineering design problems and efficient solution methods were proposed in the literature (Beightler et al, 1979). As a rule, these methods transform a non linear programming problem with non linear constraints into a convex programming problem with linear constraints. In our context the geometric programming approach is not helpful since problem (1) is already convex and linearly constrained.

4 An example

A simple example is proposed in this section to illustrate the algorithms presented above. Let X be a discrete variable taking on the values $\{0, 1, 2, 3, 4\}$ ($n = 5$). The mean and the moment of order 2 are known and are equal to 2 and 5 respectively ($m = 2$). These values were chosen equal to the first two moments of the binomial distribution with parameters $(4, 0.5)$. A sample of size $H = 20$ is given and the frequencies h_i observed are reported in Table 2.

Table 2: Sample observed

X	0	1	2	3	4
h_i	2	4	8	5	1

The two problems of interest are as follows:

Min $D_{LS} = (p_0 - \frac{2}{20})^2 + (p_1 - \frac{4}{20})^2 + (p_2 - \frac{8}{20})^2 + (p_3 - \frac{5}{20})^2 + (p_4 - \frac{1}{20})^2$

and,

Max $D_{ML} = 2 \cdot \log p_0 + 4 \cdot \log p_1 + 8 \cdot \log p_2 + 5 \cdot \log p_3 + \log p_4$

subject to:

$$p_0 + p_1 + p_2 + p_3 + p_4 = 1$$
$$p_1 + 2 \cdot p_2 + 3 \cdot p_3 + 4 \cdot p_4 = 2$$
$$p_1 + 4 \cdot p_2 + 9 \cdot p_3 + 16 \cdot p_4 = 5$$
$$p_0, p_1, p_2, p_3, p_4 \geq 0$$

The resulting distributions $\widehat{P}_{LS}$ and $\widehat{P}_{ML}$ are shown in Table 3 together with the sample distribution $\widehat{P}_h$ and the above-mentioned binomial distribution P_{bin}.

Table 3: The four distributions under consideration ($\mu_1 = 2$, $\mu_2 = 5$)

X	0	1	2	3	4
$\widehat{P}_h$	0.1	0.2	0.4	0.25	0.05
P_{bin}	0.0625	0.25	0.375	0.25	0.0625
$\widehat{P}_{LS}$	0.08285	0.1986	0.4073	0.2588	0.05325
$\widehat{P}_{ML}$	0.08594	0.1914	0.4102	0.2617	0.05080

In this case $\widehat{P}_{LS}$ could be computed easily, without using the modified simplex algorithm, by simply solving the linear system (6). Also, the Franck and Wolfe algorithm converges quickly, in less than 10 iterations, to the distribution $\widehat{P}_{ML}$. A graphical representation of $\widehat{P}_{LS}$, $\widehat{P}_{ML}$ and $\widehat{P}_h$ is provided in Figure 1. In this example the two distributions $\widehat{P}_{LS}$ and $\widehat{P}_{ML}$ turn out to be very similar. This results from the fact that the sample mean and the sample moment of order 2, which are equal to 1.95 and 4.85 respectively, are close to the desired moments. The mean is increased from 1.95 to 2 by giving less weight to the values $x_i \in \{0, 1\}$ in $\widehat{P}_{LS}$ and $\widehat{P}_{ML}$. Other examples have shown that the distributions achieved with the two objective functions selected in this study do not differ much in general.

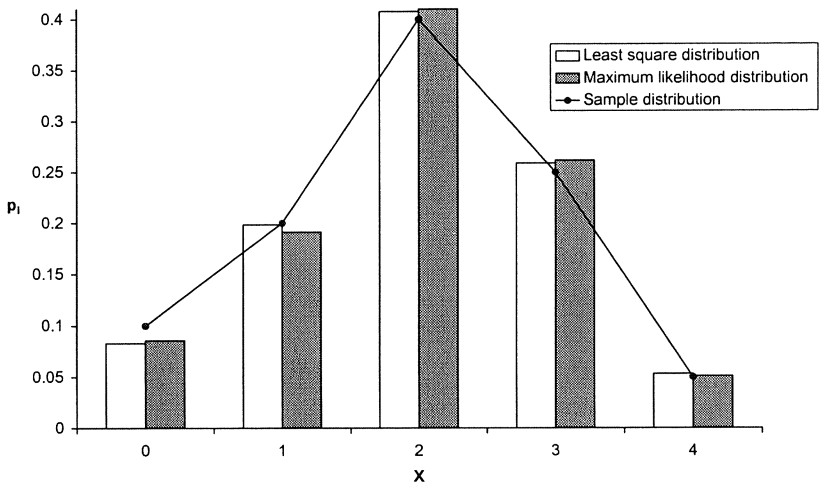


Figure 1: The three distributions $\widehat{P}_{LS}$, $\widehat{P}_{ML}$ and $\widehat{P}_h$

Let us now consider the situation where the mean of the distribution is different from 2. If $\mu_1 = 1.3125$, the solution we obtain by solving equations (5) and (6) includes one negative component and thus we must use the modified simplex algorithm to compute $\widehat{P}_{LS}$. As shown in Table 4, the main difference between $\widehat{P}_{LS}$ and $\widehat{P}_{ML}$ lies in the fact that $\widehat{P}_{ML}$ does not contain a nil component. The algorithm of Franck and Wolve will always converge to such a solution provided that there exist feasible distribution vectors with nonnil components. If $\mu_1 = 1.25$ then there is only one distribution meeting the given constraints and thus $\widehat{P}_{LS} = \widehat{P}_{ML}$. This unique distribution is also reported in Table 4. Since $p_1 = p_2 = p_3 = 0$, it follows that $D_{ML}(\widehat{P}_{ML}) = -\infty$. If $\mu_1 = 1.125$ then there is no distribution meeting the moments of order 1 and 2 ($S \cap \{0 \leq p_i \leq 1\} = \emptyset$).

Table 4: Changes in the mean of the distribution
($\mu_1 \in \{1.125, 1.25, 1.3125\}$)

	X	0	1	2	3	4
$\mu_1 = 1.3125$	$\widehat{P}_{LS}$	0.6147	0.06909	0.0107	0	0.3055
	$\widehat{P}_{ML}$	0.6249	0.0501	0.006574	0.02449	0.2939
$\mu_1 = 1.25$	$\widehat{P}_{LS} = \widehat{P}_{ML}$	0.6875	0	0	0	0.3125
$\mu_1 = 1.125$	no distribution					

To validate the approach presented in this paper, one can check that the distributions $\widehat{P}_h$ and $\widehat{P}_{LS}$ (or $\widehat{P}_{ML}$) do not differ significantly. Let P_h, P_{LS} and P_{ML} be the unknown distributions that $\widehat{P}_h$ and $\widehat{P}_{LS}$ and $\widehat{P}_{ML}$ are estimating. In our context, a goodness of fit test with the following null hypothesis:

$$H_0 : \quad P = P_h = P_{LS} \quad (\text{or } P = P_h = P_{ML})$$

is appropriate. To specify whether the null hypothesis H_0 should be accepted or rejected, one possibility is to implement the chi-squared test, with n classes x_i and with the observed and "expected" frequencies being equal to h_i and $H \cdot \widehat{P}_{LS,i}$ (or $H \cdot \widehat{P}_{ML,i}$) respectively. Here, the chi-squared statistic is distributed approximately as a χ^2 variable with $\nu = n - 1 - m$ degrees of freedom since the "expected" frequencies are computed knowing the first m moments of the distribution.

In the above example with $\mu_1 = 2$, we obtain the values $\chi^2_{LS} = 0.0838$ and $\chi^2_{ML} = 0.0694$ for the chi-squared statistic. The 0.95 quantile of the χ^2 distribution with $\nu = 5-1-2 = 2$ degrees of freedom is equal to 5.99. Thus, at the level of significance $\alpha = 0.05$, there is not enough evidence to reject the hypothesis that the three distributions $\widehat{P}_h$, $\widehat{P}_{LS}$ and $\widehat{P}_{ML}$ are the same. Note that the same conclusion does not hold when $\mu_1 = 1.25$ or $\mu_1 = 1.3125$.

5 Conclusion

In this research we have proposed a model for estimating a probability distribution knowing the first m moments of the distributions and given a sample of observations. This results in a constrained convex programming problem. To measure the similarity between the empirical and the desired distributions, two objective functions are considered, namely the least squares and the maximum likelihood functions. A modified simplex algorithm is introduced to minimize the least square function. It exploits the fact that, in our case, the Karush-Kuhn-Tucker conditions for constrained optimization are "almost" linear when the objective function is quadratic. To achieve the maximum likelihood with regard to the observed sample, we use an adapta-

tion of the Franck and Wolfe algorithm. The two proposed algorithms are very simple to implement and the distribution of interest is obtained quickly.

This study can be extented in a least two directions. First it would be interesting to better characterize the effect of the choice of the objective function on the final distribution. A second desirable extension would be to the case where the underlying random variables are continuous. However, the resulting problem formulation does not appear to be as simple as in the discrete case and further assumptions are necessary to define the shape of the distribution to be estimated.

Acknowledgements: The research leading to this paper was supported by the Natural Sciences and Engineering Research Council of Canada under grant No. A1485 and by the Carma Chair at the University of Calgary. The authors would like to thank an anonymous referee for his excellent comments and remarks.

References

[1] Beightler C.S., Phillips D.T., Wilde D.J. (1979), Foundations of Optimization, 2nd edition, Prentice-Hall, Englewood Cliff, N.J.

[2] Chvatal, V. (1983), Linear Programming, Freeman, New York, 241-242.

[3] Duffin, R.J. (1974), "On Fourier's analysis of linear inequalities systems", Mathematical Programming Study 1, 71-95.

[4] Ferson, S. (1997), "Probability Bounds Analysis Software", VIP-68, Computing in Environmental Resource Management, Gertler A. (Ed.), Air & Waste Management Association, Pittsburgh, Pennsylvania, 669-678.

[5] Hillier, F.S. and Lieberman, G.J. (1990), Introduction to Operations Research, 5th edition, McGraw-Hill, New York.

[6] Hoffman, K. and Kunze, R. (1971),Linear Algebra, second edition, Prentice-Hall, Englewood Cliffs - N.J.

[7] Minoux, M. (1983), Programmation Mathémathique : Théorie et Algorithmes, Tome 1, Dunod, Paris.

[8] Rardin, R.L. (1998), Optimization in Operations Research, Prentice-Hall International, London.

[9] Silver, E.A., Costa, D. and Zangwill, W. (1998), "Order Statistics of Independent Identically Distributed Variables when the Sum is Known",

to appear in Statistics and Computing.

[10] Silverman, B. W. (1986), Density Estimation for Statistics and Data Analysis, Chapman & Hall, London.

[11] Strijbosch, L.W.G. and Heuts, R.M.J. (1994), "Investigating Several Alternatives for Estimating the Lead Time Demand Distribution in a Continuous Review Inventory Model", Kwantitatieve Methoden, 15 (46), 57-75.

[12] Wismer, D.A. and Chattergy, R. (1978), Introduction to Nonlinar Optimization: A Problem Solving Approach, North-Holland, New-York.

Student (1999), Vol.3, No.1, 41-43

Proof Without Words: Sum of Squared Integers

Nanny Wermuth and Hans-Jürgen Schuh

Johannes-Gutenberg Universität, Mainz, Germany

When a treatment effect is to be juged from a random sample on paired observations, like measurements taken for each patient before and after treatment, the judgement may be based on rankings of the size of differences (Wilcoxon, F. (1945). Individual comparisons by ranking methods. *Biometrics*, **1**, 80-83.).

In samples of n nonzero differences the signed rank statistic, W, is a sum of variables which have values $+i$, $-i$, for $i = 1, \ldots, n$, where i denote ranks assigned to absolute values of differences. Here is a small example how to compute an observed w of W and related quantities.

l	x_l	y_l	$x_l - y_l = d_l$	$\lvert d_l \rvert$	Rank $\lvert d_l \rvert$ i	Signed rank: r_i
1	10	5	5	5	4	4
2	12	10	2	2	1	1
3	5	9	-4	4	3	-3
4	9	6	3	3	2	2

$w = \sum r_i = 4$, $w_{pos} = 7$, $w_{neg} = 3$, $w_{smaller} = 3$.

In the case of no treatment effect positive and negative signs are equally probable for each rank, therefore the mean and variance of W is obtained as $E(W) = 0$ and $var(W) = \sum i^2$ (Mosteller, F. & Rourke, R.E.K. (1973). *Sturdy Statistics.* Reading: Addison-Wesley.).

For large n the test statistics is based, equivalently, either on the negative ranks W_{neg} or the positive ranks W_{pos}, i.e. for ease of computation the one based on the smaller number of signs can be chosen. An approximately

standard normal test statistic results as:

$$\frac{W_{pos} - E(W_{pos})}{var(W_{pos})^{1/2}} = \frac{W_{pos} - n(n+1)/4}{\sqrt{(2n+1)(n+1)n/24}},$$

where

$$\begin{aligned} W_{pos} &= W + W_{neg} \\ 2W_{pos} &= W + W_{neg} + W_{pos} = W + \sum i \\ E(W_{pos}) &= \frac{1}{2}\left[E(W) + \sum i\right] = \frac{1}{2}\sum i \\ var(W_{pos}) &= \frac{1}{4}var(W) = \frac{1}{4}\sum i^2. \end{aligned}$$

To see that

$$\sum i = \frac{n(n+1)}{2}$$

an anecdote about C.F. Gauss can be used. To keep Carl Friedrich, at age of 7, quiet for some time his teacher asked him to add the numbers from 1 to 500. However this he did fast as follows:

1	2	3	$\cdots$	498	499	500
500	499	498	$\cdots$	3	2	1
501	501	501	$\cdots$	501	501	501

giving

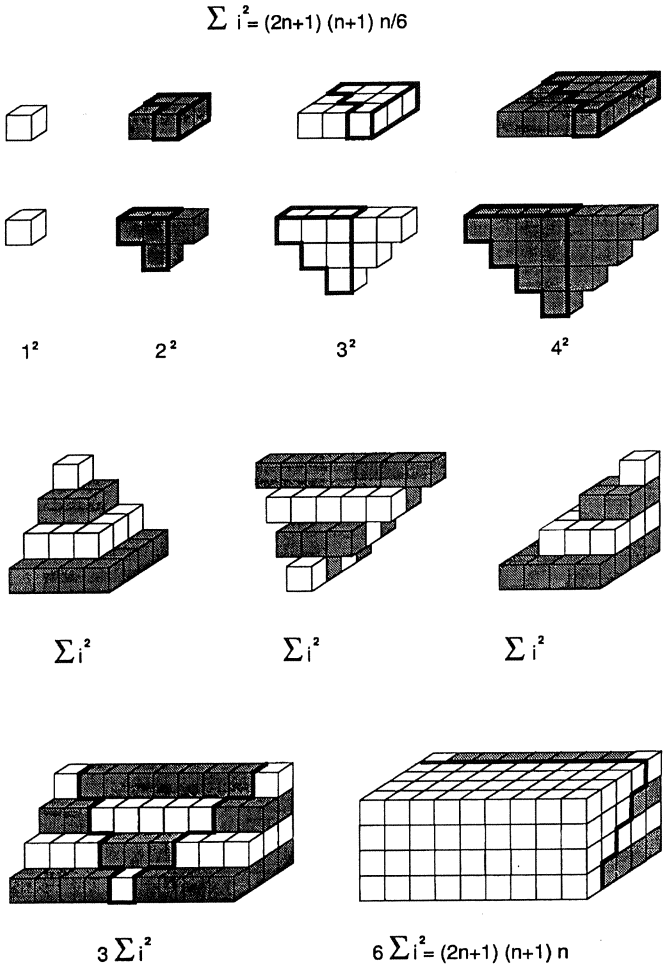
$$2\sum_{i=1}^{i=500} i = 500 \cdot 501,$$

so that, in general

$$2\sum i = n(n+1).$$

The value of $\sum i^2$ is derived in the following 'Proof without Words'.

Proof Without Words: Sum of Squared Integers



— NANNY WERMUTH, HANS-JÜRGEN SCHUH
JOHANNES-GUTENBERG UNIVERSITÄT
D-55099 MAINZ, GERMANY

Student (1999), Vol.3, No.1, 46-54

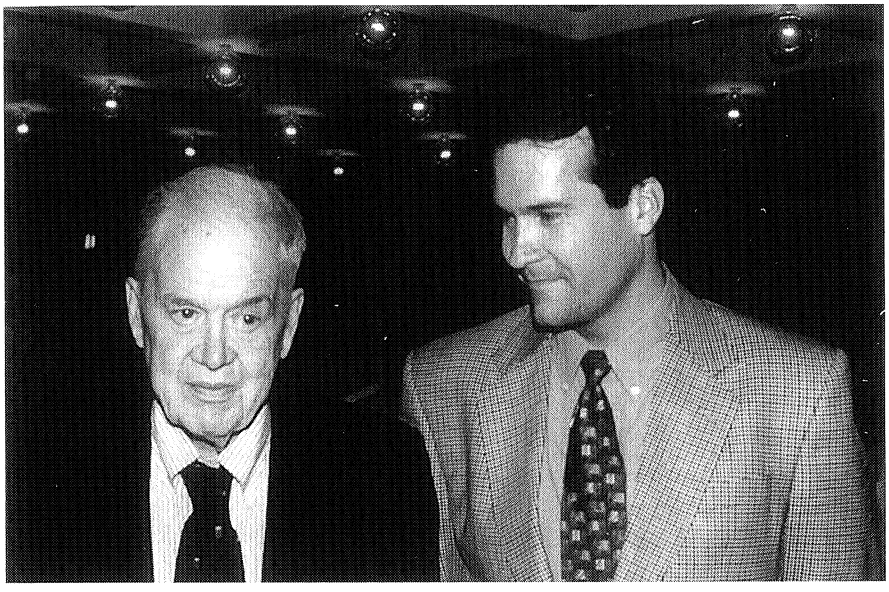
John W. Tukey, Founder of Exploratory Data Analysis

John Wilder Tukey was born in New Bedford, Massachusetts, on June 16, 1915. He earned bachelor's and master's degrees in chemistry at Brown University in 1936 and 1937, respectively. He then went to Princeton, where, in 1938 he was a Jacobus Fellow, and upon receiving his doctorate in mathematics in 1939, he was appointed Henry B. Fine Instructor in Mathematics. A decade later, at age 35, he was advanced to a full professorship. He directed Princeton's Statistical Techniques Research Group from its founding in 1956. When the Department of Statistics was formed in 1965, Tukey was named its first chairman.

He has led the way in the now burgeoning fields of Exploratory Data Analysis and Robust Estimation and his contributions to the Spectrum Analysis of Time Series and other aspects of Digital Signal Processing have been widely used in engineering and science. He has been credited with coining the word "bit" a contraction of "binary digit" which refers to a unit of information, often as processed by a computer.

In another area, collaboration with a fellow mathematician resulted in the formation of the Cooley-Tukey Fast Fourier Transform (FFT) algorithm, a mathematical technique that greatly simplifies computation for Fourier series and integrals. Other interests range through applications to such fields as behavioral sciences, geophysics and pharmaceutical research.

In 1965 John Tukey was named the first recipient of the Samuel S. Wilks Award of the American Statistical Association. He received the National Medal of Science in 1973 "for his studies in mathematical and theoretical statistics and for his outstanding contributions to the applications of statistics to the physical, social, and engineering sciences". Tukey received the Shewhart Medal of the American Society for Quality Control in 1977; the Institute of Electronic and Electrical Engineers' 1982 Medal of Honor for the Cooley-Tukey Fast Fourier Transform (FFT) Algorithm; and the American Society for Quality Control's Deming Medal in 1983. In 1989 he was elected a Foreign Member of The Royal Society (London). Princeton University honored him in 1984 with the James Madison Medal, given annually to an alumnus of the Graduate School, who has had a distinguished career, who has advanced the cause of graduate education, or who has achieved a record of outstanding public service. In 1989 he received the Monie A. Ferst Award of Sigma Xi, and in 1990 the Educational Testing Service Award for Distinguished Service to Measurement.



John W. Tukey with Stephan Morgenthaler the night of the dinner at Prospect House, Princeton University, celebrating John's 80^{th} birthday, June 19, 1995

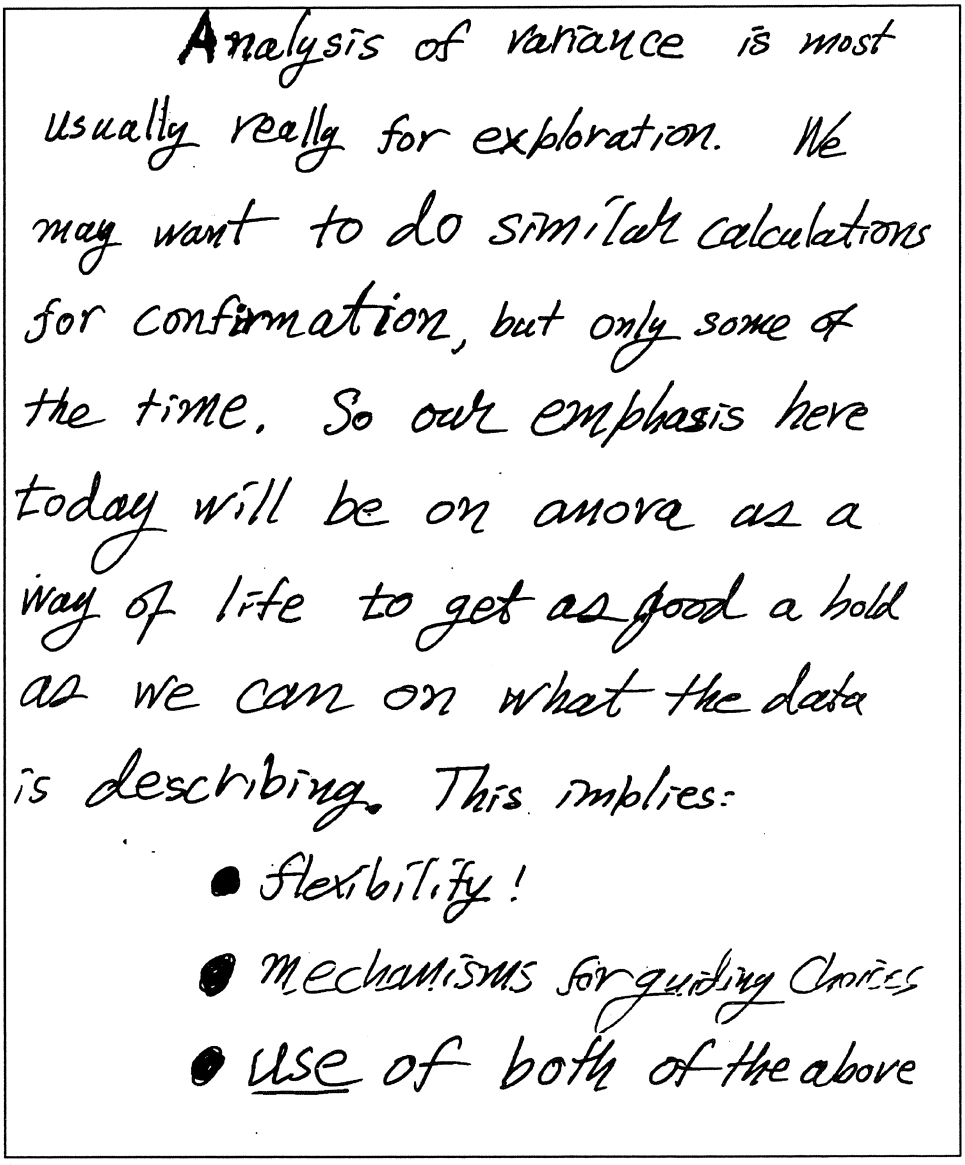
Author of Exploratory Data Analysis (now translated into Russian) and eight (to date) volumes of collected papers, he is co-author of: Statistical Problems of the Kinsey Report on Sexual Behavior in the Human Male; Data Analysis and Regression (also translated into Russian); Index to Statistics and Probability: Citation Index; Permuted Titles; Locations and Authors (four volumes); The Measurement of Power Spectra; and Robust Estimates of Location: Survey and Advances. He was co-editor of and contributor to Understanding Robust and Exploratory Data Analysis, Exploring Data Tables, Trends Shapes, Configural Polysampling, and Fundamentals of Exploratory Analysis of Variance.

John Tukey has written more than 500 technical papers and reports and he has served on editorial committees of several professional organizations. He has been represented frequently in publications such as the Annals of Mathematical Statistics (of which he was associate editor, 1950-52), Biometrics, the Journal of the American Statistical Association (JASA), Science, and Technometrics.

John Tukey has been a consultant (among others) to various pharmaceutical companies (over 40 years to Merk and Company) and, since retirement, to the Xerox Corporation (among others). He led the statistical component of NBC's election night vote projection effort in all major elections from 1960 to 1980, after which exit polls took over the role previously played by statisticians.

For more information about John W. Tukey's life, the reader is refered to an excellent book edited by D.R. Brillinger, L.T. Fernholz ans S. Morgenthaler. We wish to express our scincere thanks to Professors Louisa Fernholz and Stephan Morgenthaler for providing the journal Student with professor Tukey's handwritten transparencies and the photos.

J.W. Tukey giving a lecture at Bell Labs or Princenton (early 80's)



John W Tukey's handwriting, extract from Statistics 411 a course developped at Princeton University.

BASIC STRUCTURE.

Factorial anova, where all permitted combinations of the chosen versions of the chosen factors are represented by observed values, is the simplest--perhaps even the most important--case. But it is not alone.

It is, however, the place to start, and the place to learn as much as we can. So this is where we will begin.

Factors can be "crossed" or "nested" or have more complicated relations to one another. We shall

stick to the simplest cases, and, at least for most of what we say, will assume that making the most standard computations can be passed over — that either you know how to do them, or you have a program that does.

Analysis of variance, classical or modern, lives by <u>value-splittings</u> by separation of observed numbers into more-or-less meaningful parts. Such an identity as:

$$(\text{observation})_{ij} = m + a_i + b_j + c_{ij}$$

should be thought of as breaking-up a two-way table, entry-by-entry--one entry at a time--into 4 two-way tables.

To stress the way these tables combine to give the observations, we shall call them Overlays.

An example of such a value-splitting might be

$$\begin{pmatrix} 11 & 13 & 15 \\ 22 & 23 & 24 \\ 33 & 33 & 33 \end{pmatrix} = \begin{pmatrix} 23 & 23 & 23 \\ 23 & 23 & 23 \\ 23 & 23 & 23 \end{pmatrix} + \begin{pmatrix} -1 & 0 & 1 \\ -1 & 0 & 1 \\ -1 & 0 & 1 \end{pmatrix}$$

$$+ \begin{pmatrix} -10 & -10 & -10 \\ 0 & 0 & 0 \\ 10 & 10 & 10 \end{pmatrix} + \begin{pmatrix} -1 & 0 & 1 \\ 0 & 0 & 0 \\ 1 & 0 & -1 \end{pmatrix}$$

At least two comments are important here:

In the precomputer era, when arithmetic was painful, only summaries of overlays, beginning with the sums of the squares of their entries, namely

$9 \times 529 = 4761$ (M₂ = 4761)
$3 \times 2 = 6$ (M₂ = 3)
$3 \times 200 = 600$ (M₂ = 300)
$1 \times 4 = 4$ (M₂ = 1)

in our example. Today, with computers --and printers, as well as screens, proliferating madly, we have no excuse for not looking at the overlays themselves -- or at things that completely specify the overlays.

Data & Statistics in *Student (1999), Vol.3. No.1, 57-75*

Reliability of clinical indices

Philippe Mojon D.M.D.

Division of Gerodontology and Removable Prosthodontics, University of Geneva, Switzerland

Abstract: A major question in epidemiological studies is concerned with the reliability of clinical indices. A variety of statistical tools have been developed to assess reliability, yet none of them seems widely accepted. To assess the reliability of 9 clinical dental indices, 15 subjects were examined by 2 trained examiners and re-examined again 2 weeks later by the same examiners. The prevalence of the clinical features under scrutiny ranked from 1% to 53%. Russel and Rao index (RRI, 1940), Lerman index (1970), and the overall non-weighted and weighted Kappa (Cohens Kappa, 1960) were computed in order to evaluate the reliability of these clinical indices. Then, the clinical indices were ranked from worst to best according to each reliability index separately and the ranking thus obtained were compared. The 3 statistical techniques provided different ranking. The more complex clinical indices were ranked low by the RRI and Lerman index but were ranked better by Kappa index. The use of weights, either linear or squared, improved even further the Kappa indices of these complex clinical indices. The less reliable clinical indices scored similarly with the three statistical methods. It seems that the choice of one statistical method over the others should be based on the prevalence of the clinical feature under scrutiny and on the complexity of the clinical index.

Key words: reliability, intra-examiner agreement, Russel and Rao index, Lerman index, Kappa index.

1 Introduction

A major question in epidemiological studies is concerned with the reliability of clinical indices. Clinical indices express numerically the health situation of individuals or the degree of pathological involvement. A clinical index should be simple to use, reliable, reproducible, amenable to statistical analysis, and equally sensitive throughout the scale (see Davies, 1968 [8]). The simplest

indices are dichotomic, indicating whether a character is present or absent. More refined indices use nominal, ordinal or interval scale. Yet, any recording of an information is associated with an error. The error may come from the measurement device or it can be related to the individual who is recording the information. Thus, the evaluation of the reliability of a clinical index is necessary prior to use it as a research tool. In general, the reliability of an index depends on the clarity of the criteria (see Banting, 1993 [4] ; Howat et al., 1981 [14] ; Pitts and Fyffe, 1988 [22]), the number of categories (see Ambjörnsen et al., 1982 [1] ; Poulsen et al.; 1979 [23] ; Fleiss and Chilton 1983 [12]), and the training of the examiners (see Fleiss et al., 1991 [13] ; Kingman, 1986 [16]).

To assess the reliability of 9 clinical dental indices, 15 subjects were examined by 2 trained examiners (a and b) and re-examined again two weeks later by the same examiners (see Mojon et al., 1995 [20] ; Mojon et al., 1996 [21]). For the purpose of this paper, data accumulated by one examiner (examiner b) are sufficient. In order to evaluate the reliability of these clinical indices we used Russel and Rao index (cf. Chandon and Pinson, 1981 see [5]), Lerman index (1970 see [19]), and the overall non-weighted Kappa (Cohens Kappa, 1960 see [7]) to express the amount of similarity or differences between the 2 examinations. In addition weighted Kappa was computed for ordinal clinical indices with > 3 categories (see Kingman 1986 [16] ; Cicchetti and Fleiss, 1977 [6]). Ranking the clinical indices with respect to these indices could point out different classification which, in turn, could be confusing in terms of reliability.

In this study, the statistical unit could be the patient or the tooth. In order to use the patient as the statistical unit, each clinical index must be expressed as one central index that summarises all the teeth, such as the median. However, the use of one central index would result in a loss of information. Others have used the tooth as a statistical unit in similar studies (Kingman, 1986 [16] ; Hunt, 1986 [15] ; Bader & Shugars, 1992 [2] ; Bader & Shugars, 1993 [3]). It simplifies the computing, and increase the power of the statistical tests. However, it is reasonable to assume that the pathologies that affect the teeth are the consequence of the environment and that the neighbouring teeth are equally at risk to get the disease. Thus, the assumption of complete independence between each unit is apparently violated. Yet, in the case of reliability study, the independence means that the examiner is not influenced by the scoring of the neighbouring teeth. This can be quite an acceptable assumption.

2 Clinical indices

The tested clinical indices were as follow:

(1) dental plaque (PD)
accumulation of dental plaque is scored from 0 to 3. It is an ordinal scale.

(2) coronal caries (CC)
detection of caries on the upper part of a tooth.
score: absent, present inactive, present active
it is a nominal scale but can be also interpreted like an ordinal scale.

(3) root caries (CR)
same as CC but on the root portion of the tooth

(4) detection of restorations (RE)
aesthetic fillings are sometimes difficult to detect.
dichotomic score: absent or present

(5) detection of recurrent caries method A (OB1)
caries at the interface tooth-filling.
dichotomic score: absent or present

(6) detection of recurrent caries method B (OB2)
same as OB1 using a more refined clinical criteria.

(7) community periodontal index of treatment need (CP)
evaluation of gingivitis, calculus, and pocket depth.
the score can be considered as an ordinal scale comprising 5 categories.

(8) tooth mobility A (MO1)
an ordinal scale with 3 categories.

(9) tooth mobility B (MO2)
same as MO1 using more refined clinical criteria.

A more detailed description of these clinical indices can be found in previously published reports (Mojon et al., 1995 [20] , Mojon et al., 1996 [21]).

In order to evaluate the prevalence of the diseases, the mean frequency for each category of the clinical indices was computed from the 4 examinations i.e. examiner a and b session 1 and 2 (see Table 1).

Clinical indices	Scores				
	0	1	2	3	4
Dental plaque (PD)	2.73	41.53	51.05	4.43	-
Coronal caries (CC)	96.88	2.35	0.80	-	-
Root caries (CR)	82.43	1.00	16.60	-	-
Restoration (RE)	46.75	53.30	-	-	-
Recurrent caries A (OB1)	68.48	31.53	-	-	-
Recurrent caries B (OB2)	83.63	16.38	-	-	-
Periodontal index	18.63	28.43	11.03	32.03	09.93
Tooth mobility A (MO1)	88.83	10.15	1.30	-	-
Tooth mobility B (MO2)	75.68	23.10	1.25	-	-

Table 1: Mean frequency for each category

3 Data analysis

Intra-examiner reliability was evaluated by comparing data from the first appointment with data recorded at the second appointment. An object was defined as the recording of one of the 9 clinical indices for one session. A total of 283 teeth were examined and the data are presented in a matrix of 20 by 283. Each line represents one tooth. The first column is the identification number of the subject, the second column is the tooth number. Teeth are numbered according to their situation in the mouth, from 11 to 18, 21 to 28, 31 to 38, and 41 to 48. The following columns represent one object each. The suffix (1 or 2) refers to the examination number. The two first columns were not used during the statistical analysis. Some values are missing. It is either because the clinical index could not be applied to the tooth or because the examiners did not have enough time to finish the examination.

The aim of the statistical analysis was to provide for each clinical index three measures of intra-examiner reliability, to rank the clinical indices from worst to best reliability according to each measure separately, and to compare the ranking obtained with each measure.

The 3 indices were 2 proximity indices: Russel and Rao index and Lerman index (1976 see [19]) (both indices described in Chandon and Pinson, 1981 see [5]), and the overall non-weighted Kappa (Cohens Kappa, 1960 see [7]).

Proximity can be defined as any set of numbers that express the amount of similarity or difference between pairs of objects (Dillon and Goldstein, 1984 see [9]). Coefficients are called index of similarity when they represent a measure of whether 2 objects, i and j, are similar or close (Chandon & Pinson, 1981 see [5]). The simplest index is from Russel and Rao while Lerman (Chandon & Pinson, 1981 see [5]) proposed a different approach. His index is based on the number of co-presence expected when the two objects are independent. The probability of rejecting the hypothesis of independence

when the objects are dependent will be close to 1 when the number of observed co-presence is very large. Also, this probability is close to 0 when the number of co-presence is close to 0. Hence, this probability is an increasing function of the co-presence and it can be considered as a similarity index.

Coefficients such as Russel and Rao are sensitive to the distribution of categories in the sample. There is a need, therefore, to incorporate in a rater reliability coefficient, the agreement due to chance. Cohen (1960 see [7]) proposed to compute the excess beyond chance and called the coefficient Kappa. Kappa coefficient is not a similarity index since it does not fulfil the non-negativity property.

When the clinical index is an ordinally scaled variable, the use of weighted Kappa has been advocated (see Kingman 1986 [16] ; Cicchetti and Fleiss, 1977 [6]). All possible disagreements are treated equally when computing overall Kappa. With the weighted Kappa, a weight is assigned for each possible disagreement. This technique provides one with the possibility of awarding partial credits to pairs of disagreement that are "close" to being agreement. To generalize, the weights should be chosen so that worse is the disagreement, less contribution is brought to the Kappa value. In this study linear and squared weights were selected.

The degree of agreement can be characterised according to the value of Kappa. Fleiss (1982 see [11]) proposed that values greater than 0.75 represent excellent agreement while values below 0.40 represent poor agreement. These limits are empirical.

4 Computation

In order to be computed using a worksheet program (Excel 5.0), the Rao and Russel, and the Lerman indices had to be re-written so that they could be computed from a table of proportion. The original formulations of the 3 indices were taken from Chandon & Pinson (1981 see [5]) and Fleiss (1982 see [11]). The data were organised in a $h \times h$ table as such (see Table 2):

		Rater b						
Rater a		1	2	$\cdots$	$\cdots$	$\cdots$	h	total
	1	p_{11}	p_{12}	$\cdots$	$\cdots$	$\cdots$	p_{1h}	$p_{1\cdot}$
	2	p_{21}	p_{22}	$\cdots$	$\cdots$	$\cdots$	p_{2h}	$p_{2\cdot}$
	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
	H	p_{h1}	p_{h2}	$\cdots$	$\cdots$	$\cdots$	p_{hh}	$p_{h\cdot}$
	Total	$p_{\cdot 1}$	$p_{\cdot 2}$	$\cdots$	$\cdots$	$\cdots$	$p_{\cdot h}$	1

Table 2: Data organisation

p_{ij} are proportions.
n = total number of cases
h = number of categories of the clinical indices tested
In that context, the 3 indices can be written as in Table 3:

Russel and Rao	Lerman[1]	Overall Kappa
$RRI = \frac{\sum_{i=1}^{h} p_{ii}}{h}$	$\Pr\left(P' < \frac{k-\overline{P}}{\sigma_p}\right)$ $P' = \frac{h^2 RRI}{h-1} - \frac{1}{h-1}$	$\widehat{K} = \frac{\sum_{i=1}^{h} p_{ii} - \sum_{i=1}^{h} p_{i\cdot} p_{\cdot i}}{1 - \sum_{i=1}^{h} p_{i\cdot} p_{\cdot i}}$

Table 3: RRI, Lerman and Kappa index

Weighted Kappa indices (linear and squared) were computed using a matrix of weight W (see Table 4):

	Clinical indices categories			
	1	2	$\cdots$	k
1	1	W_{12}	$\cdots$	W_{1k}
2	W_{21}	1	$\cdots$	W_{2k}
$\cdots$	$\cdots$	$\cdots$	1	$\cdots$
k	W_{k1}	W_{k2}	$\cdots$	1

Table 4: Matrix of weights

Linear weights:

$$w_{ij} = 1 - \frac{|i-j|}{k-1}$$

squared weights:

$$w_{ij} = 1 - \frac{|i-j|^2}{(k-1)^2}.$$

Weighted Kappa is given by

$$\widehat{K}_w = \frac{p_{0(w)} - p_{e(w)}}{1 - p_{e(w)}}$$

where

$$p_{0(w)} = \sum_{i=1}^{k} \sum_{j=1}^{k} w_{ij} p_{ij}$$

and

$$p_{e(w)} = \sum_{i=1}^{k} \sum_{j=1}^{k} w_{ij} p_{i\cdot} p_{\cdot j}.$$

[1] In order to rank the pairs of objects P' is enough since the normal cumulative distribution is an increasing function

5 Results

5.1 Computed indices

The Russel and Rao, the Lerman index, and the Kappa index were computed for each pairs of objects. In Table 5 we can see that the three indices had different magnitudes. The Russel and Rao index had a range between 0.12 and 0.48, the Lerman index had a range between 0.70 and 0.83, and Kappa varied between 0.21 and 0.94. Apart from CC, all Kappa indices increased when weights were used.

Clinical index	Proximity index				
	RRI	Lerman	Kappa		
			no weight	linear	squared
PD	0.17	0.71	0.42	0.46	0.52
CC	0.32	0.83	0.21	0.17	0.20
CR	0.28	0.78	0.42	0.43	0.42
RE	0.48	0.83	0.94	-[2]	-[2]
OB1	0.43	0.77	0.72	-[2]	-[2]
OB2	0.39	0.72	0.23	-[2]	-[2]
CP	0.12	0.70	0.50	0.58	0.65
MO1	0.31	0.82	0.66	0.70	0.68
MO2	0.29	0.79	0.64	0.67	0.65

Table 5: Calculated RRI, Lerman and Kappa index

5.2 Comparison of ranking

Overall ranking:

Clinical index	RRI rank	Lerman rank	Kappa rank		
			no weight	linear	squared
CP	1	1	5	5	5
PD	2	2	4	4	4
CR	3	5	4	4	4
MO2	4	6	6	6	6
MO1	5	7	7	7	7
CC	6	9	1	1	1
OB2	7	3	2	2	2
OB1	8	4	8	8	8
RE	9	8	9	9	9

Table 6: Ranking

Table 6 shows that the ranking were different. CP and PD, the more complex clinical indices, were ranked low with the RRI and Lerman. RE was

[2] with dichotomic variables non-weighted and weighted are identical

ranked high whatever the index was. The use of weight in the computation of Kappa did not change the ranking.

The following ranking was performed separately for clinical indices having the same number of categories (see Tables 7 and 8).

Clinical index	RRI Rank	Lerman Rank	Kappa Rank
CR	1	1	2
MO2	2	2	3
MO1	3	3	4
CC	4	4	1

Table 7: Clinical indices with 3 categories.

Clinical index	RRI Rank	Lerman Rank	Kappa Rank
OB2	1	1	1
OB1	2	2	2
RE	3	3	3

Table 8: Dichotomic clinical indices

These tables illustrate that Lerman is a transformation of the RRI when the number of categories is kept constant. Thus, the RRI and Lerman provided identical ranking. Kappa produced identical ranking with dichotomic clinical indices but the ranking was different for the 3 categories clinical indices.

6 Discussion

The assessment of examiner reliability is essential prior to any clinical study (Landis and Koch, 1977 [18] ; Kingman et al., 1991 [17]). Indeed, it is unlikely that the individuals who are examining do not make any mistakes or oversight (Fleiss, 1982 [11] ; Fleiss et al., 1979 [10]). The source of the errors may be an inefficient training or the vagueness of the criteria used to identify the clinical feature. Even when all attention has been given to optimise the recording, it is probable that errors will still occur. The quantity and the magnitude of these errors, whatever the source is, should be quantified to provide adjusted estimates based on the data and some indication on reliability (Kingman et al., 1991 [17]).

The design of a preliminary study to evaluate reliability is usually straightforward. The recording of the clinical information is repeated twice (at least) and the results of the two examinations are compared. When the information recorded is normally distributed on a continuous scale, there is not much

controversy about the statistical test to use to quantify the degree of agreement (Landis and Koch, 1977 [18] ; Fleiss, 1982 [11]). However, when the information recorded is categorical, plethora of methods have been proposed (Chandon & Pinson 1981 [5]). It was the purpose of this work to compare three of these methods and to apply them on clinical data. Two indices, the RRI and the Lerman index, evolved from the theory of proximity. They evaluate the similarity between two objects on a scale from 0 to 1. Note that the computation of RRI is very close to the Sokal and Michener index since the later include the co-absences in the numerator while the former does not. In the case of clinical data, the distinction between the two indices becomes unclear. For example, the absence of caries can be interpreted as the presence of the character "free of disease". We opted for the name Russel and Rao in acknowledgement of the anteriority of their index (1940 against 1958). The third index was specifically designed to evaluate examiner agreement (Cohen, 1960 [7]). It is not an index of proximity since it can take negative values but it can be easily transformed to meet this requirement.

An index of reliability should be easy to interpret, amenable to statistical inference, and stable. The RRI is easy to interpret since it can be translated in % of agreement. The Kappa statistics can be interpreted according to the scale from Landis & Koch (1977 see [18]). The strength of the agreement is qualified as poor, slight, fair, moderate, substantial, or almost perfect according to the Kappa value. Fleiss (1982 see [11]) recommended 2 cut-off points at 0.40 an 0.75. A Kappa index below the 0.40 cut-off point is considered not clinically acceptable (poor agreement) while values greater than 0.75 represent excellent agreement. The Lerman index represents the probability of not obtaining the number of observed agreement by chance alone. Thus, one can set its own limit, say for example 10% of chance of obtaining the agreement by chance alone, which converts in a Lerman index of 0.9 or higher.

The range of values observed with the 3 indices indicated that Kappa had the widest range by far (0.21-0.94). In comparison, all the Lerman indices were between 0.70 and 0.83. Thus, the interpretation was more difficult since small changes may reflect large difference in reliability.

The use of weights in the Kappa computation did not change drastically the indices obtained with the 3-categories clinical indices, nor did it change the ranking. However, the weighted Kappa indices were higher than non-weighted Kappa when the clinical indices had 4 or 5 categories. Furthermore, the Kappa values obtained with the squared weights were better compared to those obtained with linear weights. The rationale of using weighted Kappa when evaluating reliability of ordinal clinical indices has been clearly shown (see Kingman 1986 [16] ; Cicchetti and Fleiss, 1977[6]). Yet, in light of the

improvement observed with different weighting methods, the choice of a set of weight should be justified, based on clinical ground.

Inference is easy to conduct with the RRI, since it is a proportion. The Lerman index being similar to the exact test of Fisher, inference is also possible. Statistical tests to make inference are available for Kappa statistic (Fleiss, 1982 see [11]).

As argued by Fleiss (1982 see [11]) and Lerman (Chandon & Pinson, 1981 [5]), the computation of an index of proximity should include somehow the expected agreement (agreement due to chance alone). The practical application exposed in this study did not clearly show the advantage of such a feature. The RRI index has no adjustment for expected agreement. When the number of categories is kept constant, the Lerman index is a transformation of the RRI index, and thus provide the same classification. When the number of categories and the sample size was kept constant, the classification was identical for all 3 indices except for two clinical indices with low prevalence, CC and MO2. On the other hand, the classification of all the objects, illustrated the disparity of the three classifications. The concordance between the classification could be improved slightly by discarding the clinical index CC. The clinical feature identified by this clinical index had the lowest prevalence, and this could explain the discordant results obtained. The Kappa statistic seems to control better for the expected agreement than the Lerman index does. This is illustrated by the ranking obtained by PD and CP. The expected agreement will decrease with an increasing number of categories in the clinical indices. Thus, these two clinical indices with 4 and 5 categories respectively, obtained the lowest rank according to the RRI index and low ranks according to Lerman index. However, the classification obtained with the Kappa statistic did not seem to be influenced by the number of categories.

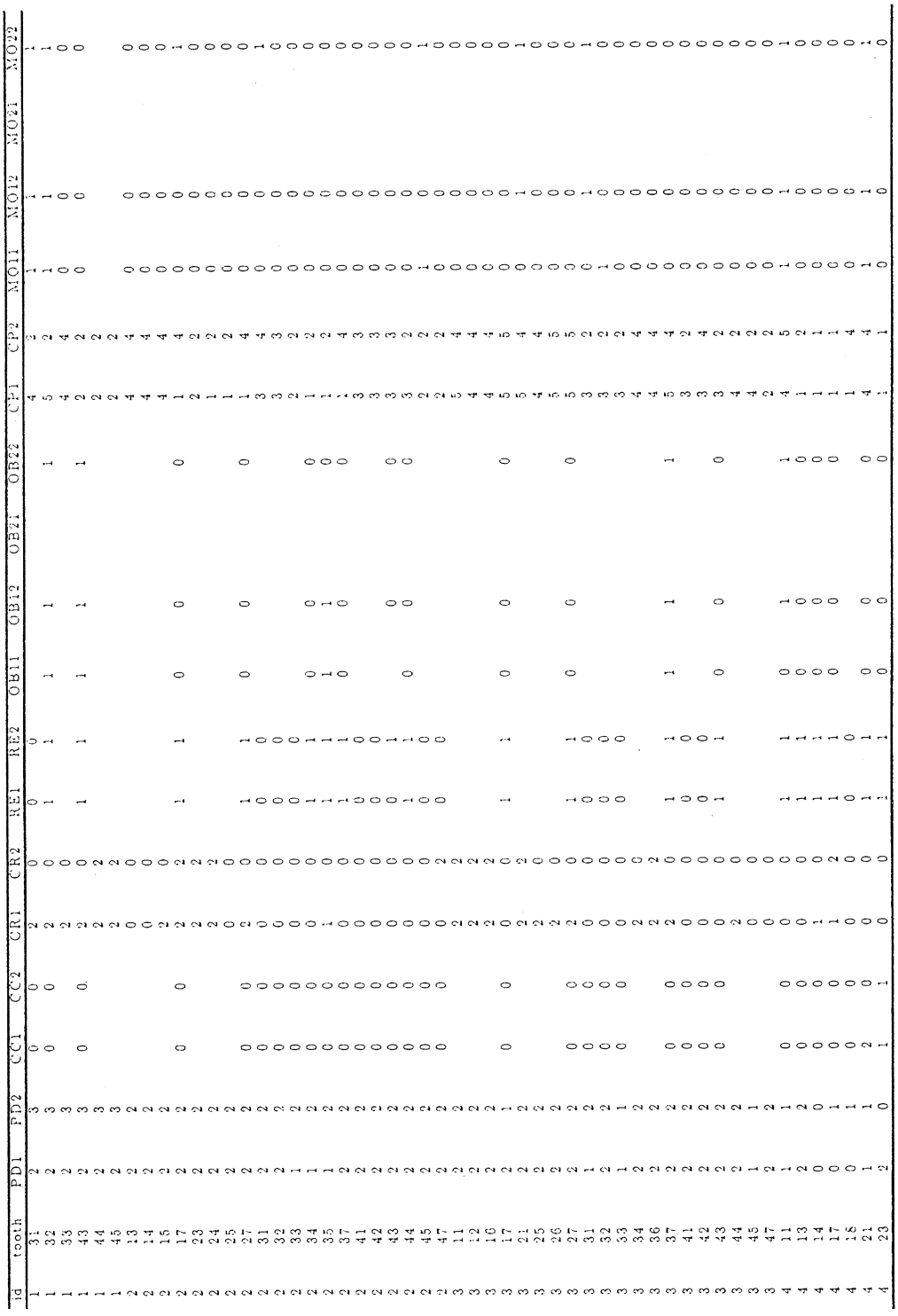
The main question was to evaluate whether the tested clinical indices were reliable enough to justify their use in epidemiological studies. Clearly, the clinical index MO2 scored poorly whatever the statistical method, and its use seems therefore questionable. On the other hand, there is little doubt that the clinical index RE was reliable. When comparing 2 clinical indices that evaluated the same clinical feature (MO1 versus MO2, OB1 versus OB2) all 3 methods agreed on which clinical index was more reliable. Apart from that, the three reliability measurements provided different answers. It is worth noticing that the Kappa index has become very popular, in psychology sciences and medicine at least, and to such an extend that a Kappa value is often a prerequisite to publish scientific reports. However, it seems that the choice of a statistical method should be base on the prevalence of the features or pathology identified, and on the complexity of the clinical index.

Acknowledgements :The author would like to thank Dr Gérard Antille for his great help and encouragement throughout the study.

This work was supported by the Swiss National Foundation for Research, Grants 32-35493.92.

Résumé: La fiabilité des évaluations cliniques reste une question essentielle en épidémiologie clinique. Il existe une variété de méthodes pour évaluer cette fiabilité mais aucune d'entre elles ne semble s'imposer face aux autres. La fiabilité de 9 indices dentaires a été évaluée en faisant examiner 15 patients par 2 dentistes et en répétant ces examens cliniques 2 semaines plus tard. La prévalence des caractéristiques ou maladies évaluées était comprise entre 1% et 53%. Pour chaque indice clinique, les indices de fiabilité selon Russel et Rao (RRI, 1940), et selon Lerman (1970) ainsi que l'indice Kappa (Cohen, 1960) ont été calculés. L'indice de Kappa pondéré selon 2 méthodes a également été calculé (Cicchetti et Fleiss, 1977). Les indices cliniques ont été classés selon les 3 méthodes statistiques, du pire au meilleur et les classifications ainsi obtenues ont été comparées. Il y avait de grandes disparités entre les 3 classifications. Les indices cliniques les plus élaborés ont obtenu des rangs bas avec le RRI et l'indice de Lerman, tandis que leur classement était meilleur selon l'indice de Kappa. L'introduction d'une pondération dans le calcul de l'indice de Kappa a encore amélioré les résultats de ces indices cliniques basés sur une échelle ordinale. Les trois méthodes ont permis d'identifier les indices cliniques les moins fiables. Il semblerait que le choix d'une méthode statistique devrait tenir compte de la prévalence du caractère ou pathologie évaluée et de la complexité de l'indice clinique.

7 The data



id	tooth	PD1	PD2	CC1	CC2	CR1	CR2	RE1	RE2	OB11	OB12	OB21	OB22	CP1	CP2	MO11	MO12	MO21	MO22
1	31	2	3	0	0	2	0	0	0					4	2	1	1	1	1
1	32	2	3	0	0	2	0	1	1	1	1		1	5	2	1	1	1	1
1	33	2	3	0		2	0							4	4	0	0	0	0
1	43	2	3	0	0	2	0	1	1	1	1		1	2	2	0	0	0	0
1	44	2	3			2	2							2	2				
1	45	2	3			2	2							4	4	0	0	0	0
2	13	2	2			2	0							4	4	0	0	0	0
2	14	2	2			2	0							4	4	0	0	0	0
2	15	2	2			2	0							4	4	0	0	0	1
2	17	2	2	0	0	2	2	1	1	0	0		0	1	4	0	0	0	0
2	23	2	2			2	2							2	2	0	0	0	0
2	24	2	2			2	2							1	2	0	0	0	0
2	25	2	2			2	0							1	2	0	0	0	0
2	27	2	2	0	0	2	0	1	1	0	0		0	1	4	0	0	0	1
2	31	2	2	0	0	0	0	0	0					3	4	0	0	0	0
2	32	2	2	0	0	0	0	0	0					3	3	0	0	0	0
2	33	1	2	0	0	0	0	1	1	0	0		0	2	2	0	0	0	0
2	34	1	2	0	0	0	0	1	1	0	1		0	1	2	0	0	0	0
2	35	1	2	0	0	0	0	1	1	1	0		0	1	4	0	0	0	0
2	37	2	2	0	0	0	0	0	0					3	4	0	0	0	0
2	41	2	2	0	0	0	0	0	0					3	3	0	0	0	0
2	42	2	2	0	0	0	0	0	0					3	3	0	0	0	0
2	43	2	2	0	0	0	0	1	1	0	0		0	3	3	0	0	0	0
2	44	2	2	0	0	0	0	1	1	0	0		0	3	2	0	0	0	1
2	45	2	2	0	0	0	0	0	0					2	2	0	0	0	0
2	47	2	2	0	0	0	2	0	0					2	2	0	0	0	0
3	11	2	2			2	2							5	4	0	0	0	0
3	12	2	2			2	2							4	4	0	0	0	0
3	16	2	2			2	2							4	4	0	0	0	0
3	17	2	1	0	0	0	0	1	1	0	0		0	5	5	0	0	0	1
3	21	2	2			2	2							5	4	0	1	0	1
3	25	2	2			2	0							4	4	0	0	0	0
3	26	2	2			2	0							5	5	0	0	0	0
3	27	2	2	0	0	2	0	1	1	0	0		0	5	5	0	0	0	0
3	31	1	2	0	0	0	0	0	0					3	2	0	1	0	1
3	32	2	2	0	0	0	0	0	0					3	2	1	0	0	0
3	33	1	1	0	0	0	0	0	0					3	2	0	0	0	0
3	34	2	2			2	0							4	4	0	0	0	0
3	36	2	2			2	2							4	4	0	0	0	0
3	37	2	2	0	0	2	0	1	1	1	1		1	5	4	0	0	0	0
3	41	2	2	0	0	0	0	0	0					3	2	0	0	0	0
3	42	2	2	0	0	0	0	0	0					3	4	0	0	0	0
3	43	2	2	0	0	0	0	1	1	0	0		0	3	2	0	0	0	0
3	44	2	2			2	0							4	2	0	0	0	0
3	45	1	1			2	0							4	2	0	0	0	0
3	47	2	2			2	0							2	2	0	0	0	0
4	11	1	1	0	0	0	0	1	1	0	1		1	4	5	1	1	1	1
4	13	2	2	0	0	0	0	1	1	0	0		0	4	2	0	0	0	0
4	14	0	0	0	0	1	0	1	1	0	0		0	1	1	0	0	0	0
4	17	0	1	0	0	1	2	1	1	0	0		0	1	1	0	0	0	0
4	16	0	1	0	0	1	0	1	0					1	4	0	0	0	0
4	21	1	1	2	0	0	0	1	1	0	0		0	4	4	1	0	1	0
4	23	2	0	1	1	0	0	1	1	0	0		0	1	1	0	0	0	0

id	tooth	PD1	PD2	CC1	CC2	CR1	CR2	RE1	RE2	OB11	OB12	OB21	OB22	CP1	CP2	MO11	MO12	MO21	MO22
4	23	2	0	1	1	0	0	1	1	0	0		0	1	1	0	0		0
4	27	1	1	0	0	2	2	1	1	0	0		0	4	4	1	1		1
4	31	0	0	0	0	0	0	0	0					3	1	0	0		0
4	32	1	1	0	0	0	0	0	0					3	1	0	0		0
4	33	0	1	0	0	0	0	1	1	1	1		1	2	2	0	0		0
4	35	1	1	0	0	0	0	1	1	1	1		1	2	1	0	0		0
4	36	2	1	0	0	2	0	1	1	1	1		1	4	4	0	0		0
4	37	1	1	0	0	0	0	1	1	0	0		0	4	4	0	0		0
4	41	0	1	0	0	0	0	0	0					3	2	0	0		0
4	42	0	0	0	0	0	0	0	0					3	2	0	0		0
4	43	0	1	0	0	0	0	0	1		0		0	3	2	0	0		0
4	44	1	1	0	0	0	0	0	1	1	1		1	1	2	0	0		0
4	45	1	1	0	0	2	0	1	1	1	0		0	2	2	0	0		0
4	46	1	1	0	0	0	0	1	1	0	0		0	4	4	0	0		0
4	47	2	1	1	0	0	0	1	1		0		0	4	4	0	1		1
5	33	3	2	0	0	0	0	0	1	0	0		0	2	2	0	0		0
5	34	3	2	0	0	0	0	1	1		0		0	2	2	0	0		0
5	43	3	2	0	0	0	0	0	0					2	2	0	0		0
5	44	3	2	0	0	0	0	0	0					2	2	0	0		0
6	11	2	2	0	0	0	0	0	0					2	2	0	0		0
6	13	2	2	0	0	0	0	0	0					3	2	0	0		0
6	14	2	2	0	0	0	0	0	0					2	4	0	1		1
6	15	2	2	0	0	2	0	0	0					4	4	0	1		1
6	21	2	2	0	0	0	0	0	0					1	1	0	1		1
6	23	2	1	0	0	0	0	0	0					2	1	0	0		0
6	24	2	1	0	0	2	0	0	1	0	0		0	2	4	0	0		0
7	31	2	2			0	0							4	4	0	0	0	0
7	32	3	2			0	0							4	4	0	0	0	0
7	33	3	2			0	0							4	4	0	0	0	0
7	34	2	2			0	0							3	3	0	0	0	0
7	35	2	2			0	2							2	4	0	0	0	0
7	38	3	2			0	0							5	5	0	0	0	0
7	41	2	2			0	0							4	4	0	0	0	0
7	42	2	2			0	0							2	4	0	0	0	0
7	43	3	2			0	0							2	4	0	0	0	0
7	44	2	2			0	0							2	1	0	0	0	0
7	45	2	2			0	0							1	1	0	0	0	0
8	11	2	2	0	0	0	0	1	1	0	1	0	1	2	2	1	0	1	0
8	12	1	1	0	0	0	0	0	0					2	2	0	0	0	0
8	13	2	2	0	0	0	0	0	0					2	2	1	0	1	0
8	14	2	2	0	0	0	0	1	1	1	1	0	1	1	2	0	0	0	0
8	15	1	1	0	0	0	0	0	1					1	2	0	0	0	0
8	16	1	1	0	0	0	0	1	1	1	1	0	1	4	4	0	0	0	0
8	17	1	1	0	0	0	0	1	1	0	0	0	0	4	4	0	0	0	0
8	21	2	2	0	0	0	0	1	1	0	0	0	0	2	2	1	0	1	0
8	22	2	2	0	0	0	0	0	0					2	2	1	0	1	0
8	23	2	2			0	0							1	2	0	0	0	0
8	24	2	2	0	0	0	0	1	1	1	0	0	0	4	5	1	0	1	0
8	25	2	2			0	0							5	6	2	2	2	2
8	26	2	2			0	0							4	4	0	0	0	0
8	27	2	2	0	0	0	0	1	1	1	0	1	0	4	4	0	0	0	0
8	31	2	2	0	0	0	0	1	0					3	3	0	0	0	0
8	32	2	2	0	0	0	0	0	0					3	3	0	0	0	0

id	tooth	PD1	PD2	CC1	CC2	CR1	CR2	RE1	RE2	OB11	OB12	OB21	OB22	CP1	CP2	MO11	MO12	MO21	MO22
8	25	2	2			0	0							5	6	2	2	2	2
8	26	2	2	0		0	0							4	4	0	0	0	0
8	27	2	2	0	0	0	0	1	1	1	0	1	0	4	4	0	0	0	0
8	31	2	2	0	0	0	0	0	0					3	3	0	0	0	0
8	32	2	2	0	0	0	0	0	0					3	3	0	0	0	0
8	33	1	1	0	0	0	0	1	1	1	1	0	1	2	4	0	0	0	0
8	34	1	1	0	0	2	2							2	4	0	0	0	0
8	35	2	2			0	0	0	0					2	4	0	0	0	0
8	38	2	2	0	0	0	0	0	0					3	3	0	0	0	0
8	41	2	2	0	0	0	0	0	0					3	3	0	0	1	0
8	42	2	2	0	0	0	0	0	0					3	3	1	0	0	0
8	43	2	2	0	0	0	0	0	0					4	3	0	0	0	0
8	44	2	2	0	0	0	0	1	1	0	0	0	0	4	4	0	0	0	0
8	45	2	2			0	0							2	4	0	0	0	0
8	46	2	2			0	0							4	4	0	0	0	0
8	47	2	2			0	0							4	2	1	0	1	0
9	11	1	1			0	0	1	1	0	0	0	0	1	1	0	0	0	0
9	12	1	2	0	0	0	0	1	1	0	0	0	0	1	1	0	0	0	0
9	13	1	2	0	0	0	0	1	1	0	0	0	0	2	1	0	0	0	0
9	14	1	2	0	0	0	0	1	1	0	0	0	0	2	1	0	0	0	0
9	15	1	2	0	0	0	0	1	1	0	0	0	0	1	1	0	0	0	0
9	16	1	2			0	0							1	1	0	0	0	0
9	21	1	1	0	0	0	0	1	1	0	0	0	0	1	1	0	0	0	0
9	22	1	1	0	0	0	0	1	1	0	0	0	0	2	2	0	0	0	0
9	23	1	1	0	0	0	0	1	1	0	0	0	0	2	2	0	0	0	0
9	24	1	1			0	2							1	2	0	0	0	0
9	25	1	1			0	0							1	1	0	0	0	0
9	31	1	1	0	0	0	0	1	1	0	0	0	0	1	1	0	0	0	0
9	32	1	1	0	0	0	0	1	1	0	0	0	0	1	1	0	0	0	0
9	33	1	1	0	0	0	2	1	1	0	1	1	1	1	1	0	0	0	0
9	34	1	1	0	0	0	2	1	1	0	1	0	1	1	1	0	0	0	0
9	35	1	1	0	0	0	0	0	1		0		0	1	1	0	0	0	0
9	41	1	1	0	0	0	0	0	1	0	0	0	0	1	1	0	0	0	0
9	42	1	1	0	0	0	2	1	1	0	1	0	1	1	2	0	0	0	1
9	43	1	1	0	0	0	2	1	1	1	1	1	1	1	1	0	0	0	0
9	44	1	2			0	0							1	1	0	0	0	0
9	45	1	2	0	0	0	0	1	1	0	0	0	0	1	1	0	0	0	0
10	13	0	2			0	0							2	2	0	0	0	0
10	16	3	3			2	2							2	2				
10	17	2	2			2	2							2	2				
10	23	3	3			2	2							1	2				
10	26	2	2			0	1							1	2	0	0	0	0
10	27	2	3	0	0	2	2	1	1	1	1	1	1	1	2	0	0	0	0
10	33	2	1			2	2							2	2				
10	34	2	1			2	2							2	2				
11	11	2	2			0	2							4	4	0	1	0	1
11	12	2	2			0	2							4	4	1	0	1	1
11	13	2	2			0	2							4	4	0	0	0	0
11	21	2	2			0	2							4	4	1	1	1	1
11	22	2	2			0	2							4	4	1	1	1	1
11	23	2	2			2	2							2	4	0	1	0	1
11	31	2	2	0	0	0	0	0	0					2	3	1	1	1	1
11	32	2	2	0	0	0	0	0	0					2	4	1	1	1	1

id	tooth	PD1	PD2	CC1	CC2	CR1	CR2	RE1	RE2	OB11	OB12	OB21	OB22	CP1	CP2	MO11	MO12	MO21	MO22
11	33	2	2	0	0	0	0	1	1	0	0	0	0	2	4	1	1	1	1
11	34	2	2			0	0							3	5	1	1	1	1
11	41	2	2	0	* 0	0	0	0	0					3	5	1	1	1	1
11	42	2	2	0	0	0	0	0	0	0	0	0	0	3	5	1	1	1	1
11	43	2	2	0	0	0	0	1	1	0	0	0	0	2	4	1	1	1	1
12	11	1	1	0	0	0	0	1	1	1	1	0	0	2	2	0	0	0	0
12	12	1	1	0	0	0	0	1	1	0	0	0	0	2	2	0	0	0	0
12	13	2	2			0	2							4	4	0	0	0	0
12	14	2	2			0	2							4	4	0	0	0	0
12	15	2	1			0	0							4	2	0	0	0	0
12	16	2	1			0	0							4	4	0	0	0	0
12	17	1	1	0	0	0	0	1	1	0	0	0	0	6	6	2	2	2	2
12	18	2	1	1	0	0	0	1	1	0	1	0	0	6	6	2	2	2	2
12	21	2	1	0	0	0	0	1	1	1	1	0	0	4	2	0	0	0	0
12	22	2	1		0	0	2							4	2	0	0	0	0
12	23	1	1		1	0	0							4	2	0	0	0	0
12	24	2	1		0	0	0							4	4	0	0	0	0
12	25	2	1		0	0	0							4	4	0	0	0	0
12	27	2	1		0	0	0							5	4	0	0	0	0
12	31	1	1	0	0	0	0	0	0					3	3	0	0	0	0
12	32	1	1	0	0	0	0	0	0					3	3	0	0	0	0
12	33	1	1	0	0	0	0	0	0					1	1	0	0	0	0
12	34	2	1	0	0	0	0	1	1	0	0	0	0	2	1	0	0	0	0
12	35	2	1	0	0	0	0	1	1	1	1	1	0	2	1	0	0	0	1
12	37	2	1	1	0	0	0	1	1	0	0	0	0	4	4	0	0	1	1
12	41	1	1	0	0	0	0	0	0					3	3	0	0	0	0
12	42	1	1	0	0	0	0	0	0					3	3	0	0	0	0
12	43	2	1	0	0	0	0	1	1	0	1	0	1	2	2	0	0	0	0
12	44	2	2		2	0	2							2	2	0	0	0	0
12	47	2	2		0	0	0							4	4	0	0	0	0
12	48	2	2	2	0	0	0	1	1	0	1	0	1	4	4	0	0	1	1
13	11	0	1	0	0	0	0	0	0					1	1	0	0	0	0
13	12	0	1	0	0	0	0	0	0					1	1	0	0	0	0
13	13	1	2	0	0	0	0	1	1	1	1	0	0	1	1	0	0	0	0
13	15	1	1	0	0	0	0	1	1	0	0	0	0	2	1	0	0	1	1
13	16	1	1	0	0	0	0	1	1	0	0	0	0	1	4	0	0	0	0
13	21	0	1	0	0	0	0	0	0					1	1	0	0	0	0
13	22	0	1	0	0	0	0	0	0					1	1	0	0	0	0
13	23	1	1	0	0	0	0	0	0					1	1	0	0	0	0
13	24	1	1	0	0	0	0	1	1	1	1	0	1	1	4	0	0	1	1
13	25	1	2	0	1	0	0	1	1	0	0	0	0	1	1	0	0	1	1
13	31	0	2	0	0	0	0	0	0					1	1	0	0	0	0
13	32	0	2	0	0	0	0	0	0					1	1	0	0	0	0
13	33	0	2	0	0	0	0	0	0	1	1	0	1	1	2	0	0	0	0
13	34	0	2	0	0	0	0	1	1	1	1	0	0	1	4	0	0	0	0
13	35	1	1	0	0	0	0	1	1	0	0	0	0	1	4	0	0	0	0
13	41	1	2	0	0	0	0	0	0					1	1	0	0	0	0
13	42	1	2	0	0	0	0	0	0					2	2	0	0	0	0
13	43	1	2	0	0	0	0	0	0					1	1	0	0	0	0
13	44	1	1	0	0	0	0	0	0					1	1	0	0	0	0
13	45	1	1	0	0	0	0	1	1	0	0	0	0	1	1	0	0	0	0
14	11	1	1	0	0	0	0	1	1	0	1	0	0	2	1	0	0	0	0
14	12	0	1	0	0	0	0	1	1	1	1	0	0	2	2	0	0	0	0

id	tooth	PD1	PD2	CC1	CC2	CR1	CR2	RE1	RE2	OB11	OB12	OB21	OB22	CP1	CP2	MO11	MO12	MO21	MO22
13	42	1	2	0	0	0	0	0	0					2	2	0	0	0	0
13	43	1	2	0	0	0	0	0	0					1	1	0	0	0	0
13	44	1	1	0	0	0	0	0	0					1	1	0	0	0	0
13	45	1	1	0	0	0	0	1	1	0	0	0	0	1	1	0	0	0	0
14	11	1	1	0	0	0	0	1	1	0	1	0	0	2	1	0	0	0	0
14	12	0	1	0	0	0	0	1	1	1	1	0	0	2	2	0	0	0	0
14	13	1	1			0	0							2	2				
14	14	1	1			0	0							2	2				
14	15	1	1			0	0	1	1	0	0	0	0	4	5	0	0	0	0
14	17	2	2	0	0	0	0	1	1	1	1	0	0	1	1	0	0	0	0
14	21	1	1	0	0	0	0	1	1	1	1	0	0	1	1	0	0	0	0
14	22	1	1	0	0	0	0	1	1	1	1	0	0	1	1	0	0	0	0
14	23	1	1	0	0	0	0	1	1	0	0	0	0	1	2	0	0	0	0
14	24	0	1	0	0	0	0							4	4	0	0	1	0
14	25	1	2	0	0	0	2							4	4	0	0	1	1
14	31	1	1	0	0	0	0	0	0					1	1	0	0	1	0
14	32	1	1	0	0	0	0	0	0					1	2	0	0	1	0
14	33	1	1	0	0	0	0	0	0					2	2	0	0	0	0
14	34	1	1	0	0	0	0	1	1	0	0	0	0	2	4	0	0	0	0
14	35	2	1			0	0							2	5	0	0	0	0
14	36	1	1	0	0	0	0	1	1	0	0	0	0	2	5	0	0	0	0
14	41	1	1	0	0	0	0	0	0					1	1	0	0	1	0
14	42	1	1	0	0	0	0	0	0					1	1	0	0	1	0
14	43	1	1	0	0	0	0	0	0					1	2	0	0	0	0
14	44	1	1	0	0	0	0	1	1	0	0	0	0	2	2	0	0	0	0
14	45	1	1			0	0							2	2	0	0	0	0
14	46	1	2			0	0							2	2	0	0	0	0
15	11	3	2	0	0	0	0	0	0					2	2	0	0	0	0
15	12	3	1	0	0	0	0	0	0					2	2	0	0	0	0
15	13	3	2			0	0							2	2				
15	18	3	2	0	0	0	0	1	1	0	0	0	0	6	6	2	2	2	2
15	21	3	1	0	0	0	0	0	0					2	2	0	0	0	0
15	22	3	2	0	0	2	0	0	0					2	2	0	0	0	0
15	23	3	2			0	0							2	2				
15	26	3	2	0	0	0	0	1	1	0	1	0	0	5	4	0	0	0	0
15	27	3	2	0	0	0	0	1	1	0	0	0	0	5	5	0	0	1	1
15	47	3	2	0	0	0	2	0	0					4	4	0	0	0	1
16	31	2	2	0	0	0	0	1	1	1	1	0	0	4	3	0	0	0	0
16	32	2	2	0	0	0	0	1	1	1	1	0	0	4	4	0	0	1	0
16	33	2	2	0	0	0	0	1	1	1	1	1	0	3	4	0	0	1	0
16	34	2	2	0	0	2	0	1	1	1	1	1	1	2	2	0	0	0	1
16	41	2	2	0	0	0	0	1	1	0	1	0	0	4	3	0	0	0	0
16	42	2	2	0	0	0	0	1	1	1	1	0	0	4	3	0	0	0	0
16	43	2	2	0	0	0	0	1	1	1	1	0	0	2	2	0	0	0	0
16	44	2	1			0	0							2	2				
16	47	2	3	0	0	0	0	1	1	1	1	1	0	4	2	0	0	1	0
17	11	1	1	0	0	0	0	1	1	1	1	0	1	2	1	0	0	0	0
17	12	1	1	0	0	0	0	0	0					1	1	0	0	0	0
17	13	1	1	0	0	0	0	0	0					1	1	0	0	0	0
17	14	1	1	0	0	0	0	0	0					2	1	0	0	0	0
17	21	1	1			2	0							2	2	0	0	1	1
17	22	1	1	0	0	0	0	0	0					2	2	0	0	0	0
17	23	1	1	0	0	0	0	0	0					1	4	0	0	0	0

id	tooth	PD1	PD2	CC1	CC2	CR1	CR2	RE1	RE2	OB11	OB12	OB21	OB22	CP1	CP2	MO11	MO12	MO21	MO22
17	24	1	1	0	0	0	0	0	0					1	1	0	0	0	0
17	31	2	1	0	0	0	0	0	0					3	3	0	0	0	0
17	32	2	1	0	0	0	0	0	0					3	3	0	0	0	0
17	33	1	1	0	0	0	0	1	1	0	0	0	0	1	1	0	0	0	0
17	37	1	1	0	0	2	0	1	1	0	0	0	0	2	2	0	0	0	0
17	41	2	1	0	0	0	0	0	0					3	3	0	0	0	0
17	42	2	1	0	0	0	0	0	0					3	3	0	0	0	0
17	43	1	1	1	0	0	0	0	0					1	4	0	0	0	0
17	47	2	2	0	0	0	0	1	1	0	0	0	0	1	2	0	0	0	0
18	31		2		0		0		0						4		0		1
18	32		2		0		0		1		0		0		4		0		0
18	33		2		0		0		1		1		0		4		0		0
18	41		2		0		0		1		0		0		4		0		1
18	42		2		0		0		1		0		0		4		0		1
18	43		2		0		2		1		1		1		5		0		0
19	11	1	1			0	0							2	2	0	0	1	1
19	12	1	1			0	0							4	4	0	0	0	0
19	13	1	1			0	0							3	3	0	0	0	0
19	14	1	2			2	2							4	4	0	0	0	0
19	15	1	2			0	2							4	4	0	0	1	0
19	21	1	1			0	2							4	2	0	0	1	1
19	22	1	1			0	0							2	2				
19	31	2	2	0	0	0	0	0	0					4	2	0	0	0	0
19	32	1	2	0	0	0	0	1	1	1	1	0	0	2	4	0	0	0	0
19	33	1	2		0	0	0							2	2	0	0	0	0
19	34	1	2			0	0							4	2	0	0	1	1
19	41	2	2	0	0	0	0	0	0					4	2	0	0	0	0
19	42	1	2	0	0	0	0	1	0	0		0		2	2	0	0	0	0
19	43	1	2			0	0							4	4	0	0	0	0
19	44	2	2			0	0							4	4	0	0	1	1

References

[1] Ambjörnsen E., Valderhaug J., Norheim P.W., Floystrand F. (1982). Assessment of an additive index plaque accumulation on complete maxillary denture. Acta Odontol. Scand. 40, 203-208.

[2] Bader J.D. and Shugars D.A. (1992). Understanding dentists'restorative treatment decisions. J. Pub. Health Dent. 52, 102-110.

[3] Bader J.D and Shugars D.A. (1993). Agreement among dentists' recommendation for restorative treatment. J. Dent. Res. 72, 891-896.

[4] Banting D.W. (1993). Diagnosis and prediction of root caries. Adv. Dent. Res. 7, 80-86.

[5] Chandon J.L. and Pinson S. (1981) In Analyse typologique, théories et applications, Masson, Paris.

[6] Cicchetti D.V. and Fleiss J.L. (1977)Comparison of the null disributions of weighted Kappa and the C ordinal statistic. App Psychol Meas; 1: 195-201.

[7] Cohen J. (1960). A coefficient of agreement for nominal scales. Educ. Psycho. Meas. 20, 37-46.

[8] Davies G.N. (1968).The different requirements of periodontal indices for prevalence studies and clinical trials. Int. Den. J. 18, 560-569.

[9] Dillon W.R. and Goldstein M. (1984). In Multivariate analysis: Methods and applications, John Wiley and Sons, New York, p 108.

[10] Fleiss J.L., Fischman S.L., Chilton N.W., Park M.H. (1979). Reliability of discrete measurements in caries trials. Caries Res. 13, 23-31.

[11] Fleiss J.L. (1982) The measurement of interrater agreement. In Statistical methods for rates and proportions, 2nd Ed, Wiley and Sons Eds, pp 212-236.

[12] Fleiss J.L. and Chilton N.W. (1983).The measurement of interexaminer agreement on periodontal disease. J. Periodont. Res. 18, 601-606.

[13] Fleiss J.L., Mann J., Paik M., Goultchin J., Chilton N.W. (1991). A study of inter- and intra-examiner reliability of pocket depth and attachment level. J. Periodont. Res. 26, 122-128.

[14] Howat A.P, Holloway P.J, Brandt R.S. (1981). The effect of diagnostic criteria on the sensitivity of dental epidemiological data. Caries Res.15, 117-123.

[15] Hunt R.J. (1986). Percent agreement, Pearsons correlation, and Kappa as measures of inter-examiner reliability. J. Dent. Res. 65,128-130.

[16] Kingman A. (1986). A procedure for evaluating the reliability of a gingivitis index. J. Clin. Periodontol.13, 385-391.

[17] Kingman A, Löe H, Anerud A, Boysen H. (1991). Errors in measuring parameters associated with periodontal health and disease. J. Periodontol.62, 477-486.

[18] Landis J.R. and Koch G.G. (1977) The measurement of observer agreement for categorical data. Biometrics 33, 159-174.

[19] Lerman I.C. (1970) In Les bases théoriques de la classification automatique, Gauthiers Villars, Paris.

[20] Mojon P., Favre P., Chung J.P., Budtz-Jørgensen E. (1995). Examiner

agreement on caries detection and plaque accumulation during dental surveys of elders. Gerodontology 12, 49-55.

[21] Mojon P., Chung J.P., Favre P., Budtz-Jørgensen E. (1996). Examiner agreement on periodontal indices during dental surveys of elders. J. Clin. Periodontol. 23, 56-59.

[22] Pitts N.B. and Fyffe H.E. (1988) The effect of varying diagnostic thresholds upon clinical caries data for a low prevalence group. J. Dent. Res. 67, 592-596.

[23] Poulsen S., Holm-Pedersen P., Kelstrup J. (1979) Comparison of different measurements of development of plaque and gingivitis in man. Scand. J. Dent. Res. 87, 178-183.