APPLIED STATISTICS IN PUBLIC HEALTH RESEARCH

16th March 1990

TNO Institute for Preventive Health Care (NIPG-TNO)

Division of Health Research
Netherlands Organisation for Applied Scientific Research

Leiden, the Netherlands

Organizing Committee:

Dr. C.C.J.H. Bijleveld
R.M. Cortel
J.J. Radder
Dr. J.L.A. van Rijckevorsel

TIME TABLE OF EVENTS

08.45	Registration and coffee	*Entrance Collegezaal*
09.45	Opening address	*Collegezaal*
10.00 - 11.00	Contributed papers	*Collegezaal*
11.00	Coffee break	*Entrance Collegezaal*
11.30 - 12.30	Contributed papers	*Collegezaal*
12.30	Lunch	*Foyer Bijlzaal*
14.00 - 15.30	Invited papers	*Collegezaal*
15.30	Tea break	*Entrance Collegezaal*
16.00	Invited paper	*Collegezaal*
16.30	Closing address	*Collegezaal*
17.00	Drinks and snacks	*Foyer Bijlzaal*

CONTENTS

MORNING-SESSION

Chair: J.L.A. van Rijckevorsel[1], F.D. Pot[2], S.P. Verloove-Vanhorick[3], C.L. Ekkers[4], A. Dijkstra[5]

08.45 Registration and coffee

09.45 Opening address by A. Dijkstra, deputy-director NIPG-TNO
Does public health research have its own methodology [1] ?

Contributed papers

10.00 F. Andries, C.C.J.H. Bijleveld
Multiple correspondence analysis of computer specialists' health complaints [2]

10.15 J.L.A. van Rijckevorsel, S. van Buuren
Imputation of missing data in a quality of life survey [2]

10.30 A.G.C. Vogels, M.M. van der Klaauw
Building statistical databases for an AIDS survey [3]

10.45 J.K.S. van Ginneken
Some applications of the life-table technique to determine indicators of health status [3]

11.00 Coffee break

11.30 H.M.E. Miedema
Nonlinear dose-response analysis with qualitative variables applied on odour data [4]

11.45 A. Bloemhoff
Nonlinear partial correlation analysis of health, social class and work indicators [4]

12.00 J.E.F.M. Frencken, M.A. van 't Hof
A mixed longitudinal design for monitoring prevalence of dental caries [5]

12.15 J.J. Radder, C.C.J.H. Bijleveld, E. Wortel
Nonlinear multivariate analyses of parental safety behaviour [5]

AFTERNOON-SESSION
Chair: J.L.A. van Rijckevorsel

12.30 Lunch
During the lunch break F.H.G. Marcelissen and D.J. van Putten will demonstrate the software package V-PROF for graphical reference profiles

Invited papers

14.00 D.N. Geary
Department of Statistics, Oxford University, UK
Dental trails and multivariate analysis

14.30 B. Goldfarb
Laboratoire de Biostatistique, Hôpital Necker, Paris, France
Spectral analysis in clinical and health research

15.00 P.A. Burrough
Netherlands Expertise Center for Spatial Information Processing
Rijksuniversiteit Utrecht, NL
Detecting geographical clusters of disease incidence

15.30 Tea break

16.00 G. Gallus
Instituto di Biometria E Statistica Medica
Facoltà di medicina e chirurgia, università degli studi
Milan, Italy
Statistical methods for monitoring rare health events

16.30 Closing address by D.A. Lievesley, director of the International Statistical Institute, Voorburg, NL
Common statistical problems in AIDS research, and why we must coordinate

17.00 Drinks and snacks

ABSTRACTS: CONTRIBUTED PAPERS

Does public health research have its own methodology ?

Atze Dijkstra, deputy-director NIPG-TNO*

* TNO Institute for Preventive Health Care

Public health is, according to Clark (1982) best identified as a social movement concerned with protecting and promoting the collective health of a community. This is its origin and this is its accomplishment. Public health can claim a key role in pioneering data-based approaches to the solution of health and community problems. It has generated the first practitioners of primary prevention. Trained as medical doctors, many public health officers intuitively practised insights which nowadays are claimed by medical sociology or health psychology.

Public health is at a crossroads where epidemiology and social science meet. Both disciplines share the same methodology of the empirical sciences. Their theories, concepts and operationalizations show remarkable differences. The fact that the clinical difference between disease and health is so much more marked than the social scientific difference between degrees of health or ill-health is reflected in the application of statistical methods.

Especially with respect to multivariate analysis epidemiologists and social scientists seem to embrace their 'own' methods. The variety in methods has impressively increased during the last decade. In the same period the application of multivariate designs in public health research has strongly increased.

These developments indicate that public health research is expanding. At the same time differentiation and specialization per discipline can hamper fruitful application of epidemiological and social scientific research results. To promote cross-fertilization, mutual understanding of each other's methodologies and research methods is a prerequisite. This workshop presents a variety of multivariate methods, applied in public health research. The aim is to make clear that all methods have their roots in the general methodology of the empirical sciences. It is of strategic importance for public health that, between disciplinary groups, methodological innovations are exchanged.

Address for correspondence: TNO Institute for Preventive Health Care, P.O. Box 124, 2300 AC Leiden, the Netherlands

Multiple correspondence analysis of computer specialists' health complaints; relations with career and job characteristics

Frank Andries*, Catrien C.J.H. Bijleveld**

* Dept. of Working Conditions, TNO Institute for Preventive Health Care
** Dept. of Statistics, TNO Institute for Preventive Health Care

Keywords: automation-personnel, job-assesment, health and well-being, multiple correspondence analysis

In 1989 a questionnaire was sent to over 5.000 persons employed in the automation branche. The aim of this inquiry was to make an inventory of possible bottle-necks in the career and working conditions of those contributing to the realization of an automation-product; each with respect to their own function and responsibility in this process. About 60 percent returned a completed questionnaire.
Using, among others, items from 'the NIPG standard-questionnaire occupation and health', we arrived at 40 items, refering to aspects of the job, health and well-being. The subjects' answers were aggregated over 32 occupations. When an occupation had an average score above the 80th percentile, the score was recoded to 2; under the 20th percentile to 0, and otherwise the score was recoded to 1. This datamatrix of occupations by recoded complaints scores, was analysed using Multiple Correspondence Analysis. The results of the inquiry show how automation-personnel typify and evaluate their career and present working conditions. An effort was made to identify risk-factors and functions-at-risk. Results show four clusters of occupations:
[1] Directors, advisers and marketing/sales-personnel; typified by a heavy workload, great responsibility and a sense of being hushed. Poor work by others and the frequent absence of certain persons contribute to the hectic character of the job. Otherwise, the job offers a challenge and, except for marketing/sales-personnel, relatively little stress is experienced.
[2] Middle-management; typified by a great workload, working under high pressure, annoyance as a result of unexpected situations, strain and fatigue. Functions at this level seem to be functions-at-risk, concerning stress.
[3] DP-specialists (System-analists; programmers, etc.); typified by lack of autonomy, especially concerning the method of working. The job offers less challenge than average and a lot of time is spent on study. Persons working in these functions are mostly at the beginning of their career.

[4] Lower-management, teachers and support-personnel; typified by complaints about direct superiors, salary and prospects. They experience less autonomy and challenge. Functions at this level seem to be functions-at-risk concerning career-possibilities (blind-alley-occupations).

Greenacre, M.J. (1984). *Theory and applications of correspondence analysis.* London: Academic Press.
Andries, F. (1990, to appear). *Automatiseren is mensenwerk.* [Computerisation is the work of humans.] Leiden: NIPG-TNO.

Address for correspondence: Dept. of Working Conditions, TNO Institute for Preventive Health Care, P.O. Box 124, 2300 AC Leiden, the Netherlands

Imputing missing data in a life style survey

Jan L.A. van Rijckevorsel*, Stef van Buuren**

* Department of Statistics, TNO Institute for Preventive Health Care
** Department of Psychometrics, Faculty of Social Sciences, University of Utrecht

Key words: missing data, imputation, log-linear model, hot-deck

Missing data are common and costly. Data with up to 30% missing are no exception in surveys measuring health. Obsolete pairwise or listwise deletion to accommodate for missing data amounts to wasting labour intensively collected material. An alternative is to fill in missing data with appropriate replacements. Doing this based on external information is called "cold-deck" imputation and, based on the observed data, "hot-deck" imputation. Advantage of imputation is that after imputation all standard statistical techniques can be applied, meanwhile the response bias is reduced and the distribution of the population is preserved. A disadvantage is that new biases can be introduced. In this paper we propose a fast hot deck imputation method for categorical data.

The life style survey pertains to approx. 5000 respondents. Considering only a few variables and a perfectly simulated dependency structure, artificially created missing data patterns of more than 5% create havoc in any model estimation of the completed data. Stepwise imputation of first the independent and next the dependent variable shows a substantial gain in model fit. The model used is a very simple loglinear model and the imputation is based on maximizing the non-stochastic simultaneous homogeneity of all observed variables. Non-probabilistic imputation of missing data based on several categorical variables simultaneously is relatively new and no fast ready made techniques are (yet) available.

De leefsituatie van de Nederlandse bevolking 1983: kerncijfers. [The life style of the Netherlands population 1983: statistics.] Centraal Bureau voor de Statistiek, Hoofdafdeling Sociaal-Culturele Statistieken. 's-Gravenhage: Staatsuitgeverij, CBS-publicaties (1983).

Address for correspondence: Dept. of Statistics, TNO Institute for Preventive Health Care, P.O. Box 124, 2300 AC Leiden, the Netherlands

Building statistical databases for an AIDS survey; controlling complexities in data collection

Ton G.C. Vogels*, Theo M. van der Klaauw**

* Section of Child & Adolescent Health Care, TNO Institute for Preventive Health Care
** Dept. of Statistics, TNO Institute for Preventive Health Care

Key words: statistical databases, HIV, adolescents

The study 'Health, Behaviour and Relations among Adolescents' was designed to provide for a strong empirical base for adequate health education with regard to sexually transmitted diseases in general and AIDS in particular; secondly, it was intended to provide for a better estimation of the risks for adolescents of HIV-contagion.
In this study, data on sexual behavior, attitudes and knowledge are collected by means of a questionnaire that is to be completed in the classroom.

The process of data collection is complicated by several factors. The study aims to generalize over sex, age, type of education, religious orientation of schools involved and social economic status, for 6 distinct Dutch regions as well as for the country as a whole. Several relevant variables (e.g. homosexuality and sexual experience) are expected to have little variation; to ensure a sufficient number of observations a large sample (n=12.000) had to be approached. Data collection required cooperation of school administrators. Because of the delicate content of the questionaire a rather high refusal rate was expected related to religious orientation of schools. The research questions implied that many concepts had to be investigated. Yet, for practical reasons, the questionaire had to be completed within 45 minutes. The data had to be collected in cooperation with more than 45 different organizations (the Sentinel Stations for Youth Health Care); data collection had to be completed within 3 months.
To ensure a controlled collection of data on all relevant variables, several measures had to be taken. These include using 8 different versions of the questionnaire on the basis of 4 different modules and 2 age groups and the building of APRI (Automatic Project Administration System) by means of which the representativeness of the sample was safeguarded.
These measures and their consequences will be discussed.

Vogels, T. & Danz, M. (1989). *Gezondheid, gedrag en relaties. Draaiboek.* [Health, behaviour and relations. Scenario.] Leiden: NIPG-TNO.
Vogels, T., Van der Klaauw, Th. & Van Laarhoven, M. (1989). *Werken met APRI. Handleiding en procedures.* [Using APRI. Manual and procedures.] Leiden: NIPG-TNO, internal report.
Vogels, T., Van der Vliet, R.W.F., Danz, M.J. & Hopman-Rock, M. (1990). *Verslag vooronderzoek Gezondheid, Gedrag en Relaties; een onderzoek van de Peilstations Jeugdgezondheidszorg.* [Report preliminary research Health, Behaviour and Relations; a study by the Sentinel Stations for Youth Health Care]. Leiden, NIPG-TNO/NISSO.

Address for correspondence: Section of Child & Adolescent Health Care, TNO Institute for Preventive Health Care, P.O. Box 124, 2300 AC Leiden, The Netherlands

Some applications of the life table technique in the analysis of health status

Jeroen K.S. van Ginneken*

* Dept. of Public Health and Epidemiology, TNO Institute for Preventive Health Care

Key words: life table, mortality, morbidity, life expectancy

The most common application of the life table method is in the field of mortality; the result of such a life table is the life expectancy at birth and at other ages. During the past 10 to 20 years a number of techniques have been developed that make the life table method also suitable for analysis of causes of death and morbidity. In the presentation several of these newer applications will be discussed and examples of their use will be given. One of these applications with respect to causes of death, is the determination of gains in life expectancy due to elimination of a cause of death. More recent innovations are determination of life expectancy of those who die of a particular cause, and gains in life expectancy due to elimination of a cause for the so-called saved population. Another application of the life table method is in the field of morbidity where it has been used to construct a measure or index which combines information on both mortality and morbidity. An example of such a measure is the life expectancy free of disability.

Van Ginneken, J.K.S., Bannenberg, A.F.I. and Dissevelt, A.G. (1989). *Gezondheidsverlies ten gevolge van een aantal belangrijke ziektecategorieen in 1981-1985: methodologische aspecten en resultaten.* [Loss of health due to a number of important diseases in 1981-1985: methodological aspects and results.] Leiden/Voorburg: NIPG-TNO/CBS.
Wilkins, R. and Adams, O.B. (1983). Health expectancy in Canada in the late 1970s: demographic, regional and social dimensions. *American Journal of Public Health, 73*, 1073-1080.

Address for correspondence: Dept. of Public Health and Epidemiology, TNO Institute for Preventive Health Care, P.O. Box 124, 2300 AC Leiden, the Netherlands

Non-linear dose-response analysis with qualitative variables applied on odour data

Henk M.E. Miedema*

* Dept. of Social Scientific Environmental Research, TNO Institute for Preventive Health Care

Key words: dose-response, canonical analysis, optimal scaling, non-linear, fuzzy coding

In dose-response analysis a set of dose variables is related to a set of response variables. In the simplest case each set contains only one variable. The method presented is especially suited for applications in which qualitative variables are involved (e.g. type of source as one of the dose variables or subjective health reports as response variables), and/or in which the dose-response relation may be non-linear. Doses and responses are transformed to become maximally alike. Several types of restrictions can be imposed on the respective functions that transform the doses and responses. It may be required, for instance, that the transformation of the doses is an additive combination of the transformations of the individual doses. The interpretation of the transformations of the functions depends on whether the variables are qualitative or quantitative. Applications on an artificial data set and on dose and reponse data from environmental odour research are presented.

Gifi, A. (1981). *Nonlinear multivariate analysis*. Leiden: Dept. of Datatheory FSW.
Miedema, H.M.E. (under review) Equalization by scaling and transformation.

Address for correspondence: Dept. of Social Scientific Environmental Research, TNO Institute for Preventive Health Care, P.O. Box 124, 2300 AC Leiden, the Netherlands

Nonlinear partial canonical correlation analysis of health, social class and work indicators, a comparison with Cumulative Incidence Ratios

Anneke Bloemhoff*

* Dept. of Epidemiology and Occupational Health Care, TNO Institute for Preventive Health Care

Key words: social class, disability, working conditions, nonlinear partial correlation analysis

The impact of social class on disability is unclear. On the one hand the disabled are characterised as older labourers with little education, who have done hard physical labour for a long time. On the other hand there is the teaching-profession with a high incidence of disability; the educational level of teachers is relatively high and they are under a mental rather than physical strain.
As a part of the research programme on socio-economic differences in health (financed by the Ministry of WVC), the NIPG/TNO has initiated a study to analyse the relationships between social class and disability. The objectives of this study were twofold: (1) to analyse the relationship between social class and incidence of disability, and (2) to estimate disability risk of social class, both adjusted for the confounders age and working conditions.
The data pertained to a historical cohortstudy by De Winter (in preparation), where 2791 male employees of four companies completed a questionnaire on person- and job-characteristics, perceived working conditions and health. During a follow-up period of five years, data became available on the incidence of disability.
The following variables were included:
1. education level (four levels) as an indicator of social class;
2. incidence of disability;
3. age (two levels);
4. work-strain (three levels);
5. work-environment (three levels);
6. job-organisation (three levels);
7. management and colleagues (three levels);
8. appreciation of the job (three levels).

Two different analysis methods were used. The first method consisted of estimating the nonlinear partial canonical correlation (Van der Burg, 1985, pp.63-69) between disability and educational level, stratified by age, while adjusting for the covariates working conditions.
The second method consisted of the multivariate estimation of independent Cumulative Incidence Ratio's (RR) of education level, age and working conditions.

The results of both methods will be discussed and compared. The conclusion is that social class has a limited inpact on disability (partial rho = 0.09), with age (RR=4.6) and work-strain (RR=2.8 and RR=1.8) as significant confounders.

De Winter, C.R. (in preparation). *Afscheid van de werkplek; verzuim en werknemersmeningen over arbeid en gezondheid als voorspellers van uitval uit het werk.* [Absenteeism and workers' opinions on their work and health to predict job drop-out] Leiden: NIPG-TNO.
Van der Burg, E. (1985). *CANALS.* Leiden: Dept. of Datatheory FSW.

Address for correspondence: Dept. of Epidemiology and Occupational Health Care, TNO Institute for Preventive Health Care, P.O. Box 124, 2300 AC Leiden, the Netherlands

The application of the mixed-longitudinal study design in monitoring the prevalence of dental caries

Jo E.F.M. Frencken*, Martin A. van 't Hof**

* Dept. of Community Dental Health and Epidemiology, TNO Institute for Preventive Health Care
** Dept. of Medical Statistics, University of Nijmegen.

Key words: development, mixed-longitudinal design, age-period-cohort analysis, dental caries

In studying developmental processes, variables other than chronological age must often be taken into account if differences in the developmental patterns of the group under study are to be characterized. In particular, in addition to age, appropriate models for developmental studies may have to incorporate cohort and period effects (Van 't Hof et al., 1976, Palmore, 1978)
The more traditional designs keep one of these time components constant and this results in the confounding of the two factors remaining. In this way age effects are confounded with cohort effects in the cross-sectional design and with period effects in the pure longitudinal design, while cohort and period effects are confounded in the time-lag design. The mixed-longitudinal design was developed to overcome the difficulties inherent in each of the more traditional approaches. This design however, does not completely avoid the problem of confounding, although it does provide a structure in which isolation of the contribution of age, period and cohort effects can be achieved (Van 't Hof et al., 1976). The Age-Period-Cohort procedure intends to separate the effects of the three time-components age, period and cohort. However, for confirmation and/or adjustment, external (prior) information must always be taken into account (Kleinbaum et al., 1982, pp. 130-134).
The mixed-longitudinal design and A.P.C.-procedure (in this study using analysis of covariance) where used to detect a possible upward trend of dental caries experience in a child population between 1984 and 1988. It resulted in a strongly significant age effect (which served as prior information, as caries is considered an age related disease in children), a period effect (change in the application of the diagnostic criteria over the 4 year period) and a cohort effect (decrease in the caries experience).

Kleinbaum, D.G., Kupper, L.L. and Morgenstein, H. (1982). *Epidemiologic Research, Principles and Quantitative Methods*. New York: Van Nostrand Reinhold.
Palmore, E.(1978). When can age, period and cohort be separated. *Social Forces, 57*, 282-295.
Van 't Hof, M.A., Prahl-Andersen, B. and Kowalski, C.J. (1976). A model for the study of developmental processes in dental research. *Journal of Dental Research, 55*, 359-366.

Address for correspondence: Dept. of Community Dental Health and Epidemiology, TNO Institute for Preventive Health Care, P.O. Box 124, 2300 AC Leiden, the Netherlands

Multiple correspondence analysis, nonlinear multiple regression and nonlinear discriminant analysis of parental safety behaviour

J.J. Radder*, C.C.J.H. Bijleveld*, E. Wortel**

* Dept. of Statistics, TNO Institute for Preventive Health Care
** Dept. of Public Health and Epidemiology, TNO Institute for Preventive Health Care

Key words: parental safety behaviour, multiple correspondence analysis, nonlinear regression

We analyze the relation between parents' reported safety behaviour and six determinants of that behaviour (Wortel, Ooijendijk & Stompedissel, 1988). As the determinants were mixed dichotomous and nominal variables, and as the reported safety behaviour had been recorded on ordinal measurement level, and as no obvious analysis method for this type of problem stands out, several nonlinear analyses were performed to investigate the relation between safety behaviour and its determinants. Measurements had been obtained from 1129 parents with preschool children for several safety measures; we chose to analyze the safety behaviour 'removing teapot/coffeepot', in relation to the confounding variable 'education' and six determinants. The safety behaviour variable had three categories: 'unsafe behaviour', 'relatively safe behaviour' and 'safe behaviour'. A total of 754 parents had no missing measurements, these were selected for further analysis.
In order to explore the supposed confounding effect of education, we performed multiple correspondence analysis (Greenacre, 1984). After adding an interactively coded combination of education and safety behaviour, the analysis showed that such an interaction between education and safety behaviour was absent. In fact, education played little or no role at all.
Those who themselves, or whose partner thought safety measures unnecessary, exhibited unsafe behaviour; those who thought that they would not succeed, who considered the child too old for safety measures, and, to a lesser extent, who thought safety measures wouldn't help and would be an inconvenience, exhibited relatively safe behaviour.
For investigating the relative importance of the six determinants and the confounder education in predicting safety behaviour, we performed nonlinear multiple regression analysis (Gifi, 1981) on the same variables, thereby relating safety behaviour on the one hand to the behaviour determinants on the other hand. It turned out that 'succeed', 'necessary' and 'partner' were identified as

most important. Again, education had no relation whatsoever with safety behaviour.
Nonlinear discriminant analysis (Van der Burg, 1985, pp. 82-85) served to identify those variables on which the safety behaviour categories could be distinguished. On the first discriminant dimension, safe behaviour was contrasted with less safe and unsafe behaviour; variables that contributed most to this contrast were again 'succeed', 'necessary' and 'partner'. On the second discriminant dimension, relatively safe behaviour was contrasted with unsafe behaviour; the variable that contributed most to this contrast was partner: only the unsafe reported that they think their partner does not find the safety measures necessary. Thus, the various nonlinear analyses supported the same global conclusions; the various analyses highlighted different interrelationships between the variables.

Greenacre, M.J. (1984). *Theory and applications of correspondence analysis.* London: Academic Press.
Wortel, E., W.T.M. Ooijendijk and I. Stompedissel (1988). *Preventie van privé-ongevallen bij 0-4 jarigen.* [Prevention of preschool children's accidents at home] Leiden: TNO Institute for Preventive Health Care.
Van der Burg, E. (1985). *CANALS.* Leiden: Dept. of Datatheory FSW.

Address for correspondence: Dept. of Statistics, TNO Institute for Preventive Health Care, P.O. Box 124, 2300 AC Leiden, the Netherlands

Graphical reference profiles: a method for analysis and presentation of questionnaire data

F.H.G. Marcelissen*, D.J. van Putten*

* Dept. of Epidemiology and Occupational Health Care, TNO Institute for Preventive Health Care

Background

Many Occupational Health Centers in the Netherlands organize "periodic medical and workplace surveys". During such a survey a self administered questionnaire is used for obtaining insight into the health problems and working conditions as experienced by the employees. Together with other available information on health and work (eg. biometric data and workplace survey) the questionnaire provides the occupational physician and industrial hygienist with basic information about health and work.

Graphical Reference Profiles

Instruments and methods for the analysis, interpretation and presentation of large numbers of items and subscales from questionnaires are not readily available for every day (occupational) practice. We developed the Graphical Reference Profile (GRP) method, which has the following advantages:

- it gives a comprehensive visual presentation of series of items and subscales;
- the difference between scores of different studygroups is shown;
- confidence intervals are shown to facilitate inferences;
- confouders can be corrected for by means of indirect or direct standardization;

The method will be illustrated using a menu driven software package.

Applications

Although the questionnaire is developed for use in Occupational Health Care settings, it can also be applied to other settings using other questionnaires.

Address for correspondence: Dept. of Epidemiology and Occupational Health Care, TNO Institute for Preventive Health Care, P.O. Box 124, 2300 AC Leiden, the Netherlands

ABSTRACTS: INVITED PAPERS

Dental trials and multivariate analysis

D.N. Geary*, E. Huntington**, R.J.Gilbert**

* University of Oxford
** Unilever Research, Port Sunlight Laboratory, Merseyside

Some background on dental disease is presented.
Four dental clinical trials conducted between 1974 and 1986 are considered. The trials compare toothpastes, with three or six pastes per trial. Between 1400 and 3000 children enter each trial. Complete data for a given child consist of five oral health variables measured on each of three or four occasions, over a period of three years. The five variables are plaque, DEFS (decayed surfaces), DEFT (decayed teeth), gingivitis and calculus.
The complete data are transformed towards multivariate normality. Multivariate analyses of covariance (MANCOVAs) show up significant effects.
The MANCOVA analyses are based on data from 72% of the children who entered the trials: only those who provided complete data measured by the same clinician each year. Missing data are examined for any lack of randomness which might invalidate conclusions from the MANCOVAs. It turns out that no serious imbalance occurs between pastes in any of the trials, with respect to missing values, though a significantly greater proportion of boys than girls have missing values in one of the trials.
If a trial could somehow be limited to involve only subjects who are most likely to show up differences between toothpastes, this could lead to some economy. This possibility, of economizing by involving only selected subjects in the trials, is investigated. In the given trials, pre-selection of only girls, or only children who brush their teeth more than twice per day on average, seems to discriminate almost as well as when no selection is imposed and data from all subjects are used.
Differences in paste effects with respect to decay might become statistically significant in a caries trial before the standard three years' duration. If so, a decision may be made to stop the trial early, claiming a significant effect and saving time, effort, money and tooth decay.
Sequential testing for paste differences, based on (univariate) caries increments, is carried out retrospectively for each of the four trials. This testing takes account of the repeated measurements of caries from year to year. In two of the trials it may have been possible to draw 'safe' conclusions about paste differences in less than three years.

Address for correspondence: Department of Statistics, University of Oxford, 1 South Parks Road, Oxford, OX1 3TG, UK

Spectral analysis in clinical and health research

B. Goldfarb*

* Laboratoire de Biostatistique et d'Informatique Medicale, Hôpital Necker, Paris, France

I Principles

Time is a reference variable for epidemiological studies, as for clinical trails. For a long time it has been considered as a fixed factor, although many parameters or indices present cyclic variations and cannot be compared only at specified moments. In such situations, repeated measures along time must be recorded and analyzed. Such chronologic studies give series of highly correlated data for which conventional methods are unsuitable. Spectral analysis looks for the basic cyclic components of a process or of a time series, and for their relative contribution to the total variance. A Fourier-like decomposition of the time series gives the so-called periodogram which is constituted by a succession of peaks. The variance of the process is "explained" by the corresponding frequencies. Thus, as in any frequency spectrum, each rhythm is represented by its frequency, reciprocal of its period. A particular process is the white noise, or discrete purely random process, consisting of a time-ordered sequence of mutually independent and identically distributed random variables. It appears then as the reference for non-cyclic behaviour; it has been chosen to test the absence of periodic component. When a cyclic component (identified by the highest peak of the periodogram) has been recognized as significant, a differentiating filter may be applied to the series. The spectrum of the new series presents the same maxima except the one used for filtering. Identification and testing can be repeated until only white noise results.

II Application

Some complications in chronic haemodialysis might be attributed to the dialysis membrane. But the notion of bio-incompatibility remained mainly based on observations restricted to dialysis sessions; developments all along interdialytic phases were not taken into account. A trail was designed to define more precisely bio-incompatibility on the basis of the evolution of body temperature during one week, compared in two groups of patients dialyzed on two different types of membranes: Cuprophane (CUP) and Polyacrylonitril AN-69 (AN-69). This clinical criterion was chosen as it summarizes the knowledge on biological perturbations. Series of 29 observations (7 days with 4 measures each, plus one) were analyzed on 32 patients. Curves of temperatures reveal the well-

known circadian rhythm, and also peaks of temperature during the twelve hours following dialysis. Classical comparisons between the two groups show only a significant difference in the hyperthermia following dialysis, and a significant difference in dialythermic amplitudes. The periodograms obtained from the raw data of the two groups, show the circadian rhythm, and only in the CUP group (which presents the relative hyperthermia following dialysis) a secondary peak. We have differentiated the series according to the 24-hour cycle. The resulting series were compared first to a white noise. In the CUP group it was significantly different from a white noise, while it was not in het AN-69 group. The periodogram of the filtered series in the CUP-group shows a secondary cycle that corresponds to the succession of dialysis sessions. We can then conclude the existence of two added cycles, the physiologic circadian rhythm and a dialysis-induced rhythm, for patients dialyzed on a bio-incompatible material, while on another type of material the classical thermic regulation is not modified by an extra cycle. Subgroups and individual analysis proved the sensitivity of spectral analysis.

III Conclusion

Spectral analysis appears as the appropiate specific methodology for any study involving a criterion with known or suspected cyclic variations. It may be used to investigate the evolution of biological functions, epidemiologic incidence data, the presence of side-effects in clinical trails, and also in case of repeated administrations of treatments along time. Graphical aspects enhance its applications. However some statistical problems remain to be solved, especially concerning robust methods of estimating the spectrum, and identifying outliers.

Address for correspondence: Laboratoire de Biostatistique et d'Informatique Medicale, Hôpital Necker, 149 Rue de Sèvres, 75015 Paris, France

Detecting geographical clusters of disease incidence

P.A. Burrough*

* Netherlands Expertise Center for Spatial Information Processing, Rijksuniversiteit Utrecht, NL

Introduction: the problem of detecting clusters

Rare diseases, such as certain cancers, can be caused by many factors. In older people who have worked under many different conditions, who have travelled and maybe, moved house several times, it is almost impossible to establish exactly any direct link between a geographical source of a potential cancer-inducing factor in the environment and incidence of disease. With rare child cancers, however, the possibility of establishing some kind of link is potentially feasible because children are more likely than adults to have spent most of their lives in one geographical area (a house, street, district).

When apparent clusters of rare diseases occur, people are often quick to attribute them to nearby sources of perceived evil - the nuclear reprocessing plant, the garbage incinerator chimney, the dioxene polluted effluent from the garbage tip that drains into the local pond. Before such intuitive and often emotional associations can be accepted as even moderately likely, there are at least two important points to be considered. The first is: is it scientifically feasible that the perceived evil is capable of causing the observed disease? This is a problem for the medical specialists and I do not propose to go further with it. The second is: is there an unusually high (possibly statistically significant) incidence of the disease at the location under consideration? The problems of detecting and determining the significance of clusters of rare events is the subject of this paper.

Point pattern analysis

When a rare child cancer is detected, the incidence can be recorded as an event occurring at a given location in geographical space (the child's home). In principle the data of all events for a given time period can be seen as a set of all points with coordinates X and Y, with an attribute A covering the geographic area under consideration. If A = 0 then the disease did not occur; if A = 1, then the disease has been detected. Mapping all points with each point coloured white for A = 0, and black for A = 1 will give a point distribution map which can be analysed by eye. Although such a map may apparently show clusters, we cannot say how significant or important they are unless we know something about the way the data have been collected and stored, (in particular the spatial units of aggregation), and the probability of the disease occurring. We should be particularly careful to avoid deriving

hypotheses from a data distribution, and then testing them, because of bias. It is much better to use an independent test of significance to test for the presence of clusters.

Data aggregation

Before the days of electronic databanks it was not possible to record all data about a population with reference to a detailed geographical coordinate system. Therefore data about all aspects of the population were linked to local authority areas, census districts etc. This caused local variations in incidence to be smoothed away. If the number of incidences of a condition in these aggregated units was large, and they were distributed evenly over the unit, then the units were a good way of reducing data complexity. Units such as local authority areas or census tracts are good ways of displaying the variations in density of numbers of jobless, per capita income, house prices, political views, at a sub-national or regional scale. They are not suitable for displaying the variation of properties that may vary spatially within the aggregated units. For example a map of European countries showing population densities will not show the clustered population areas of S.E. England, Paris and the Ruhr. In particular, if a disease is estimated to occur in one person in 10.000, then if two enumeration areas A and B have populations of 100.000 and 20.000 respectively, then they are likely to be 10 and 2 incidences, respectively. If these incidences are smeared out over the whole area, then no clusters can be detected. If the location of the disease incidence is more precisely known, then it is possible to ask whether or not the incidences are clustered.

As data recording and data storage techniques improve it is no longer necessary, nor desirable to aggregate social and epidemiological data before analysis. Many countries are building large, detailed data sets so the problem of spatial aggregation before analysis can be removed. But it is essential to consider spatial aggregation during analysis. A single incidence, tied to a single household, is not a cluster. Ten incidences spread over a city of 500.000 people may be too small to be thought of as a cluster. Clearly we should adopt a spatial unit that is independent of the bureaucratic organization. One solution is to use a circle of a given radius as the basic spatial unit. A grid of points is laid over the area in question and for each point the circle is laid over it. The number of incidences failing within the circle are counted and tied to the point. The procedure is repeated for each point on the grid.

In order to avoid problems of changing spatial resolution, the analysis is repeated for a series of incrementally increasing circles. Each circle of larger radius results in a different set of counts. This is computationally tedious, but quite possible on modern computers, as Openshaw (1988) has demonstrated.

Testing for clusters

The significance of each count at each grid point for each circle radius can be obtained by comparison with a statistical model expressing the probabilities

that the disease would occur purely at random. Although more statistical work needs to be done here, Openshaw (1988) suggests that a Poisson distribution is suitable as the basis for the test statistic. Each count at each grid point is then compared with the test statistic in order to evaluate the likelihood that such a count value occurs by chance alone (Miller and Kahn 1962, page 380). Values that have a low probability of occurring (say 2 in 1000 or p = 0.002) can be highlighted and displayed as a map.
Openshaw (1988) describes the construction of a prototype Geographical Analysis Machine (GAM) which derived counts, tested probability levels and displayed results as circles on maps. When probability levels did not exceed 0.002 then no circle was drawn. Regions of densely overlapping circles, on the other hand indicate places where clusters not only occur, but where the clusters are relatively insensitive to spatial aggregation as given by circle diameters. Openshaw considers that these "hot spots" are the real clusters that have been determined independently without reference to external hypotheses or local perceived evils. In his examples he not only detected the much disputed Sellafield "hot spot" for acute lymphoblastic leukaemia, but also there was a very strong suggestion that a previously unnoticed cluster occurs in the area of north east England known as Tyneside. The presence of such stable clusters allows hypotheses to be set up about the relationship between the cancer incidence and environmental factors which can be examined further.

Criticism of GAM and further developments

No method of analysis can be better than the data used, and it is likely that all large data sets may contain mis-typed data or poorly-referenced site information. In many cases the data used for analysis are only a sample. Unless point counts are related to geographical coordinates there will always be an extra source of error. If for example, locations are recorded by postal codes or other non-exact means. So far, very little work seems to have been done on comparing data collected at different points in time, or exploring the consequences of using data in which disease incidence changes through time. The method of significance testing is an area for statistical research. Although the Poisson distribution seems to be a sensible replacement for the Monte Carlo simulation used originally, there is as yet, no unambiguous means of setting up the test statistic. This is clearly an area of further research.
Openshaw (1988) has suggested various developments of GAM, including a super-computer version to speed analysis, and a system that would work automatically, albeit much slower, on an 80383-based personal computer. He has also suggested modifying the spatial search to include certain target populations around each point in order to overcome rural-urban differences. Another alternative is to adjust circle diameters so that they always contain a fixed number of incidents, and then to assess significance. In both cases, simulation is used to handle the multiple significance testing problem.

Conclusion

Openshaw et al (1987) have demonstrated the construction and use of a methodology for analysing spatial point patterns and determining the significance of spatial clusters. The method is independent of spatial aggregation and is free of intuitively developed hypotheses. The method is still experimental but is very worthy of further study because, if proved, it could be of immense value in detecting the spatial occurrence of rare, but potentially avoidable diseases. It would be sensible if funds were made available to conduct research into the feasibility of the methodology in the Netherlands.

Miller, R.L. and Kahn, J.S. (1962). *Statistical analysis in the geological sciences.* Wiley, New York.

Openshaw, S., Charlton, M., Wymer, C. and Craft, A. (1987). A mark 1 Geographical Analysis Machine for the automated analysis of point data sets. *Int. J. Geographical Information Systems* 1: (4), 335-358.

Openshaw, S. (1988). Building more refined geographical analysis machines for the automated analysis of cancer data; a review of progress, problems and opportunities. *Proc. of a Congress on Quantitative Applications in Medical Geography*, University of Lancaster, September 1988.

Address for correspondence: Netherlands Expertise Center for Spatial Information Processing, Rijksuniversiteit Utrecht, Postbus 80.115, 3508 TC Utrecht, NL

Statistical methods for monitoring rare health events

G. Gallus*

* Instituto di Biometria E Statistica Medica, Facoltà di Medicina e Chirurgia, Università degli Studi, Milan, Italy

Health surveillance as a whole represents a fundamental prerequisite not only to be able to detect new environmental risk factors, but also to evaluate the efficacy of health interventions. It will become a topic of growing concern in developed countries.
Given the great variety of health conditions and their mutual relationships, the problem in itself appears to be rather complex. So far surveillance of health events has been considered for very specific problems, but in the future it is expected to become a highly structured system.
After the thalidomide epidemic, congenital malformations represent the first field in which health surveillance procedures have been set up, and in several countries they have been operating for many years.
The main emphasis of this intervention will be on the discussion of the problem from the statistical viewpoint.
The reference procedures remain those proposed for quality control in industrial environments, such as for example the CUSUM scheme, but some procedures, like the SETS method, have been specifically proposed for health applications.
The most relevant methods will be briefly presented and properly compared.
Their suitability to fit into the health requirements will be finally discussed.

Address for correspondence: Instituto di Biometria E Statistica Medica, Facoltà di Medicina e Chirurgia, Università degli Studi, Via Venezian 1, 20133 Milan, Italy

Common statistical problems in AIDS research and why a co-ordinated effort is therefore required

D.A. Lievesley*

* International Statistical Institute, Voorburg, NL

Many of the difficult questions about AIDS are essentially statistical. The purpose of statistical research on AIDS is to investigate what factors are important, what changes are needed in behaviour and attitudes, and what policies should be adopted, in order to reduce the spread of the infection. Statistical research on AIDS will thus fall into a number of different categories:

1 the acquisition of reliable information on sexual behaviour and attitudes and on knowledge of AIDS amongst different subgroups of the population; the assessment of the impact of campaigns to heighten awareness of AIDS and/or change behaviour.

2 the estimation of the current incidence of AIDS cases and of the number of people infected but not yet diseased (HIV seropositives) and the spatial dimensions of these cases.

3 the identification of parameters which are important in modelling the future spread and incidence of the disease and the determination of relevant models, in order to make predictions.

4 the measurement of the accuracy of the tests used to identify AIDS cases and those with HIV infection.

5 assessing the burden which AIDS cases will put upon the medical, educational and social services of communities, as well as the impact upon national economies (particularly due to the loss of people of productive age); and contributing to actuarial research on AIDS.

It will be discussed what rôle the International Statistical Institute can play through its research and education programmes, its committee structure and its networks to contribute to world wide efforts to estimate, forecast and combat the incidence of AIDS. A number of different ways in which the ISI might contribute including:

- bringing about international co-operation in the collection and interpretation of data on AIDS and in the modelling work;

- making recommendations on the type of studies and data which are required and perhaps drawing up international standards of guidelines;
- providing a central advisory service on statistical research on AIDS - particularly for the developing world;
- collecting together a data bank of relevant studies and data;
- acting as an independent non-political pressure group to ensure that resources are devoted to statistical research.

Address for correspondence: International Statistical Institute, Postbus 950, 2270 AZ Voorburg, NL