Abstracts

Bayesian Logistic Regression for Modeling Probabilities in a Two-way Layout Using Income Data

Ibrahim Abdalla (United Arab Emirates Univ, United Arab Emirates)

The objective of this paper is to obtain improved estimates of probabilities in a two-way cross-classification of nominal variables. Analysis is based on data made available from Abu-Dhabi Emirate Family Expenditure Survey, 1997, to estimate probability of low income, given education level, employment sector and other attributes, by employing a Bayesian approach using logistic regression. Model parameters are estimated using Markov Chain Monte Carlo (MCMC) simulation implemented using BUGS software. Each possible combination (i, j) is assumed to be a realization of a binomial random variable with a given sample size and unknown parameter. The sample proportion, the maximum likelihood estimate, is the classical estimator for this parameter. However, the estimate may possess undesirable features, particularly, when the data are sparse. An alternative is to fit a fixed effect model to the data, seeking relationships between the probability at each combination and certain levels of the nominal variables involved, or values of other attributes. This would produce smaller standard errors, but does not take into account the uncertainty associated with model parameters. Bayesian estimation provides better estimates that compromise between the two approaches.

Estimation of Lifetime Distribution from Incomplete Data and Follow ups
Abutaleb Ahmed (Jordan University of Science & Technology, Irbid, Jordan)

The motivation for the problem considered in this paper comes from a survival analysis problem in quality-assurance systems that are becoming more popular. For example, when an automobile failures occur within the automotive warranty period, a manufacturer can develop a record of mileage to failure from owners' request for repair. When no failures occur during the warranty period the owner naturally will not report the mileages, and it may be inferred that no record of failures. By using a follow-up survey data can be acquired to include a partial record of nonfailures . A method of estimating life time parameters is proposed for analyzing this kind of data under various scenarios. Assuming an exponential and Weibull lifetime distribution under the censoring model.

Investigating Item Orderings in Test Data

L. Andries van der Ark and Wicher P. Bergsma
(Tilburg University, The Netherlands)

Tests and questionnaires consisting of a list of polytomous items with ordinal answer categories are frequently used in the social and behavioral sciences to measure properties of respondents that are of interest to the researcher. It may be desirable that the order of the difficulty of the items (measuring an ability) or the attractiveness of the items (measuring a trait or attitude) is the same for every individual in the population of interest. Sijtsma and Junker (1996; also see Sijtsma and Hemker, 1998) called this property invariant item ordering (IIO) and they review situations in testing where IIO is required, such as research to differential item functioning, person fit analysis and adaptive testing. The aim of this paper is to define IIO in a statistical sense and to construct a statistic to test the hypothesis that IIO holds for a set of items. If IIO is rejected the statistic can be used to guide the test constructor which items to delete from the item pool in order to obtain an IIO.

There are many ways to define the ordering of the items. We consider likelihood ratio ordering (LRO), stochastic ordering (SO), and expected score ordering (ESO). The test constructor should give the justification of the type of ordering. Lehmann and Rojo (1992) showed that LRO implies SO, and SO implies ESO. This means that if an item ordering exists then this ordering is expressed by the ordering of the expected item scores.

For our population of interest we define θ as a multivariate latent trait containing all factors that influence a person's response on the set of items. Then, IIO means that the item ordering is invariant for θ. It is often desirable that θ is unidimensional, which means that the test measures only one latent trait.

Because θ is a latent variable it cannot be tested directly. We show that if an observed variable Y and the item scores are conditionally (on θ) independent then an SO invariant for θ implies an SO invariant for Y, and an ESO invariant for θ implies an ESO invariant for Y. However, an LRO invariant for θ does not imply an LRO invariant for Y.

If we have an appropriate choice for Y, we may obtain a table like Table 1. Table 1 gives the expected scores of six items with three ordered answer categories, $E(X_j)$, and the expected item scores given $Y = y(y = 1, 2, 3)$. The values of $E(X_j)$ indicate that the item order is $(1, \ldots, 6)$ but the bold faced values indicate that the item ordering is not invariant for Y. This suggests that IIO does not hold. To test whether or not the violations in the table are due to random fluctuations, we estimate a model where the violations of the item ordering are constrained to be equal.

Table 1

Item (j)	$E(X_j)$	$E(X_j\|Y=1)$	$E(X_j\|Y=2)$	$E(X_j\|Y=3)$
1	0.2	**0.2**	0.2	**0.2**
2	0.4	**0.1**	**1.0**	**0.1**
3	0.8	0.8	**0.6**	1.0
4	1.0	1.4	**0.4**	1.2
5	1.6	**1.8**	1.4	1.6
6	1.8	**1.6**	1.9	1.9

This model is called the *IIO-model*. For Table 1 this means estimating a model where $E(X_1|Y=1) = E(X_2|Y=1)$, $E(X_2|Y=2) = E(X_3|Y=2) = E(X_4|Y=2)$, and $E(X_1|Y=3) = E(X_2|Y=3)$. For several reasons we prefer *minimum discrimination information* estimates (Ireland and Kullback, 1968) which minimize the discrimination information statistic $I = \sum m_i log(m_i/n_i)$, where m_i are the estimated frequencies and n_i are the observed frequencies. This model is called the *IIO-model*. The constraints in the IIO-model are due to violations of inequality constraints. As a result I does not have an asymptotic chi squared distribution. To test whether the IIO-model should be rejected, we use the bootstrap to determine the distribution of I. The quality of the bootstrap distribution has to be further investigated, but the first results are promising.

If the IIO-model is rejected, a backward selection algorithm using the minimum discrimination information statistic I can be applied to delete the items that cause the violations of IIO. The general idea is that if we have J items, we compute $I_{(-j)}$, the minimum information statistic of the set of items with the j-th item deleted, for $j = 1, \ldots J$. Item j^* corresponding to $I_{(-j*)} \# I_{(-j)}$ is deleted. The cycle is repeated until the test constructor is willing to accept the IIO-model.

References

1. Ireland, C.T. and Kullback, S. (1968): Minimum Discrimination Information Estimates. *Biometrics*, **33**, 283-300.
2. Lehmann, E.L. and Rojo, R. (1992): Invariant Directional Orderings. *Annals of Statistics*, **20**, 2100-2110.
3. Sijtsma, K. and Hemker, B. T. (1998): Nonparametric Polytomous IRT Models for Invariant Item Ordering, with Results for Parametric Models. *Psychometrika*, **63**, 183-200.
4. Sijtsma, K. and Junker, B. W. (1996): A Survey of Theory and Methods of Invariant Item Ordering. *British Journal of Mathematical and Statistical Psychology*, **49**, 79-105.

Blockmodeling 2-Mode Networks

Vladimir Batagelj (University of Ljubljana, Slovenia)
Patrick Doreian (University of Pittsburgh, USA)
Anuška Ferligoj (University of Ljubljana, Slovenia)

The blockmodeling problem is a special kind of clustering problem searching for a clustering (partition) of units (actors) of network that reveals the internal structure of the network.
In 1991 we proposed optimizational approach to blockmodeling problem of conventional 1-mode networks where relations are defined over one set of units. The approach was based on criterion functions compatible with structural and regular equivalence [4,3]. In 1993 we extended this idea to the other types of blocks - the generalized blockmodeling [2]. In 1996 we added to the approach the fitting of the given network to a pre-specified blockmodel [1,5]. We also studied some special blockmodels [6].
In the paper we present an extension of our generalized blockmodeling approach to the blockmodeling of 2-mode networks [7]. Such a network contains measurements on which units from one set have ties to units in the other set. The generalized blockmodeling approach of 2-mode networks considers two partitions (for each set of units), different types of blocks, and the possibility of fitting to prespecified blockmodels. The criterion function is addapted to this blockmodeling problem and optimized by the relocation method.
The approach will be illustrated with several examples.

References

1. Batagelj, V. (1996): 'MODEL 2 – Program for Generalized Pre-Specified Blockmodeling', manual, Deparment of Mathematics, University of Ljubljana.
2. Batagelj, V. (1997): Notes on Blockmodeling. *Social Networks*, **19**, 143–155.
3. Batagelj, V., Doreian, P., and Ferligoj, A. (1992): An optimizational approach to regular equivalence. *Social Networks*, **14**, 121–135.
4. Batagelj, V., Ferligoj, A., and Doreian, P. (1992): Direct and indirect methods for structural equivalence. *Social Networks*, **14**, 63–90.
5. Batagelj, V., Ferligoj, A., and Doreian, P. (1998): Fitting pre-specified blockmodels. In *Data Science, Classification, and Related Methods*, Eds., C. Hayashi, N. Ohsumi, K. Yajima, Y. Tanaka, H. H. Bock, and Y. Baba, Tokyo: Springer-Verlag, 199-206.
6. Doreian P., Batagelj V., and Ferligoj A. (2000): Symmetric-acyclic decompositions of networks. *Journal of Classification*, **17**, 3-28.

7. Everett, M.G. and Borgatti, S.P. (1992): Regular colouring of digraphs, networks and hypergraphs. *Social Networks*, **14**, 1–35.

Network Analysis of Reuters News about the Terrorist Attack on September 11, 2001

Vladimir Batagelj and Andrej Mrvar (University of Ljubljana, Slovenia)

On September 11, 2001 the terrorist attacks in New York (Twin Towers) and Washington (Pentagon) happened. Reuters news were tracked for the next 66 days after the attack. For each day the news were transformed into a network using Centering Resonance Analysis (CRA). See: `http://locks.asu.edu/terror/`
CRA is a new text analysis technique developed by Steve Corman and Kevin Dooley at Arizona State University. It uses natural language processing and network text analysis techniques to produce abstract representations of texts.
We will present the network analysis of the CRA networks using program Pajek.
Pajek is a program (for Windows) for large network analysis and visualization. It is freely available, for noncommercial use, at its site:
`http://vlado.fmf.uni-lj.si/pub/networks/pajek/`

Customer Interaction and Interface Analysis and the Following Consequences – Case 'Parkschlössl'

Thomas Benesch, Klemens Franz, Kerstin Gnaser, Alexander Nischelwitzer, and Peter Unterkofler (FH Joanneum, Graz, Austria)

In the past years the booking rate of not only this hotel but most other small and medium sized hotels in the region especially the ones depending only on a single season (summer) has decreased steadily. The typical guests who were former frequenters of the tourist area have gotten old of age and the younger generation is lured to the Mediterranean via cheep flights and the guarantee of sunshine. The potential guest is still out there but the competition has gotten harder to convince him.

The motivation for this study was to answer the question of how can a hotel have a better customer focus to stand up to the increasing competition? And how can it improve its customer relations to receive an increase of bookings. Which tools are necessary? And how to other hotels, which are in a similar situation, interact with their customers?

The study was focused on the analysis of four major factor of customer interaction: traditional mail, e-mail, and the unidirectional approach of brochures and web pages. This was done with the help of a fake mail request, a fake e-mail request, a fax questionnaire, a hotel webpage analysis combined with a search engine analysis. The mail and email analysis were focused on response time, quantity and quality of information given. The webpage analysis had the emphasis on the HTML code browser compatibility and load time. The questionnaire was sent out to each hotel in the study to acquire information on the use of Internet, online booking systems, their expectations and feedback they receive from their guests.

The results show that most hotels make good use of traditional mail and that there are only few possible improvements in the answer. The potentialities of email on the other hand are unexplored yet by most hotels. The homepage and search engine analysis showed that only half of the hotel use meta tag information which is a factor, that helps improve the position in search rankings. In addition the search engine analysis showed that, when leaving out the hotel name in searches the possibility of getting a result is significantly lower than with the name. A good description and a lot of keywords could help to put Parkschlössl in a better position among the search rankings.

In conclusion the study suggests a design alignment between used media paving the way for a corporate design and a better media integration. In addition an increase of information provided especially on the webpage with focus not only on general information about the hotel but details on the rooms, meals and so on. Information on activities one can engage in around the area, cultural and historical information with links to other sites interesting to the visitor. All of this can be combined

with multimedia elements so the user can visualize better how it would be coming for a holiday. Interactivity would be another important key factor to keep users interested in the website. An interesting game focused on the hotel or the village of Millstatt would also be a good idea to increase traffic on the website.

Internet Surveys: How Can they Contribute to Improve Traditional Survey Data and Administrative Databases Information

Silvia Biffignandi and Monica Pratesi (University of Bergamo, Italy)

During the 90s years a great amount of research has focused on the use of administrative data for statistical purposes and to the integration of various administrative databases. Recently, integration between administrative data and survey data is both under theoretical investigation and under empirical implementation (see Biffignandi 1994 and 1995). The integration between survey and administrative databases let to assign to the surveys new roles and let arise new problems regarding survey methodology.

This paper focuses on the possible new role and content of the surveys and on the new methodological problems in survey research, with special attention to Internet surveys. The presentation is carried out by taking into account the state of art in Italy and by discussing the situation of some Italian databases (such as Ministero delle Finanze, balansheet database, structural statistics survey) and of Internet survey which are planned in the next future. The authors on the basis of their research experiences give a profile of the contribution of Internet survey both from the economic information point of view and from the methodological point of view.

References

1. Biffignandi, S. (1994): Integrazione di archivi e sistema statistico delle imprese. In *Atti del Convegno della Societa Italiana di Statistica*, San Remo, aprile 1994.
2. Biffignandi, S. and Martini, M. (a cura di) (1995): Il registro statistico europeo delle imprese in Europa. Esperienze e metodi per la sua costruzione in Italia. In *Il registro statistico europeo delle imprese in Europa. Esperienze e metodi per la sua costruzione in Italia*, F. Angeli, Milano.
3. Biffignandi, S. (2000): L'analisi statistico economica dei gruppi d'impresa: problemi e risultati. In *Tecnologie informatiche e fonti amministrative nella produzione di dati, a cura di C. Filippucci*, Franco Angeli, Milano.
4. Biffignandi S. and Pratesi M. (2001): Internet surveys: timeliness of data collection and individual survey period length. Proceedings of the International Conference Methodology and Statistics, Ljubljana 17-19 September 2001.
5. Biffignandi S. and Pratesi M. (2002): The respondents' profile in a Web survey on firms in Italy. *Metodološki zvezki*, **18**, Ljubljana, Slovenia (forthcoming).

Combining Segmentation and Region Clustering for Image Categorization
Janez Brank (J. Stefan Institute, Ljubljana, Slovenia)

Introduction. Collections of images, also known as pictorial databases, are becoming more and more common in many situations. One of the ways to handle massive collections of images better is to organize them into categories, with each image being assigned to one or more categories. The goal of image categorization (or image classification) is to partly automate this task. Thus we assume that the set of categories is known and fixed in advance, and that a set of images with supposedly correct category memberships is available. Our goal is then to train a model for each category, which will enable us to determine, for any particular image, whether it should belong to this category or not.

This problem can be seen as lying at the intersection of image retrieval and machine learning, and it can benefit from techniques from both fields. The chief underlying issue is that of representation: how should individual images be represented so as to allow learning algorithms to successfully (and efficiently) learn models with good predictive abilities? Since machine learning algorithms typically cannot work directly with pictorial data, images will need to be represented with structures that learning algorithms can handle. In this aspect we can build on the considerable amount of work on representation that has already been done in the field of image retrieval. Although in practice images are often represented by textual descriptions or keywords (when such descriptions are easily available), preparing such descriptions for a large database is often impractical. In such cases, content-based descriptions (i.e. based on the appearance of the images themselves) must be used.

Several well-known kinds of content-based image representations are based on relatively straight-forward ways of describing an image with a vector. This is a desirable approach as numerous machine learning algorithms can work with vectors. Such descriptions include e.g. histograms, which are obtained by simplifying the image into a relatively small number of colors and determining the proportion of the image covered by each color. However, this representation entirely disregards the distribution of pixels of each color around the image (e.g. scattered vs. compact). Several other representations have been proposed to address this. For example, autocorrelograms (Huang, Kumar, and Mitra, 1997) store probabilities that a pixel, chosen randomly at distance d from a randomly chosen pixel with color c, will itself be of color c; this should be computed for all c and for several small values of d. The vector of the resulting values gives some indication both of the amount of a color present in the image, and of its distribution in space.

However, in addition to these 'global' vector-based representations, the image retrieval has also proposed various more complex image representations, many of them based on segmentation. Segmentation is the process of partitioning an image into several (perhaps disjoint) areas, known as regions, such that each region

is approximately homogeneous in appearance (i.e. in color or in texture). While segmentation is still a subject of research, the existing approaches to segmentation have already been used with good results for image retrieval purposes. Informally, it seems convincing to say that segmentation brings out some information about the image that the simpler, global representations would probably miss; it is therefore interesting to try using representations based on segmentation as a basis for image categorization.

However, segmentation-based descriptions of images have a slightly more complex structure than typical machine learning algorithms expect. Each image consists of a set of regions, and each region has an associated description (usually a short vector arising from the segmentation procedure). Several measures of similarity between segmented images have been used in image retrieval, but few machine learning algorithms can work with an arbitrary similarity measure (the nearest-neighbor method can, but its classification accuracy has been found to be unpromising in our earlier work (Brank, 2001). In this paper we propose two approaches to using segmentation-based image descriptions in combination with the support vector machine (SVM) algorithm (Cortes and Vapnik, 1995), which has achieved a wide popularity in the machine learning literature in the past years and has been found to work well in many domains.

Generalized kernels. The original form of the support vector machine assumes that instances to be categorized are d-dimensional real vectors. Each instance is either positive or negative, and SVM attempts to separate them with a hyperplane such that positive instances lie on one side of the plane, negative instances on the other side, and margin of the plane (i.e. the distance from the plane to the nearest instances) is as wide as possible. It turns out, however, that it is not necessary to work with the vectors directly, as long as we can compute dot products between them. Thus one can obtain nonlinear models by mapping the training vectors nonlinearly into a new vector space, as long as a function (called a kernel) can be found that efficiently computes dot products of the images of vectors under such a mapping.

A kernel can in some sense be seen as a similarity measure: the dot product of two vectors is larger if they point into a similar direction. However, not any arbitrary similarity measure can be used as a kernel because it might not correspond to a dot product in some vector space. We can, however, use a slightly different formulation of the optimization problem on which SVM training is based. This has been introduced in (Mangasarian, 2000) as generalized kernels. In its simplest form, it can be seen to be equivalent to representing an instance I by the vector $(s(I, I_j))_j$ of its similarity values to some set of instances $\{I_j\}_j$, typically the training set. Thus two instances are considered similar if they exhibit a similar pattern of similarity values to a fixed set of training instances. To use this approach for image categorization, we have employed the WALRUS segmentation method (Natsev, Rastogi,

and Shim, 1999) and the IRM (integrated region matching) similarity measure (Li, Wang, and Wiederhold, 2000).

Region clustering. Image segmentation algorithms typically also produce a short vector describing each region. For example, such a vector may contain the average color of the region and other related information. However, regions from different images are quite unrelated, their vectors might be quite different, and consequently it is difficult to compare images. To obtain a common representation space for all images, we propose to cluster the descriptions of all regions from all training images. Each resulting cluster corresponds to a group of similar regions, probably from several different images. To approximately describe an image, we can now simply note what percentage of that image is covered by regions from each particular cluster. Thus each image is now described by a single (probably sparse) vector, and these vectors can be used to train a SVM classifier on. The BIRCH clustering algorithm (Zhang, Ramakrishnan, and Livny, 1996) has been used in this approach, primarily because the same algorithm is used in WALRUS for the segmentation itself.

Experiments. We used a set of images from the misc database, which has already been used in image retrieval literature (Natsev, Rastogi, and Shim, 1999). This is a diverse collection of 9907 small photographic images (typically around 128×96 pixels large). We selected a subset of 1172 images and manually assigned each of them into one of the following 14 categories: butterflies, flags, sunsets, autumn, planets, flowers, automobiles, Earth from space, mountains, clouds, sea, surfboards, sailboats, prairie animals. The LibSvm program (Chang and Lin, 2001) was used to train SVM models, its main advantage being a built-in capability to handle multiclass problems (for each pair of categories, it trains a classifier to distinguish between these two categories; to classify a new instance, it is shown to all these classifiers and the category with the largest number of votes wins). As a baseline technique to compare our proposed approaches with, we chose the autocorrelogram representation, which has been found to work well in comparison to histograms (Huang, Kumar, and Zabih, 1998) and several other global vector-based representations (Brank, 2001). The results reported here are based on average classification accuracies (and their standard errors) after stratified tenfold cross-validation.

These experiments surprisingly show that the generalized kernel approach does not yield a significantly different classification accuracy in comparison to the autocorrelogram-based approach, while the generalized clustering method actually performs significantly worse. A closer examination suggest that the partition of regions into clusters is not sufficiently stable, i.e. if the cluster centroids were stored and regions redistributed to the nearest centroid, the partition would change considerably. This can easily cause two similar images to be treated as quite different. As one way to avoid the influence of this instability we considered including the

test images into the clustering process as well. This amounts to a form of transduction, and it is therefore also fair to introduce transduction into the SVM learning process as well. In the transductive setting, the region clustering approach achieves a slightly higher classification accuracy than the autocorrelogram-based representation.

Conclusion. Our experiments indicate that despite the usefulness of segmentation-based representations for image retrieval, it is more difficult to make good use of such representations for image categorization. This may possibly also be in part due to the machine learning methods used, and could be tackled by e.g. multiple-instance learning.

Further work that we hope to complete by the time of the conference is to test the performance of a combined representation including both features from the generalized kernels approach and from the region clustering approach.

1. Brank, J. (2001): Machine Learning on Images. *Information Society 2001*, Ljubljana, 152-55.
2. Chang, C.C., and Lin, C.J. (2001): *LibSVM: A Library for Support Vector Machines (version 2.3)*. Dept. of Comp. Sci. and Inf. Engineering, National Taiwan University, April 2001.
3. Cortes, C. and Vapnik, V.N. (1995): Support-Vector Networks. *Machine Learning*, **20**, 273-297, Sept. 1995.
4. Huang, J., Kumar, S.R., and Mitra, M (1997): Combining Supervised Learning with Color Correlograms for Content-based Image Retrieval. *Proc. 5th ACM Int. Multimedia Conf.*, 325-334.
5. Huang, J., Kumar, S.R., and Zabih, R. (1998): An Autmoatic Hierarchical Image Classification Scheme. *Proc. 6th ACM Multimedia Conf.*, 219-228.
6. Li, J., Wang, J.Z., and Wiederhold, G. (2000): IRM: Integrated Region Matching for Image Retrieval. *Proc. 8th ACM Multimedia Conf.*, 147-156.
7. Mangasarian, O.L. (2000): Generalized Support Vector Machines. In: A. J. Smola et al. (eds.), *Advances in large margin classifiers*, MIT Press, 2000, 135-146.
8. Natsev, A., Rastogi, R., and Shim, K. (1999): WALRUS: A Similarity Retrieval Algorithm for Image Databases. Proc. ACM SIGMOD, 395-406.
9. Zhang, T., Ramakrishnan, R., and Livny, M. (1996): BIRCH : An Efficient Data Clustering Method for Very Large Data-Bases. *Proc. ACM SIGMOD*, 103-114.

Web-Based Learning of Statistics: Does it Work?

Lea Bregar, Mojca Bavdaž Kveder, and Irena Ograjenšek
(University of Ljubljana, Slovenia)

One of the most significant characteristics of modern societies is the constant need to revise and update knowledge on one, and to upgrade skills on the other hand. The need to continuously revise and update knowledge also results in necessary revisions of old and development of new pedagogical concepts and delivery formats, adapted to the needs of continuous learning. Arising from these needs, traditional distance learning was successfully introduced several decades ago. With the recent fast development of information and telecommunication technology, this traditional form of distance learning went through the process of transformation and adaptation to emerge as a new concept of e-learning, depending heavily on the advantages of the Internet.

From the viewpoint of statistical education, such development is more than welcome. The potential advantages include: course interactivity based on the use of different communication channels and tools; possibility to create new knowledge by applying statistical tools to real-life data in order to tackle real-life problems; study flexibility in terms of time, place and speed; depth and variability of available course topics.

This paper attempts to investigate whether the current practice of design and delivery of web-based statistical courses provides a sufficient and appropriate framework for deployment of these benefits form learners perspective. In order to achieve that, general technology-based model is proposed first, linking concepts of constructivist theory and distance learning. This model is then used as a comparison standard in the exploratory study of the current state of affairs of statistics on the web.

Our survey shows that at present, very few comprehensive statistical courses can be found. Web-based courses have been far too often diminished to a simple on-line delivery of syllabi and traditional learning materials (textbooks, case studies, exam sheets, etc.) Their development and implementation are usually due to dedicated enthusiasts and not a result of systematic efforts at the institutional level. Being a novelty, their proponents usually have a hard time trying to ensure their recognition and application in the framework of traditional universities. A search for detailed information is further made difficult by vague terminology used in the field, as well as by subordination of statistics to mathematics and related disciplines.

All in all, given the present situation in the hyperspace, it is hard for learners to conduct a meaningful search for an appropriate statistical course. It is even harder for them to find a course tailored to their needs and expectations. Lack of terminological clarity is certainly an important obstacle. The main problem, however, lies somewhere else: in the lack of courses with sufficient learners support (be it

through communication or through course design).
Presently, statistics on the web can be of use predominantly to teachers, and perhaps advanced learners, who are able to master a search in a user-unfriendly environment dealing with a very demanding topic. It is therefore our suggestion, that a well-organised web directory, which could be used without problems even by less skilled learners, should be developed for web-based courses on statistics. Among a large number of benefits thus achieved, two should be specially emphasised. Firstly, for learners a search and selection of a proper course would be much easier. Secondly, researchers and practitioners in the field could exchange information much more quickly and efficiently, which would further stimulate the development of the field.

Compositional Time Series

Matevž Bren (University of Maribor, Slovenia)
Vladimir Batagelj (University of Ljubljana, Slovenia)

Compositional data are special kind of data representing parts of some whole. For example: students enrolling faculties for the first time, bank clients subscribing monthly for different kinds of loans, year productions of different car factories, different energy sources consumption or production... They obey the (unit) sum constraint, that have to be considered analyzing such data with statistical methods. In our paper we'll present a compositional approach to time series analysis – in many cases in time series data we are interested just in dynamics of proportions, and not the amounts.
We'll illustrate our approach on real data.

References

1. Aitchison, J. (1986): *The Statistical Analysis of Compositional Data*, Chapman and Hall, New York, 416.
2. Brunsdon, T.M. and Smith, T.M.F. (1998): The Time Series Analysis of Compositional Data, *Journal of Official Statistics*, **14**, No. 3, 237-253.
3. Grunwald, G.K., Raftery, A.E., and Guttorp, P. (1993): Time series of continuous proportions, *Journal of the Royal Statistical Society*, Series B (Statistical Methodology), **55**, No.1, 103-116.

The Relationship between the Relevance Quotient and the Indices of Risk in a 2X2 Exposure-Affection Table

Maurizio Brizzi (University of Bologna, Italy)

The relevance quotient: definition and main properties

The relevance quotient (RQ) is a symmetric measure of association between events, introduced by Costantini (1979) and developed by other Authors, such as Belcastro and Guala (1998) and Brizzi (1998). The RQ is defined as the ratio between the joint probability of a set of events (or the joint density, if the sample space is continuous) and the corresponding probability (or density) under stochastic independence. In particular, if A and B are two events with positive probability $P(A)$ and $P(B)$ respectively, the RQ between A and B is:

$$Q_{A,B} = \frac{P(A \cap B)}{P(A)P(B)} = \frac{P(A|B)}{P(A)} = \frac{P(B|A)}{P(B)} \tag{1}$$

When A and B are disjoint, they have an empty intersection, and their RQ is obviously equal to zero; when they are independent, the numerator and the denominator of (1) become equal, and then $Q_{A,B} = 1$. When the RQ overtakes this value, there is a positive association between these events. The value of $Q_{A,B}$ may be estimated using the relative frequencies instead of probabilities; the estimated RQ is then:

$$\hat{Q}_{A,B} = \frac{f(A \cap B)}{f(A) \cdot f(B)} \tag{2}$$

Decomposition of the indices of risk based on relevance quotients

One of the simplest way of studying the association between a potentially dangerous factor and the incidence of a pathology is the well known exposure-affection contingency table, from which we can derive many indices of risk, such as the rate ratio (RR) and the Odds Ratio (OR), developed by many Authors, such as Agresti (1990):

Table 1: Exposure-affection contingency table

	Affected	Not Aff.	Total
Exposed	n_{11}	n_{12}	n_{10}
Not Exp.	n_{21}	n_{22}	n_{20}
Total	n_{01}	n_{02}	n

$$RR = \frac{n_{11}/n_{10}}{n_{21}/n_{20}} = \frac{n_{11} \cdot n_{20}}{n_{21} \cdot n_{10}}; \quad OR = \frac{n_{11} \cdot n_{22}}{n_{21} \cdot n_{12}}. \tag{3}$$

Now, we consider the four relevance quotients corresponding to the exposure-affection events, and estimate them by applying (2) to Table 1:

$$\hat{Q}_{ij} = \frac{n_{ij} \cdot n}{n_{io} n_{oj}} = \frac{n_{ij}}{\tilde{n}_{ij}},\ i = 1, 2; j = 1, 2 \tag{4}$$

where $\tilde{n}_{ij}$ is the theoretical value of n_{ij} under stochastic independence.
It is easy to demonstrate that both RR and OR may be written as a function of these quotients, and more precisely:

$$RR = \frac{\hat{Q}_{11}}{\hat{Q}_{21}}; OR = \frac{\hat{Q}_{11}\hat{Q}_{22}}{\hat{Q}_{21}\hat{Q}_{12}} \tag{5}$$

If we use the notation: $A_{ij} = \left\{ \begin{array}{c} \hat{Q}_{ij}\ ,\ \text{if } i = j \\ 1/\hat{Q}_{ij}\ ,\ \text{if } i \neq j \end{array} \right\}$, we can write:

$$RR = A_{11} \cdot A_{21}; \quad OR = A_{11} \cdot A_{12} \cdot A_{21} \cdot A_{22}. \tag{6}$$

If we take the natural logarithm of each member in (6), we have an additive decomposition:

$$LRR = \ln RR = \ln A_{11} + \ln A_{21}; \tag{7}$$

$$LOR = \ln A_{11} + \ln A_{12} + \ln A_{21} + \ln A_{22}. \tag{8}$$

We can then evaluate the contribution of each cell to the value taken by the RR and the OR, respectively.
It can be easily noticed that most important contribution derives from the cell n_{11}; if we consider only the RQ between the events 'exposed' and 'affected' we get a very simple index of risk:

$$QR = Q_{11} = \frac{n_{11}}{\tilde{n}_{11}} \tag{9}$$

Studying the sample distribution of the indices of risk
We have then tried to study and compare the distribution of the indices QR, RR, OR, and their logarithmic transformations, following two different approaches: an 'exact' approach, developed by considering all the possible 2 x 2 contingency tables we can build under certain constraints, and a simulation approach with the

aid of the package 'GAUSS'. In particular, the indices QR and RR are perfectly cograduated, and the logarithmic transformations are asymptotically normal with different speed of convergence.

References

1. Agresti, A. (1990): *Categorical Data Analysis*, John Wiley and Sons, New York.
2. Belcastro, A. and Guala, E. (1998): Le λ-inferenze induttive: caratterizzazione e proprietà. 'Statistica', **LVIII**, 4, 603-628 (in Italian, with abstract in English).
3. Brizzi, M. (1998): Il ruolo del quoziente di rilevanza nelle inferenze λ-predittive. 'Statistica', **LVIII**, 4, 629-639 (in Italian, with abstract in English).
4. Costantini, D. (1979): *The relevance quotient*, 'Erkenntnis', **14**, 149-157.

Issues for Concern When Using ML-based Methods to Handle Missing Data?
Peter Mulhall, Brendan Bunting, and Gary Adamson
(University of Ulster, N. Ireland)

The performance of the Expectation Maximization (EM) algorithm and Full Information Maximum Likelihood (FIML) were tested within a simulation study which mimicked conditions that could be expected in a general field trial, with approximately 25% of data missing at random. A factorial design tested the performance of the two methods with three levels of parameter estimate magnitude (i.e. model quality) and with three levels of sample size (100, 400, 800). As expected, the results show that both EM and FIML recaptured the original parameter estimates and standard errors with a high degree of accuracy, but the EM procedure poorly recaptured the original fit statistics. However, performance for both methods was enhanced with increased model quality and sample size. Adjustments to the Minimum Fit Function are proposed which significantly improve the EM- based fit statistics.

Introduction to Facet Analysis

David Cairns (Macquarie University, Sydney, Australia)

The purpose of this presentation is to introduce facet analysis to an audience who may not be familiar with this approach. Facet analysis is the analysis approach of Facet Theory, which also includes facet design.

Facet Theory was developed by Louis Guttman and is best described as an integrated approach or systematic framework for research particularly suited to, but not limited to, social and behavioural sciences. Its main features are that it formally links the design and analysis of scientific questions and promotes cumulative research and the development of scientific laws. Facet Theory includes both facet design and facet analysis. This presentation is mainly concerned with facet analysis as an approach to data analysis. This approach includes representing multivariate data as a visual map (using similarity structure analysis, multiple scalogram analysis and partial-order scalogram analysis) and interpreting these maps in terms of partitions and regions. It is this concept of interpretation through partitioning which is unique to facet analysis. By way of introduction, this presentation uses an example of data that is well-known and well-understood by data analysts, namely, data from a psychological experiment. It is hoped by using this example that facet analysis can be seen in familiar terms.

Most commonly experimental data in psychology consists of manipulated categorical independent variables with measured numerical dependent variables. Analysis routinely involves plots of means followed by various ANOVA techniques producing an array of interpreted p-values to test hypotheses. This is a well-known and accepted approach to the analysis of data. The hypotheses will also be tested using the less well-known facet analysis approach which confirms hypotheses through the partitioning of spatial maps of the data.

The example consists of typical experimental data from a 2x3x2 experiment with repeated measures on the last two factors. The experiment investigated the effects of speed of presentation, difficulty level and mode of presentation of a number of passages on self-rated comprehension levels of 30 subjects. Mode of presentation (reading, listening) was a between group factor with speed of presentation (slow, medium, fast) and difficulty (easy, hard) were both within subject factors. Hence the design is a mixed design. It will be shown that both the ANOVA approach and the facet analysis approach can evaluate the hypotheses involving these factors. It will be shown that although the facet analysis does not provide p-values it does provide further graphical insight into the data.

Using the traditional statistical model three main effects, three two-way interactions and one three-way interaction were evaluated. Using this statistical model the data are split into systematic components and random errors.

The traditional approach resulted in a non-significant mode of presentation effect

and significant speed and difficulty main effects. Of the two-way interactions mode by speed was significant as was speed by difficulty but the mode by difficulty interaction failed to reach significance. The three-way interaction was not significant, although the interaction graph did suggest that something was going on. Interpreting the main effects revealed that comprehension was greater for easy passages and passages read at slow or medium speeds. The first two-way interaction revealed that comprehension was always higher for the easy passages compared with the hard passages particularly when the speed was fast. The second two-way interaction revealed that comprehension was higher when the passages were being listened to except when the speed was fast when the comprehension was greater for the reading mode of presentation. There is some level of difficulty to incorporate these separate findings into a global statement for the results of the experiment.

Facet analysis provides an alternative way of testing these hypotheses that does not depend on p-values. Facet analysis uses a statistical visualization model whereby data are represented either by a row space map, a column space map or a joint space map. A number of statistical techniques including smallest space analysis (also known as multidimensional scaling), multiple scalogram analysis (also known as multiple correspondence analysis) are employed by facet analysis. However, other statistical techniques such as principal components analysis and preference analysis are applicable to use with the facet analysis approach. The facet analysis confirmed the traditional results but also gave a visual insight into what was going on in the three-way interaction which was not obvious in the traditional analysis.

The facet analysis of the experimental data visually confirmed that easy passages were separated from hard passages and that slow and medium passages were separated from fast passages. These hypotheses were confirmed through the clear partitioning of the space diagrams into regions separating the types of passages in terms of speed and difficulty. Perhaps the main advantage of the facet approach was in looking at the three-way map showing the joint information of mode of presentation, speed and difficulty. Here, although the traditional analysis was non-significant, the graphical analysis revealed that structure was present and revealed why the interaction was not significant. The graph revealed three types of patterns; a pattern for readers whose comprehension scores were dependent on the easy vs hard distinction; a pattern for listeners whose comprehension scores were dependent on the slow, medium vs fast distinction; and a third pattern which combined both readers and listeners whose comprehension scores were dependent on a combination of the speed and difficulty factors.

The facet analysis was presented as giving a visual insight into the data not available by the traditional analysis approach. Certainly the visual approach complemented the traditional approach. In conclusion, the facet approach will be discussed in terms of the many other forms of data which it is applicable to including survey data and qualitative data. Programs which can be used for the facet analysis

approach will be briefly discussed.

Predicting Random Level and Seasonality of Hotel Prices. A Structural Equation Growth Curve Approach

Germà Coenders, Josep Maria Espinet, and Marc Saez (Univ. of Girona, Spain)

This article examines the effect on price of different characteristics of holiday hotels in the sun-and-beach segment, under the hedonic function perspective. Monthly prices of the majority of hotels in the Spanish continental Mediterranean coast are gathered from May to October 1999 from the tour operator catalogues. Hedonic functions are specified as random-effect models and parametrized as structural equation models with two latent variables, a random peak season price and a random width of seasonal fluctuations. Characteristics of the hotel and the region where they are located are used as predictors of both latent variables.
Besides hotel category, region, distance to the beach, availability of parking place and room equipment have an effect on both peak price and seasonality. 3-star hotels have the highest seasonality and hotels located in the southern regions the lowest. This could be explained by a warmer climate in autumn in these regions, and the model is expanded to include the effect of climate on prices.

Multivariate Modelling with Latent Variables in Experimental Design Application to Spruce (Picea abies K.)

Dario Cziráky, Anamarija Pisarović, and Tugomir Filipan
(Department for Environmental Sciences, IMO, Zagreb, Croatia)

An important problem in large-scale forestry experiments concerns the choice of tree development measures (indicators) and methods for testing the effects of experimental treatments such as soil correctors and fertilisers on tree growth. In the standard experimental design for testing of treatment effects on tree development, ANOVA and dummy-variable regression techniques are routinely applied to individual tree development indicators such as total tree height or trunk diameter. Multivariate techniques that can combine several tree development indicators and test treatment effects on all indicators simultaneously are, however, rarely used in the forestry research.

On the other hand, structural and functional relationships among different tree development indicators such as diameter and height are extensively researched in the literature (e.g., Arney, 1985, Wensel et al., 1987). Other tree dimensions such as crown and stem grown can also be structurally modelled (e.g., Kellomki et al., 1999). For recent literature on general forest growth modelling see inter alia Botkin (1993), Bossel and Kriger (1994), Vanclay (1994) and Adler (1995). Specific applications of functional modelling of the relationships among tree development indicators often focus on modelling tree diameter as a function other tree measures (Gourlet-Fleury and Houllier, 2000) or on prediction of particular indicators such as volume increment (Walters and Hann, 1986). Models of ring width and volume distribution along the bole (Mitchell, 1975; Houllier et al., 1995), or vertical distribution of annual ring area (Curbert, 1999) all follow similar lines of (mainly deterministic) analysis, while Fox et al. (2001) develop a growth model that explicitly incorporates stochastic structure of the growth indicators.

This line of research shows clear relationship between different tree development indicators suggesting that they all measure the same underlying (latent) variable, which is tree development. The problem of how to test for treatment effects within classical experimental design immediately follows. Namely, we can either test for the treatment effects on separate tree development indicators (e.g., height or diameter), or make an attempt to combine multiple indicators into a single (latent) variable and then test for treatment effects on a composite (latent) tree development variable.

Testing for treatment effects on different indicators, separately, is most common in practice but problematic in several ways. First, if multiple indicators are used to separately measure the effect of applied treatments on the overall tree development, there is a good chance that the results will be ambiguous. That is, it might be the case that different treatments show strongest effects on different measures,

for example, one treatment might appear superior vis-a-vis other treatments insofar the effect on tree height goes, while another treatment might be superior in respect to tree diameter or yet another measure of tree development. Across time, when growth dynamics are taken into account, this ambiguity can be even larger. A known example is the application of soil correctors with high amino acid content which are likely to have stronger effect on tree height increment in the earlier growth phases then correctors with lower amino acid content. Such treatments, however, tend to show less discriminating effect on other measures (e.g., trunk diameter).

The second approach based on multivariate modelling, while substantively appealing, is often difficult to apply in practice partly due to the lack of appropriately taken tree measures for such purpose or due to inadequately applied statistical techniques. The methods most commonly used in multivariate modelling with latent variables are based on linear structural equation modelling which belong to the general class of covariance structure analysis. Rather powerful methods are available for modelling multivariate Gaussian variables but these methods generally fail to account for complex nonlinearities in the relationships among particular tree development indicators that are themselves not normally distributed.

This paper analyses data containing three indicators of tree developments measured in two time points (end of years 1999 and 2001) from an ongoing multi-annual forestry experiment carried out since 1992 at a site in Fuegenberg, Zillertal (Tyrol, Austria). The dependence of multivariate techniques on distributional assumptions is accounted for by detailed preliminary descriptive analysis and data transformations aiming at assuring Gaussian distribution of the modelled variables. A panel (longitudinal) latent variable model is then developed and estimated using maximum likelihood method for the joint 1999 and 2001 sample. The panel model was estimated as a general LISREL model using full information maximum likelihood technique, which allowed statistical evaluation of models specification and provided empirical guidance to further model modification. We further show how to compute latent scores for the overall tree development variable and used this variable in testing for treatment effects. The results proved to be robust in respect to alternative specifications of the estimated model further providing unambiguous indication of the treatment effects. The final analysis included multigroup estimation that aimed at comparing model specification in individual treatment subsamples.

The main results from this paper suggest that multivariate methods could be successfully used in experimental design with good potential of providing unambiguous and more substantively interpretable results. The main requirement for using methods described in this paper is multivariate normality of tree development indicators. Our approach to dealing with the empirically observed deviations from normality was to transform the data, for which purpose the normal scores technique was used. This technique proved highly satisfactory though other, perhaps

theoretically more justifiable techniques (e.g., Box-Cox), could be used alternatively. Further research in this direction could focus on application of multiple comparison procedures and on adjustments of the significance levels. Such adjustments might be needed because the latent-scores variables ware obtained from a multivariate statistical model that has been build or modified already on inferential grounds. Additionally, alternative paths of model building and different estimation methods could be considered. Data with more pronounced time series dimension might be modelled by extensions of the presently widely used structural equation models such as mixture modelling (Muthen, 2002) and growth mixture modelling (Muthn et al., 2001). Most importantly, the data gathering process and decision of which tree development indicators need to be measured would be improved by knowing how each measure fits into the covariance structure giving rise to overall tree development.

Graph Distance in Multicriteria Decision Making Context

Lavoslav Čaklović (Zagreb, Croatia)

In this article we are introducing a dissimilarity measure on a certain class of oriented graphs arising in the context of group decision making. Group consensus is done using Potential Method (PM) developed by author. Main step in PM is to determine normal integral (value function) X of the preference flow F on the set of alternatives by solving the normal equation

$$B^\tau BX = B^\tau F, \quad \sum_{i=1}^{m} X_i = 0$$

where B is incidence matrix of the graph.
Two decision makers can define two different preference flows on the set of alternatives. If they induce the same value function they are considered equivalent. In this context, equation

$$\delta(F_1, F_2) := \|X_1 - X_2\|$$

gives a good dissimilarity measure on the set of preference flows over the same set of alternatives.
We used this measure to define distance matrix for a set of individual preferences in a certain Multi Criteria Group Decision problem. During the research students were asked to give preference flows, for a certain number of criteria, over the set of their teachers. The experiment was organized at two different places. Students of the first group, 29 of them, were allowed to select criteria and alternatives on their own choice, while students of the second group, 48 of them, were forced to select all criteria and all alternatives. Web interface of the questionnaire was (is) available one the local server. Students were allowed to see only their own ranking after processing their input.
In both cases, dissimilarity matrix of individual preference flows was calculated for each group and `Statistica` software was used for clustering. In the first group, outlayers were discovered and eliminated from the group consensus flow. Fine analysis of criteria ranking was not possible due to the lack of information. In the second group two clusters were present and group consensus generated by them were not significantly different. This lead to conclusion that dissimilarity measure, defined above, is very sensitive as a function of preference flow.
Another applications of the method, not discussed here are: discovering hidden conflicts in society, testing a homogeneity of micro-social structures, like sport teams . . .

Clustering Symbolic Objects Described with Distributions

Simona Korenjak-Černe and Vladimir Batagelj (Univ. of Ljubljana, Slovenia)

In standard data analysis objects are usually described with vectors. Each component of a vector corresponds to a special descriptor of some property of object – variable. Clusters of similar objects are described with the mean value (average, median, mode) for each component over cluster. More detailed information about cluster offers the description with distributions. Such description is a special kind of symbolic object.

In our contribution we shall present two adapted clustering methods based on such symbolic descriptions of units and clusters. The adapted leaders' method is used to reduce a large dataset of units to smaller set of clusters represented by their symbolic descriptions – leaders. These leaders are clustered further using an adapted agglomerative hierarchical method to reveal their interconnections.

The advantages of the proposed methods will be presented by analyzing a concrete dataset.

References

1. Batagelj (1988): Generalized Ward and related clustering problems. In: Bock, H.H. (Ed.), *Classification and Related Methods of Data Analysis*, North-Holland, Amsterdam, 67–74.
2. Bock, H.H. and Diday, E. (2000): *Analysis of Symbolic Data. Exploratory methods for extracting statistical information from complex data.* Springer, Heidelberg.
3. Hartigan (1975): *Clustering Algorithms*. Wiley, New York.

On the Evolution of Signed Social Networks
Patrick Doreian (University of Pittsburgh, USA)

To the extent that social relations are signed - in contrast to being only positive - it is important to study them. In turn, the study of signed networks is helped by having sound substantive ideas, good measurement and relevant data analytical tools. At face value, we have all three. 'Structural balance theory', in its various guises, provides the substantive foundation. Data have been collected for signed relations and we have a variety of flexible and useful tools for analyzing such signed network data. Some of these results are presented. Yet balance theory, measurement and signed blockmodeling tools are insufficient for establishing a general theory of balance and associated tools. In its empirical form, the core feature of structural balance theory is its focus on dynamics. Using some of the few through time data sets that exist, modest support for balance theory can be found. But there is even more contradictory evidence and this information raises a variety of important questions for the empirical study of balance dynamics. It is fruitful to pursue these questions. Some recent, and provisional, results are presented together with an agenda for the study of the evolution of signed networks.

A Parametric Regression Model by Minimum L_2 Criterion: A Study on Hydrocarbon Pollution of Electrical Transformers

Alessandra Durio and Ennio Davide Isaia (University of Turin, Italy)

The purpose of our work is to investigate the use of L_2 distance as a theoretical and practical estimation tool for parametric regression models. According to Scott (2001) we consider the simple regression model $Y = a0 + a1x + \varepsilon$, where ε has a density $f(\varepsilon \mid \theta)$, and, in order to estimate the parameter of the density of the errors, we apply the minimum L_2 criterion (briefly L_2E) which consists in minimizing

$$\hat{\theta}_{L_2E} = arg \min_{\theta} \left[\int f(\varepsilon \mid \theta)^2 d\varepsilon - 2n^{-1} \sum_{i=1}^{n} f(\varepsilon_i \mid \theta) \right] \qquad (1)$$

If we furthermore make the assumption that $f(\varepsilon \mid \theta)$ is the density of a $N(0, \sigma^2)$ minimizing equation (1) reduces to minimize

$$\hat{\theta}_{L_2E} = arg \min_{\theta} \left[(2\sigma\sqrt{\pi})^{-1} - 2n^{-1} \sum_{i=1}^{n} \phi(\varepsilon_i \mid \theta) \right] \qquad (2)$$

Since $\varepsilon_i = y_i - a_0 - a_1 x_i$ the L_2E criterion allows us to obtain the estimates $\hat{a}_0$, $\hat{a}_1$, and $\hat{\sigma}$ simultaneously.

The approach based on the L_2E criterion is particularly helpful in all those situations involving the study of large data sets. Handling large samples with a consistent numbers of outliers and or extreme values, situations in which a maximum likelihood regression models are frequently unstable, L_2E criterion allows to detect fractions of bad data (Scott, Simar, and Wilson, 2002). After explaining the use of the methods with some simulated examples, we will point out the results for an industrial case study. In order to prevent and mitigate the hydrocarbon pollution of electrical transformers, the firms operating in the field of pollution contamination risk usually run chemical analysis on the oil of the transformer themselves. The method based on L_2E criterion will be applied to study the relations between the PCB's level and several technical characteristics of the oil. To estimate the parameters of the model, according to the L_2E criterion, we implemented some routines with **R** software.

References

1. Bowman, A. W. and Azzalini, A. (1997): *Applied Smoothing Techniques for Data Analysis*, Clarendon Press, Oxford.
2. Reiss, R. D. and Thomas, M. (1997): *Statistical Analysis of Extreme Values*, Birckhauser, Basel.
3. Scott, D.W. (2001): Parametric Statistical Modeling by Minimum Integrated Square Error, *Technometrics*, **43**.

4. Scott, D.W. (2002): Multivariate Density Estimation,Wiley, NewYork.
5. Scott, D.W., Simar L., and Wilson P. (1995): Modeling the Stochastic Frontier with a Non- parametric Criterion, *Actes des XXXIV es Journees de Statistique*, Bruxelles et Louvain-la-Neuve.
6. Wand, M.P. and Jones, M. C. (1995): *Kernel Smooting*, Chapman & Hall, London.

Can Simulation Techniques Contribute to Microsociological Theory? The Case of Learning Matrices

Anselm Eder (University of Vienna, Austria)
Walter Gutjahr (Institut for Statistics and Decision Support Systems, Austria)

Nearly all simulation techniques have one big disadvantage in common: oversimplification, leading to unrealistic results. And nearly all simulation techniques have one big advantage in common: oversimplification, which is the only way to express a theoretical model in clear terms. The oversimplification in our simulation model consists in assuming two interacting partners whose actions/reactions are determined by only two sets of parameters: 1.) a matrix of reaction probabilities which is updated according to the subjective evaluation of each reaction of the partner to the actor's behaviour at each interaction, and 2.) a payoff matrix, reflecting each partner's subjective evaluation for each pair of action/reaction, which remains stable over a longer sequence of interactions. Applications of this model to several problem areas, such as socialization agents, game theory approaches, and Ant Colony Organization in previous publications of the authors, have shown some practical results. Here, we want to focus on three paradigms which we believe can be deducted from our work:
1.): The law of sociodynamics: Social systems whose organization is similar to the model conditions formulated above, tend to a decrease in entropy as they get older, in contrast to physical systems, which seem to do the opposite.
2.): Cultural values prevail over individual behavioural dispositions: We believe to have found an argument that social systems with properties similar to our model assumptions are likely to provide individuals with behavioural dispositions which depend a lot more on the (culturally more stable) values attributed to behaviour than on initial behavioural dispositions, reflected in reaction probabilities.
And 3.): The slower the learning process, the better the results. Optimal solutions to typical dilemma situations such as the iterated prisoner's dilemma are found more frequently when individuals show 'silly' behaviour, i.e. the rate at which they change their behavioural dispositions is low, whereas a quick learning rate results more often in sub-optimal results for both partners.

Segmentation Analysis based on a Logit Transform of a Dichotomous Dependent Variable

Luigi Fabbris and Maria Cristiana Martini (University of Padua, Italy)

Segmentation analysis is a multivariate statistical method aimed at stepwise partitioning of a sample on which a dependent variable and a set of predictors were observed. If the dependent variable is dichotomous, the conventional segmentation analysis is problematic. Let us consider a dependent dichotomous variable y and a predictor $x = x_i, i = 1, \dots K$ with K response categories, observed on n sample units; in the asymmetric model ($Y \leftarrow X$), x is one of the p predictors which are candidates for the segmentation. In this paper we suggest the *logit* transform of the dependent variable frequencies, that is:

$$\text{logit}(\pi(y|X)) = \ln \frac{\pi(y|X)}{1 - \pi(y|X)} \tag{1}$$

where $\pi(y|X)$ is the expected value of y conditioned to a set of predictors, and ln is the natural logarithm. The probability $\pi(y|X) = \Pr(y = 1|X)$ varies between 0 and 1. From now on, for simplicity, we will indicate π instead of $\pi(y|X)$. The partition of the sample is evaluated, at each analytic step, with reference to the difference between the *logit* transforms of the proportion of the y variable in the two sub-samples considered for partition. Then, the best partition of the categories of x is the one which maximizes the difference between the logit transform of the proportions of y. The proportions of y are calculated for each combination of the response categories of the variable x in two groups. This corresponds to maximizing the value of $\delta[logit(\pi(y|x1, x0))]$:

$$\max\big[\text{logit}(\hat{\pi}(y|x_1)) - \text{logit}(\hat{\pi}(y|x_2))\big] = \max\big[\ln \frac{\hat{\pi_1}(1 - \hat{\pi_2})}{(1 - \hat{\pi_1})\hat{\pi_2}}\big] = \tag{2}$$

$$\max\big[\ln \frac{n(y|x_1)[n - n(y|x_2)]}{[n - n(y|x_1)]n(y|x_2)}\big] \tag{3}$$

where x_1 and x_0 are two complementary sets of the response categories of x, and $\hat{\pi}(y|x_i) = p(y|x_i) = p(yi)$ is the proportion of y conditioned to the group of response categories in x_i.

The criterion-function $\delta[logit(\pi(y|x1, x0))]$ is infinite when the frequencies of y are accumulated in one of the two groups, and generate a null frequency in the complementary group. The 'perfect partition' is what we want to obtain by a partitioning procedure, but the infinite value may also be caused by a combination of a low frequency of the event $y = 1$ in the population, together with a small size of the group x_i. Different solutions to this problem are considered:

- Adding a constant to the absolute frequencies in formula (3);
- Testing the statistical significance of the groups which are formed using the $\delta[logit(\pi(y|x1, x0))]$ criterion;
- Defining lower bounds for the frequencies n_i of the groups x_i, in order to stop the analysis in sub-samples of exiguous size.

The performance of the proposed approach to segmentation analysis is evaluated and compared to more traditional partitioning methods by means both of simulated data and of an application on a real data set.
This approach to segmentation analysis is based on an approximation of the relative risk, and hence it has a very immediate interpretation as a tool to search for causes of social or health risks.
The methodology of analysis is examined in statistical terms also with reference to look-ahead option for interaction detection, monotone relationship between the dependent variable and an ordinal predictor, the so-called 'premium for tree symmetry', differences between binary and ternary segmentation, 'pruning' options to simplify the final tree.

A Data Mining Analysis of Slovenian Dynamic Enterprises

Bogdan Filipič (J. Stefan Institute, Ljubljana, Slovenia)
Viljem Pšeničny (GEA College of Entrepreneurship, Portorož, Slovenia)

Fast growing companies, also known as dynamic enterprises or gazelles, are the most vital subjects of modern economies. They have been intensively studied in the recent years and several multidimensional models have been developed to predict their growth and performance. Following the research results for the US and EU dynamic enterprises, we defined 320 variables to describe in detail 134 most successful dynamic enterprises selected from the database of Slovenian companies. For the analyzed enterprises we examined a number of growth factors and indicators (Pšeničny, 200, 2001). However, due to a high number enterprise characteristics, statistical analysis of the collected data was not sufficient to reveal how and why the companies grow. To obtain a better insight into the company growth, we performed a data mining analysis and extracted several models to predict the growth (Pšeničny and Filipič, 2002).

Data mining is a novel field of computer science dealing with the extraction of implicit, previously unknown and potentially useful information from data. It relies on the methodology of machine learning (Mitchell, 1997), which includes induction of decision trees, classification rules, regression models and other types of models. The models derived with these techniques represent generalizations of the input data and can be used for classification, prediction and explanation of the explored phenomena. A well explored machine learning approach is learning from examples, also referred to as inductive machine learning. In this approach, examples of problem situations are submitted to a learning system which induces a general description of the underlying concepts useful for problem solving. The resulting concept descriptions can have the form of decision trees or if-then rules. Learning examples can often be very naturally described with attributes and classes. Attributes represent features from the considered domain, while class defines how an example with given attribute values is treated or classified.

In analyzing the data on Slovenian dynamic enterprises we used an increasingly popular machine learning software Weka (Waikato Environment for Knowledge Analysis) (Witten and Frank, 2000), developed at the Department of Computer Science of the University of Waikato, Hamilton, New Zealand, and accessible at `http://www.cs.waikato.ac.nz/ml/weka`.

Descriptions of the companies were treated as learning instances, where enterprise characteristics were considered as attributes, and their performance values as classes to be predicted with the derived decision trees. Several growth indicators were examined, such as David Birch economic growth index (DaBEG) (Birch, 1987, 1993), growth rate of income, growth rate of profit before taxes, growth rate of return on equity, growth rate of return on assets, and others. The induction

procedure was tuned to obtain understandable decision trees of a reasonable complexity. Their prediction accuracy was tested both on learning examples and on test examples via cross validation.

The induced models point out the most relevant growth factors and clearly show the relationships between them and the observed indicators. In our view, the results are important not only for dynamic entrepreneurs managing growing enterprises but also for investors and venture capitalists. Since most dynamic enterprises grow very fast, it is of the greatest importance to avoid crucial mistakes in managing their growth. We believe that applying this method to a representative sample of fast growing companies from any country could help to run the gazelles with less risk and uncertainty.

References

1. Birch, D. (1987): *Job Creation in America*. New York: Free Press Macmillan.
2. Birch, D. (1993): Dynamic Entrepreneurship and Job Creation. The U.S. Experience. In Derek, F. A. (Ed.), *Dynamic Entrepreneurship in Central and Eastern Europe*, 1322. Hague: DELWEL Publishers.
3. Mitchell, T. M. (1997): *Machine Learning*. New York: McGraw Hill.
4. Pšeničny, V. (2000): Characteristics of Fast Growing Small Enterprises in Slovenia and Conditions for Their Even Faster Growth. In Vadnjal, J. (Ed.), *Dynamic Entrepreneurship for New Economy, Conference Proceedings*, 93135. Portorož: GEA College of Entrepreneurship.
5. Pšeničny, V. (2001): Lessons from the Most Dynamic Enterprises in Slovenia and EU Member States. In Vadnjal, J. (Ed.), *Dynamic Entrepreneurship for New Economy*, 2nd International Conference Proceedings, 1929. Portorož: GEA College of Entrepreneurship.
6. Pšeničny, V. and Filipič, B. (2002): A Data Mining Approach to the Modeling of Dynamic Enterprise Growth. To appear.
7. Witten, I. H. and Frank, E. (2000): *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*. San Francisco: Morgan Kaufmann.

The Likelihood Ratio Test Statistic for Ordering Hypothesis: A Comparison of Asymptotic Methods, Parametric Bootstrap and Bayesian Methods

F. Galindo-Garre, J.K. Vermunt, and M.A. Croon
(Tilburg University, The Netherlands)

In the social sciences, the variables and relationships studied are often of an ordinal nature. Sometimes, methods for nominal analysis are used that ignore information about the order of the categories. Other statistical tools available, such as correspondence analysis or log-linear and log-bilinear association models, assign scores to the categories of the ordinal variables, but these models are not strictly ordinal unless the score parameters reflect the direction o the association. In contrast, nonparametric approaches have been developed that permit the definition of more intuitive hypotheses. The nonparametric methods impose log-linear inequality restrictions on the probabilities. Several methods of estimation and testing have been developed (for example, see Robertson, Wright, and Dykstra, 1988; Croon, 1990,1991; Dardanoni and Forcina, 1998, and Vermunt, 1999). However, these methods have not been extensively used until now.

To test models with log-linear inequality constraints, the statistic used is the likelihood-ratio statistic (LR). This statistic follows, under some regularity conditions, an asymptotic chi-squared distribution with k degrees of freedom, where k represents the difference between the number of parameters in the two models. Unfortunately, when inequality constraints are imposed, k cannot be determined, because the number of model parameters depends on the sample. The asymptotic distribution of LR is not a single chi-square distribution, but a mixture of them called the chi-bar squared distribution. The computation of the weights of the chi-bar squared comprises the main difficulty of the asymptotic approach. The analytical solutions are only available if the number of restrictions is smaller than 5. Different methods to approximate or simulate the weights of the chi-bar-squared distribution have been developed. In this paper, several of these procedures will be explained.

In situations in which the asymptotic distribution of the test statistic is not known or difficult to calculate, bootstrapping methods can provide an alternative to estimate the distribution of the test statistic. These methods have been successfully applied to the testing of models for ordinal categorical data. For example, Ritov and Gilula (1993) proposed such a procedure in maximum likelihood correspondence analysis with ordered category scores. Vermunt and Galindo (2001) showed that parametric bootstrapping offers reliable results when applied in order-restricted row-column models. However, the parametric bootstrap is strongly affected by the model and can be too liberal in some situations. To solve these problems some modified parametric bootstrap procedures have been proposed. Another alternative can be to use Bayes factor or Bayesian P-values with diffuse priors.

The purpose of this paper is to give an overview of methods to estimate and test

models with inequality constraints. Concerning the testing, different methods to approximate the asymptotic distribution will be compared to both the bootstrap and the bayesian method. The methods discussed will be illustrated with the analyses of empirical data.

References

1. Croon, M.A. (1990): Latent Class Analysis with Ordered Latent Classes. *British Journal of Mathematical and Statistical Psychology*, **43**, 171-192.
2. Croon, M.A. (1991): Investigating Mokken Scalability of Dichotomous Items by Means of Ordinal Latent Class Analysis. *British Journal of Mathematical and Statistical Psychology*, **44**, 315-331.
3. Dardanoni, V. and Forcina A. (1998): A unified approach to likelihood inference on stochastic ordering in a nonparametric context. *Journal of the American Statistical Association*, **93**, 1112-1123.
4. Robertson, T., Wright, F.T., and Dykstra, R.L. (1988): *Order Restricted Statistical Inference*, New York: John Wiley.
5. Ritov, Y. and Gilula, Z. (1993): Analysis of Contingency Tables by Correspondence Models Subject to Order Constraints. *Journal of the American Statistical Association*, **88**, 1380-1387.
6. Vermunt, J.K. (1999): Nonparametric Models for Ordinal Data. *Sociological Methodology*, **29**, 187-223.
7. Vermunt, J.K. and Galindo, F. (2001): RC Association Models with Ordered Row and Column Scores. Submitted.

Testing for no Effect and Selection Model in Functional Linear Regression Model

Aldo Goia (Universitadel Piemonte orientale, Italy)

The functional regression model is a regression model where the link between the response (a scalar) and the predictor (a random function) is expressed as an inner product between a functional coefficient and the predictor.
To characterize the models we consider the random variable (r.v.) (X, Y) defined on the probabilized space $(\Omega, \mathcal{A}, \mathbb{P})$, where Y is a real r.v. and X is a square integrable real random function defined on some compact set $\mathcal{T}$ of $\mathbf{R}$, and we take:

$$Y = \mu + \int_{\mathcal{T}} \psi(t) X(t) \mathrm{d}t + \varepsilon \tag{1}$$

where μ is a real constant, ψ is a square integrable real function defined on $\mathcal{T}$, and ε is a real r.v. such that $\mathbb{E}(\varepsilon) = 0$. Model (1) may be written as

$$Y = \mu + \Psi(\{X(t),\ t \in \mathcal{T}\}) + \varepsilon \tag{2}$$

where Ψ is a continuous linear operator.
In order to estimate the functional coefficient ψ, or the operator Ψ directly, some estimators are introduced (Functional Principal Component Regression Estimator, Smooth Principal Component Regression Estimator, Penalized B-Splines Estimator).
We first focus our attention on the Functional Principal Component Regression Estimator, and in particular our aim is to apply and compare some different selection methods, which have been proposed in classical regression field.
Furthermore we are interested in testing the no effect of the model, i.e. the nullity of the functional coefficient. We introduce some tests: while the first one is based on the cross-covariance operator of (X, Y), the second is based on a pseudo-likelihood ratio test statistic. The methods are illustrated and compared through simulations.

References

1. Cardot, H., Ferraty, F., and Sarda, P. (1999): Functional Linear Model. *Statistic and Probability Letters*, **45**, 11-22.
2. Cardot, H., Ferraty, F., and Sarda, P. (2002): Spline Estimators for the Functional Linear Model. *Preprint*.
3. Cardot, H., Goia, A., and Sarda, P. (2002): Testing for no effect in functional linear regression models, some computational approaches. *Preprint*.
4. Cardot, H., Ferraty, F., Mas, A., and Sarda, P. (2002): Testing Hypothesis in the Functional Linear Model. *Scand. J. of Statis.*, to appear.

5. Ramsay, J.O. and Silverman, B.W. (1997): *Functional Data Analysis*. Springer-Verlag.

A Unified Approach to Biplots

John C. Gower (The Open University, London, UK)

Biplots are concerned with presenting information on the variables and units, typically for data presented in a data-matrix X. The archetypal biplot is represented by Cartesian coordinate axes in which the calibrated axes (usually orthogonal) represent the variables and the units are represented as points. There are two questions that relate to Cartesian axes (i) given a case x, that is a row of X, where is the corresponding point P and (ii) given a point P what are the associated values of x? Statistical biplots may differ from this classical set-up in several ways. Firstly we usually have only an approximation to the full-dimensional Cartesian representation. This induces non-orthogonal axes and complicates the answers to (i) and (ii). Secondly, we may use unusual metrics based on dissimilarity coefficients or on variants of Mahalanobis distance. Thus, extensions are needed to embrace methodologies where X is represented by various forms of metric and nonmetric multidimensional scaling. Thirdly, we may wish to include categorical variables in X, and these may be either of purely nominal form or they may be constrained to be ordered categories. An extension to the Cartesian system is needed to cope with categories.

It will be shown how all these needs may be handled in a unified way. Orthogonal projection is a key concept in the use of Cartesian axes; the more general concept of the nearest point to a set is invoked to handle the generalisations. For quantitative variables we end up with calibrated axes, possibly nonlinear, and for categorical variables, sets of labelled points representing the different category-levels. The cases continue to be represented by points.

Successful Recruitment for Online Access Panels

Anja S. Göritz (University of Erlangen-Nürnberg, Germany)

An Online Access Panel is a pool of Internet users who have signed-up to take part in WWW-based studies from time to time. Because participants are pre-recruited, surveys and experiments can thus be realized quickly and in a cost-efficient and methodologically favourable manner. A 4x2-factor recruitment experiment was conducted in a University-based Online Access Panel. The number of successfully recruited panelists was the dependent variable. The medium of solicitation was varied at four levels: Each time 500 randomly selected individuals were sent an invitation either by e-mail, letter, or fax, or they received a flyer. Furthermore, half of the solicitations mentioned a lottery into which the newly registered panelists would be entered, whereas the other half of the solicitations did not mention this lottery. There were marked differences in recruitment success between the four different means of contact. With the exception of flyers there were no significant differences due to the lottery information. On the basis of a cost-benefit analysis, recommendations for recruiting panelists are derived.

Alternative Applications of the Cox Model in Outcome Research

Dario Gregori and Patrizia Rozbowsky (University of Trieste, Italy)
Miriam Isola (University of Udine, Italy)
Alessandro Desideri (Department of Cardiology, Main Hospital, Veneto, Italy)

This paper focuses on the usage of the Cox regression model for estimating relevant quantities (e.g.: mean, median) in the analysis of two major outcomes in clinical research: costs and quality of life. This approach, which dates back to early 90's, let the analyst to avoid common problems that arise with the classical linear regression models (robustness, presence of outliers, normality assumptions, need for log-transformation).
Particular emphasis will be given to the case of multicenter studies, which are characterized usually by a great heterogeneity among centers with respect to these quantities. These issues can be easily approached in the Cox setting. Additional information gained analyzing the center-effects is also very useful for the researchers.
All topics will be illustrated on the basis of the COSTAMI dataset, which is a randomized, parallel groups multicenter study aimed at comparing two strategies of early discharge of patients with an uncomplicated AMI (Acute Myocardial Infarction) in terms of events, costs and quality of life.

Collaborative Approach to Web Access Analysis

Marko Grobelnik and Dunja Mladenič (J. Stefan Institute, Ljubljana, Slovenia)

This paper describes an approach to collaborative analysis of Web accesses to a Web site based on visits of the users to a set of Web pages. For our purpose, the Web accesses were given by pairs of user identification and page identification. We have developed a recommendation system that enables sharing information about the visits of other users to the same Web site and a system for clustering the users based on the similarity of their interests. Additionally we also cluster the Web pages based on their content and based on the user visits.

Introduction

Web access analysis is an evolving area strongly influenced by the growing number of well organized Web sites where the owners are interested in improving the site quality and visibility. Each access to the Web site leaves a footprint in the file of the Web site server, containing at least information about the user/computer (ip number) that the request came form and the Web page URL that was requested. Usually we can also find time of the request, URL of the page that the request was made from (for instance following a hyperlink). Sometimes there is additional information obtained by tracking the individual users, for instance by requesting user identification in order to access the Web site. This large amount of information is a valuable source of information that is in Web access analysis addressed using Data Mining methods. One of the most popular problem in Web access analysis is finding the most common sequences of Web pages visited one after another, in one session of the user.

In this paper we address the problem of sharing information between the users by identifying the users with similar interests and identifying similar pages. This kind of approach is usually referred to as collaborative approach to user modeling (Mladenič, 1998) and can be used for different problems, such as movie or book recommendation (Maes, 1994). We extend it on the problem of Web access analysis combining it with clustering of Web pages based on either visits or the page content. We compare the proposed collaborative approach to clustering Web pages to the content based approach and show directions for combining them. The experiments were performed on the data from Portuguese National Statistics Office having Web site for providing statistical data to the registered users. The Portuguese National Statistics Office (INE) is the governmental agency who is the keeper of national statistics and has the task of monitoring inflation, cost-of-living, demographic trends, and other important indicators has gone for Data Mining. Their managers believe that Data Mining can tell them more about the users of Infoline (`www.ine.pt`), the web site that makes statistical data available to the Portuguese citizens (Jorge and Alves, 2001).

Problem definitions

More specifically, we attacked four problems:

1. Identification of user profiles and groups of similar users based on their behavioral patterns. The user profiles were build in two ways, both using weighted sparse vector representation: (1) by using URLs clicked by the users, and (2) the text contents of the visited web pages. In the first step, the log file in the form of (User-Id, Url-Id) pairs was transformed into bipartite graph representing relationships between users and web pages. Out of the graph sparse feature vector for each user was extracted. Features represented URLs (and words from the documents in the second representation) of the web pages clicked by the users. Since the particular users visited only minor subset of all web pages, vectors included only minor part of non-zero components. For the similarity function between the vectors the classical cosine similarity was used. In the following steps of the analysis various clustering methods were applied on the vector set to identify user groups and user profiles.
2. Calculation of the web page recommendation based on the behavioral patterns of the users. The result was a list of URLs for each web page on the web site of type: 'Users who visited this page, visited also the following pages...'. We used similar methodology as the one from the first problem. First, we created bipartite graph of user and web-page relationships. Next, we constructed a sparse feature vector for each web page represented with the user-ids visited that page. Using features vectors we created page recommendations using k-nearest-neighbour method.
3. Visualization of the most characteristic user groups and web pages based on the behavioral patterns of the users. We used two types of visualization: (1) combination of K-means clustering and simulated-annealing based multidimensional-scaling, and (2) visualization of hierarchical partition of objects produced by top-down hierarchical clustering.
4. Calculation of the user and web page importance weights. We slightly adapted the well-known PageRank algorithm for the identification of the importance weight of the vertices in the bipartite graphs.

References

1. Alipio, J. and Mario, A.A (2001): *End of Phase I summary on INE Infoline: a Sol-Eu-Net end-user Phase I project*, Sol-Eu-Net IST-1999-11495 Progress Report.
2. Maes, P. (1994): Agents that Reduce Work and Information Overload, *Communications of the ACM*, **37**, 30-40.

3. Mladenič, D. (1998): Text-learning and intelligent agents, *IEEE EXPERT, Special Issue on Applications of Intelligent Information Retrieval*.

Joint Modelling of Longitudinal and Event Time Data
Robin Henderson (Lancaster University, United Kingdom)

Methods for the combined analysis of survival time and longitudinal data are reviewed and illustrated. 'Joint modelling', as the topic has become known, integrates methods developed to handle dropout in longitudinal trials with techniques developed for survival data analysis with intermittent time-dependent covariates subject to measurement error. The talk will cover quick and simple exploratory methods, modelling strategies and their limitations, estimation and diagnostics.

Collecting Ego-Centered Network Data via the Web

Valentina Hlebec, Katja Lozar Manfreda, and Vasja Vehovar
(University of Ljubljana, Slovenia)
Zenel Batagelj (CATI, Ljubljana, Slovenia)

Personal interview is a typical data source for ego-centered social networks owing to their complex data structure. However, Web data collection can substantially reduce the costs and time required for such data collection.
It is especially suitable for special populations familiar with Internet tools. Special attention to questionnaire design is needed if the respondent will complete the Web questionnaire by him or her self. One trial in the collection of ego-centered networks via the Web was done in the annual RIS (Research on Internet in Slovenia) Web survey conducted by the Faculty of Social Sciences, University of Ljubljana in summer 2001. Respondents were randomly split into four groups. Each group received a name generator for one type of social support: material, informational, emotional support or social companionship, and a set of questions regarding alters for each alter they named in the network generator. Dropouts from the questionnaire are studied with respect to the number of listed alters and characteristics of respondents. Preliminary analysis shows that the Web can be used as a data collection method for ego-centered social networks. However, special attention is required when designing the graphic layout of name generators as well as with the wording of instructions. In particular, the number of alters should be limited in some substantial way since respondents who name many alters tend to quit the questionnaire before answering additional questions regarding these alters.

Comparing Demographic Variables across Nations: Concepts and Indicators
Jürgen H.P. Hoffmeyer-Zlotnik (ZUMA, Mannheim, Germany)

In attitude measurement functional equivalence of the measurement can be controlled by a process of forward and back translation. In the case of socio-demographic variables the procedure of translation and back-translation is not promising because these variables depend on societal systems and on culture but not on language. Behind socio-demographic variables often one can find a set of different concepts. These concepts depends on national structure and on national institutions and organizations. For the measurement of socio-demographic variables one need knowledge about the different concepts, the cultural structure behind the variables and the national indicators of measurement. With this knowledge functional equivalence of the measurement of socio-demographic variables can be reached by harmonizing the variables. In the presentation the way from national concepts to indicators useable in international comparison is shown for different variables: income, education, household and race.

Conceptualizing and Measuring Culture in Surveys: Case Study in Republic of Croatia

Jasna Horvat (University Josip Juraj Strossmayer, Osijek, Croatia)
Sanda Katavič (Fakultet za turistički i hotelski management u Opatiji, Croatia)
Martina Mikrut (University Josip Juraj Strossmayer, Osijek, Croatia)
Irena Ograjenšek (University of Ljubljana, Slovenia)

Questions regarding culture as a specific resource constitute a wide field of surveys and require interdisciplinary approach. The emphasis is put on economics and applied business studies on the one hand and economic geography on the other hand, but they would also have a strong base in political economic sociology and 'cultural studies'. However, a problem shared by all the mentioned disciplines is the fact that there are different meanings of 'culture', so that the development of different measuring instruments which would focus on the presence of the culture in the society and economy is necessary. The authors' intention was to develop a measuring instrument which would primarily measure the perception of cultural indicators, interest in specific programs/attractions within the context of urban culture, as well the personal values. The measuring instrument includes the specific quality of culture concept in the city of Osijek. The answers of randomly chosen respondents in Osijek were collect applying pencil and pen method ($n = 470$). Except for testing the mentioned instrument, the intention was to provide a critical review of all the usage segments of the developed instrument with the purpose of its improvement.

Connectedness of Political Actors in Newspaper Reporting: The Newspaper Delo Case

Matej Jovan (Parsek d.o.o, Slovenia)

In our research we applied several methods to obtain and analyse data related to simple 2-variable model:

1. analysis of value determined statements (Osgood et al., 1973),
2. Thurstone scale (Thurstone and Chave, 1928),
3. network analysis (signed graphs) (Doreian and Mrvar, 1996), and
4. discriminant analysis.

Osgood's method refers to content analysis, which was in our case specifically appropriate as primary data source was text (newspaper articles). It represented basic methodology within we were able to obtain or/and prepare data sufficiently enough for further analysis. With this method solely we were able to assign values to one indicator of independent variable (value-rational agency).
Creation of Thurstone scale helped us to assign values on the second indicator of independent variable (ideological orientation) though providing extensively valid measurement.
With application of network analysis (signed graphs) we obtained data for the dependent variable (connectedness). We were able to conduct network analysis on the basis of convergent definitions of relations by Heider and Osgood.
After we had obtained all data, we were able to test our hypothesis. In this case the discriminant analysis was applied.

Analysis of Different Approaches to Subgroup Discovery
Branko Kavšek and Nada Lavrač (J. Stefan Institute, Ljubljana, Slovenia)

Developments in the area of machine learning called descriptive induction have recently gained much attention, especially the field of subgroup discovery. The task of subgroup discovery is to discover subgroups of a population of individuals that are statistically ”most interesting”with respect to some property of those individuals. In this paper we try to analyze different approaches to subgroup discovery. Recently developed algorithms in this field, such as MIDOS, CN2-SD and DIARY are studied and their suitability for different types of problems is discussed.

REFII Model – Model for Recognition Patterns in Time Series
Goran Klepac (Raiffeisen bank, Zagreb, Croatia)

REFII model is an authorial mathematical model for recognition patterns in time series.
It is important to say that REFII model is not a closed system, meaning that we have a finite set of methods. It is in the first place a model for a transformation of values of time series, which prepares data that are used by different sets of methods in a domain of problem space in order to solve problems.
The purpose of the model is to:

- discover seasonal oscillation,
- discover cyclic oscillation,
- discover rules from time series,
- discover episodes from time series,
- discover similarity of time segments,
- discover correlation between time segments,
- discover rules from in domain of finances from time series,
- connect time series and standard data mining methods,
- analyze time series with the help of data mining methods (clustering of time segments, classification of time segments)

The mathematical background is focused on two basic elements of time series: shape and area beneath the curve. All submodels, which solve specific problems in different domains, use these two elements in all algorithmic procedures.
REFII model is able to analyze every time series, which is represented by values. In the first step of the analysis, after preparing the original data, we make a transformation of time series in REFII model values. The next step is to select an appropriate method inside REFII model to analyze data. The selection of the method is determined within the scope of the analysis.
Methods could be focused on discovering: seasonal oscillation, cyclic oscillation, hidden rules, episodes, similarity of time segments, correlation of time segments, clusters, or links between time segments.
The advantage of REFII model is its possible application in many different areas like finance, medicine, voice recognition, face recognition, text mining. As an illustrative example of how the model could be efficient, the accent in the presentation will be put on its application in banking and finance.
REFII model could be successful in discovering rules of bank clients' behavior, when their behaviors depend on a time dimension. These rules could be connected

with behavior when using transactional services, behavior with money management, prediction of using a new banking service in an expected period of time, or similarity between a client and a market group. We could integrate REFII model in a query language for time series, and in that case we would get a powerful tool for creating complex algorithmic procedures based on pattern recognition methodology. A large number of different methods could be used as elements of the query language, but we could also use every of them like a single method for a single analysis.
The results of the analysis could be presented in IF-THEN form, which means that it is possible to use an expert system shell and metarules to analyze knowledge from time series analysis.
REFII model is an open system, which could be used like a connection between time series and other mathematical, statistical, and data mining methods.
Main reasons why the system has been developed are:

- there is no unique methodology in data mining for time series analysis,
- the existing system for time series analysis, that is a part of statistics, doesn't provide the answers in the spirit of the data mining methodology, and doesn't give the solutions on how to connect the standard data mining methods with time series,
- when we have a problem in a domain of specific business area and time series, we have to develop a whole new system whose concept is often not compatible with other similar solutions,
- to find an efficient system that will be able to predict events in time series, search for similar patterns, search for seasonal and cyclic oscillation, find rules from time series, find correlation between time segments, or two or more time series,
- to use the potential of time series in medicine (EEG - searching for patterns of mental illness, EKG - searching for patterns of heart illness), face recognition (REFII model in 3D space), voice recognition, text mining, and multimedia data mining.

The first results of using REFII model are good. The concept could exist as an application that integrates a variety of sub-models, as a query language for time series, or as modules that are integrated within other applications.

Stability of Measurement of Egocentered Networks by Telephone Mode
Tina Kogovšek and Anuška Ferligoj (University of Ljubljana, Slovenia)

Data about personal networks and their characteristics are increasingly used in social science research, especially in research about the quality of life, social support and similar (e.g., Fischer, 1982; Marsden, 1987; Iglič, 1988a, 1988b; van der Poel, 1993; Hojnik-Zupanc et al., 1996a, 1996b; Schweizer et al., 1998). Since all data about a person's social network are usually obtained from the respondent himself, the quality (reliability and validity) of such measurements is a very important issue. Comprehensive studies (e.g., Ferligoj and Hlebec, 1999; Hlebec and Ferligoj, 2001; Kogovšek et al., 2002) have already established that different factors (e. g., type of social support, measurement method, respondent characteristics, respondent mood and others) have significant effects on reliability and/or validity of measurement of complete and egocentered networks.

Media Reporting about Slovenian Police

Tadeja Kolenc (Office of Director General of the Police, Ljubljana, Slovenia)

In the paper the media reporting about Slovenian Police, its representatives and units, about its tasks and its activity, either generally or specific topic will be discussed. For certain period of time the Slovenian Police PR analyzes, how polices work and activity monthly reflects in Slovenian mass media. Beside direct individual experiences with police the media reporting about polices efficiency, professionalism, legitimate procedures and also successfulness is one of the major public opinion makers. And by all means, how the public evaluate the police, determines a quality of partnership between public and police. Police can not fulfill its duty and obligations well without public support and cooperation. Therefore is important to know, how the police work is presented in media as a creator of social reality and as mediator of many social, political and economic interests.

'There is no event, if media didnt report it.' All the power of media is showed in these sentence. Already media has overtaken the role just to inform the public, but they also create in form the public perception and reaction on social reality. News is not a true mirror of reality, in our home the interpretation of its different forms has been brought day by day. In certain way media filtrates the enormous amount of news for public and for that reason some of media experts named media as gatekeepers. Media decide what is important in the world and at home, including the events on police field of activity. The roles of media in reporting about police are various, cause of their specific social and psychological impact on public: media can mobilize public power works for or against the police, they can demonize or glorify police procedures and results or they can apologize or accuse acts of police.

Analysis of media reporting about Slovenian Police includes more than six hundred articles per month. And the best method for analyzing is a content analyses. So, every analyzed article has been codified and as such entered into database. We try to achieve, that the meaning of codes are as much as possible near to the contents of articles and that the data does not loose to much meanings in this transformation. For each article the following variables are entered into Database: a headline of an article, media, destination (on which unit(s) is an article addressed), the technique of public relation has been used, topics, the value of an article and its headline (it is positive, neutral or negative for police), week and named representatives of Slovenian police in the article.

The goal of this paper is presentation how media reporting about Slovenian police. I will try to point out the differences between media according to the topics, such as crime, traffic, police procedures, cooperation with other formal institution etc. According to many studies the most preferred topic in media is crime. Is this also true for Slovenia I will answer on this question in my article.

In the analysis I will be concentrated only on reporting of major media as Tele-

vision and Radio Slovenia, commercial television POPtv, daily newspapers Delo, Dnevnik, Večer and Slovenske novice, weekly magazine Mag, Mladina, Demokracija and others national periodical magazine from January to June 2002. There are two sources for my analysis, the first used for press is the clipping, which is prepared in public relations of Slovenian Police, and the second used for the electronic media is the clipping of Government Public Relation and Media Office of Slovenia.

A Nonparametric Mean Residual Life Estimator:
An Example from Marketing Research

Kai Kopperschmidt (A.C. Nielsen, Frankfurt, Germany)
Ulrich Pötter (University of Bochum, Germany)

The mean residual life (mrl) function dynamically describes the average time to an event, depending on the time since the previous event. It provides a forecast in parallel with the development of the underlying process. From a theoretical point of view, the mrl characterizes the distribution of the process completely, but in contrast to other characterizations like the hazard rate, it has a direct interpretation in terms of average behavior.

We use Kaplan–Meier integrals (weighted averages of residual times) to construct a nonparametric estimator of the mrl. We use results from Stute (1995) to describe the asymptotic behavior of this estimator and derive an approximate variance formula.

We present a small simulation study and apply the estimator (and the variance formula) to data pertaining to purchase time behavior from the Homescan Panel, A. C. Nielsen, Germany.

References:

Stute, W. (1995): The Central Limit Theorem Under Random Censorship, *Ann. Statist.*, 23.

Web Fragmentation: A Network Analysis Approach

Gašper Koren and Matej Kovačič (University of Ljubljana, Slovenia)
Zenel Batagelj (CATI, Ljubljana, Slovenia)

Idea of resemblance of the World Wide Web (WWW) as a 'technical' and also as social structure with the social networks is not new. However, WWW was usually considered as a structure of web pages/web sites connected together with hyperlinks. That offers great opportunity of visualization of WWW as a network of web sites, but on the other hand omits the most important part which constructs network as a live construct – its' users!

WWW can be regarded as two-mode network, where first set of units are web pages and second set of units are web page visitors. Relation in this case connects each web page with its users. We can further transform this network into undirected network of web pages (and/or users). Relation between two web pages is common user of those two web pages and strength of this relation is number of users in common.

Importance and need for such information was increased with commercialization of Internet and appearance of web advertising. As one of the biggest advantages of web advertising is mentioned precise targeting of target population. Because in general we do not have socio-demographic data about each user and his identification to serve him with personalized web ads, we are trying to estimate those characteristics with matching different information from several sources: content of web site, data about visitors from meta-data of the web[1] combining with web surveys among web site user.

Matching of databases from different sources in social and especially in marketing research will likely be 'hot' topic in future. Obviously we are loosing control over response rates in classical survey research and 'learning without questionnaires' will be more and more important. Internet as an open structure with its enormous repository of secondary data about users and their behavior is great source for future research. But... what about our privacy?

[1] Specially designed computer application we developed, can capture user's IP Address, type and version of OS and Browser used, screen and color resolution, estimation of the connection speed, use of certain plug-ins (Java, Flash...), etc.

Impact of Food Additives on the Number of Somatic Cells in goats' Milk: Statistical Analysis of the Experiment

Katarina Košmelj and Drago Kompan (University of Ljubljana, Slovenia)
Andrej Blejec (National Institute of Biology, Ljubljana, Slovenia)

Introduction

Milk should consist as few micro-organisms such as somatic cells (SC) as possible. For cow milk, the standard for the permissible number of SC exists, for goat milk it is still under study. At the Zootechnical Department of Biotechnical Faculty an experiment was undertaken to assess the effect of three different food additives on the quality of goat milk. Breeders were interested in the possible reduction effect of these additives on the number of SC. These treatments were: additive DHA (of fish origin), additive ALFA (of plant origin), additive EPA (of fish origin), no additives (control treatment). The experimenters were interested in the following two questions:

- Which of these treatments reduces the number of SC in goat milk? If more than one, ascertain which is the best.
- For the best additive under study assess the duration of the reduction effect, i.e. how long has the food additive some reduction effect on the number of SC?

53 animals were involved in the experiment. It lasted 60 days and was divided into three time periods: period of animal adaptation (1st-10th day), period of food additive application (11th-15th day), period after food additive application (16th - 60th day). The number of SC was assessed daily for the morning and evening milking. From the 21st day on assessment took place in five-day interval only.

Description of the data

The data for each goat is in the form of a time series with non-equidistant spacing. Many time series have extreme outliers, missing values are common, too. The variability of SC in goat milk is very high, the variability exists between the animals and within the time span of individual animals.

Statistical analysis

The database obtained by this experiment is rather unusual and standard statistical methods can not be used directly. We overcame this problem in three steps:

Step 1: For each animal its time series was plotted. The great variability suggests the use of transformations. First we used logarithmic transformation, then we standardized the obtained time series in a special way.

Step 2: to evaluate the effect of the treatments we calculated two medians on the standardized series: Me1 for the period of food additive application (11th-15th day) and Me2 for the period after the application (16th to 20th day). One-way

ANOVA and Duncans test confirmed that treatment ALFA is the best. The same results were obtained for Me2.
Step 3: Further analysis was focused on ALFA treatment only. We started again with the graphical display and calculated the time series of the medians. The display shows that this time series has negative values in the period of food additive application and afterward. Using robust regression and Wilcoxons nonparametric test we proved that this effect lasts up to 30 days after the end of additive application.

Reliability and Validity of Response Scales : Effect of Category Numbers and Category Labels on Response Behavior

Dagmar Krebs (University of Giessen, Germany)

The content part: The measurement of this paper is discussed on the basis of repeatedly measuring attitudes in the same respondents. From each respondent there are at least two measures on political efficacy as well as on ethnocentrism. The methodological part: The paper describes the effect of the number of categories in response scales. Even versus uneven numbers of categories can have an effect on (a) response behavior and (b) reliability and validity of the measurement by either method. First, attention is given to the effect of a 4point versus a 5point response scale, both scales verbalizing each scale point. In the evaluation of the two measurement methods, descriptive measures as well as measurement models are compared. Second, the paper addresses the effect of labeling response categories by either positive or negative numbers, providing verbal labels only for the endpoint of the response scale. Since respondents were asked to decide to which degree each item is a reflection of their own attitude, the verbal labels on the end points were 'concurs not at all' and 'concurs completely'. Compared are two 7point scales, where in the first measurement scale points are labels by the numbers 1 to 7 while in the second measurement scale points are labeled by the numbers -3 to +3. Again, descriptive measures as well as measurement models are compared. to evaluate both methods. Results: The data are presently not exhaustively analyzed. However, some interesting points can already be reported here.

- item non-response is stronger in reaction to the 5point scale compared to the 4point scale
- the meaning of the middle category in the 5point scale is not at all clear it can be observed that the scale with category value labels ranging from -3 to +3 yields more positive responses that the scale with category value labels from 1 to 7. This points to the question of unipolarity versus bipolarity of attitude scales.
- additionally, dispersion is lower in the scale suggesting bipolarity by providing values from -3 to +3.
- correlations between items of the items of the -3 to +3 labeled response scale are higher and as a consequence
- reliabilities (Cronbachs α) coefficients are higher than in the scale where scale points are labeled by numbers 1 to 7.

World in Figures: A Multilingual Internet Application for Statistical Data Dissemination with an Online Data Analyzer

Sandor Kuti and Andras Vag (Heureka Research Ltd., Nagykovacsi, Hungary)

World in Figures project had developed an online system of survey and statistical databases, extended with a couple of advanced tools to facilitate research, education, business and knowledge management both on European and sub-country level. The project with its multilingual solutions serves the global audience with a 21-century research infrastructure. The instruments and techniques contribute and help to discover impacts, associations, causes and effects of socio-economic development, environmental issues and governance. With its wide scope the project provides a solid background to describe possible future alternatives. The innovation of World in Figures can be described as follows:

- It provides a wide range of data about different fields of life. The specific aspect is that thousands of variables describing totally different phenomena are in a single database, which allows a common, multivariate analyses.
- Traditional statistics as well as other data, like sociological survey results, ecological data, etc. are easily accessible via the Internet. Statistical data and survey result figures focus on society, economy, politics, culture, human values, and the natural environment etc. The scale of stored statistical variables is not limited.
- Apart from the data access, it offers data analyzing options. The online statistical data analyzer has basic and advanced modules, containing data presentation and visualization modules, forecasting functions, download options, etc.
- and also the chance for researchers to employ advanced methods, e.g. to discover hidden associations between variables or forecasting economic, social or environmental trends.
- Data access is free, except for some research and business statistics. Moreover: the user can find the latest statistics and time series and the methodological descriptions of the variables.

How World in Figures contributes to the societal developments and to the futures progressive efforts?

- It helps researchers and analysts in the field of science, education or business, with a new and effective online service, which facilitates the understanding of human life, accelerates the discovery of hidden causal relationships etc. The system facilitates research processes and forecasting as well as helps decision makers and politicians. The system contributes to the deepening of the integration processes, too.

- With the publicity of information it leverages democracy and helps societal control, education, and participation of the citizens in self-governance. The project supports civil communities to get acquainted with their socio-economic and natural environment and the scope of their actions.
- The usage of the tools is cost-effective. Services for the citizens are free, and research processes will be significantly faster. The objective for the future, in general, is the development and the long-term management of a European level knowledge-base of all the accessible and valuable statistical data, survey results and other information expressed in numbers via the Internet.

Integrating Respondents' And Questions' Characteristics Within A Single Item Nonresponse Model – Possible Approaches

Andrej Kveder (CATI, Ljubljana, Slovenia)

Integration of different aspects of the item nonresponse within a single model, although intuitively simple, has proved to be a challenging task. The main emphasis of this paper is the discussion of some possible analytical approaches, their strengths and weaknesses.

The identification of the nonrespondents has been one of the main driving forces of item-nonresponse analysis. Advancements in data adjustment techniques require and can handle more detailed information about the item-nonrespondents for the prediction of missing data. The majority of the analyses are concerned with the identification of the nonrespondents and thus their socio-demographic characteristics, since they are available to the researcher in every household survey. However, the socio-demographic characteristics have neither the explanatory power nor sufficient substantive justification to explain why item-nonresponse occurs. The field of item-nonresponse research is developing along two major lines: aiding the adjustment techniques with additional information and the reduction of item-nonresponse in the phase of survey design. Analytical approaches to studying item-nonresponse try to aid both: the results of more complex models could function as a better leverage for data imputation, on one hand, and as a new insight in understanding the circumstances in which the item-nonresponse occurs, on the other hand. This paper discusses some analytical approaches that combine respondents' and questions' characteristics within a single analytical model.

The analyses were performed on the Fertility and Family Survey data including 18 different countries and 116.897 cases. For the analysis, a meta-database of 423 questions was constructed. Every question was described according to its context, data type, and topic. Exploratory models were run predicting the item-response using combined power of nonrespondents' and questions' characteristics. Several different approaches were tested and evaluated.

The integration of different aspects of item-nonresponse within a single model offers a possibility of evaluating the weight of nonrespondents' characteristics against other explanatory factors like the typology and topic of each question. The results presented in this paper offer a critical overview of some possible types of analysis and the implications of their use.

Exploring Different Methods for Measuring Formal and Informal Social Networks in Knowledge Organisations

Danielle De Lange, Filip Agneessens, and Hans Waege
(University of Gent, Belgium)

Different types of social relations have been found to have an important influence on the performance of employees in organisations. This paper focuses on a comparison of different methods for acquiring information on advice, cooperation, friendship, adversarial and superficial networks in knowledge organisations.
We investigate the applicability of three distinct measurement methods to acquire these network data. Firstly, employees were presented a short description of a specific situation in which social relations with their colleagues might play a significant role. They had to indicate if (or how often) this specific situation occurred with each of the colleagues. Secondly, respondents were asked to indicate whether a specific relational concept (in this case advice or friendship') applied to each of their relations with their colleagues. Thirdly, we provided respondents with four semantic differentials (e.g. distrust-trust) on which they needed to position their relationship with the other employees.
Whether these different measurement instruments capture distinct aspects of the relationship between employees, or whether they measure the same underlying concepts, is one of the major concerns of this paper. The aim of this paper is twofold. First of all, we want to know to what extent these different measurement instruments (situations, concepts and opposite adjectives) overlap. Secondly, we would like to find out to what degree these different methods as a whole give us conceptually different and complementary information. To the extent that items are correlated within one method and between methods we need to investigate which of these different methods is best suited for our content related purposes. The criteria used for selecting the most appropriate method are minimal item non-response - i.e. from the viewpoint of measuring complete networks - and maximum relational diversity with a minimum of questions.

Rule Learning for Subgroup Discovery
Nada Lavrač (J. Stefan Institute, Ljubljana, Slovenia)

Rule learning is typically used in solving classification and prediction tasks. However, learning of classification rules can be adapted also to subgroup discovery. The paper presents a methodology for subgroup discovery and an analysis of algorithms adapting classification rule learning to subgroup discovery.

On the Normalization of Contingency Indexes based on the χ^2 Statistic
Antonio Mango (University of Napoli, Italy)

Any statistical textbook and any statistical package for Social Sciences gives, as a normalized contingency index based on the χ^2 statistic, the $Cramèr$ or the $Tschuprov$ solution when the contingency table is not squared or it does not have couples of marginal row and column frequencies with equal coordinates.
Both solutions are biased because they systematically give underestimated values of this coefficient, as it will be shown in the paper, and refer to tables that differ for the real dimensions and for the effective marginal frequencies of the table.
The following tables may offer an idea of the bias the solutions are affected by:

	A	B	Σ
a	50	10	60
b	20	20	40
Σ	70	30	100

$Table\ 1$

	A	B	C	Σ
a	30	20	10	60
b	7	13	20	40
Σ	37	33	30	100

$Table\ 2$

	A	B	C	D	Σ
a	30	20	7	3	60
b	7	13	3	17	40
Σ	37	33	10	20	100

$Table\ 3$

(1)

if we apply to them the χ^2 function, we have, respectively:

$$\chi_1^2 = 12.70, \ \ \chi_2^2 = 15.75 \ \text{ and } \ \chi_3^2 = 18.40 \tag{2}$$

according to $Cramèr$, $V^2 = \frac{\chi^2}{N\sqrt{q-1}}$, where $q = min\,[r, c]$, being r and c respectively the number of rows and columns of the contingency table, we have:

$$V_1^2 = .127, \ \ \mathrm{V}_2^2 = .157 \ \text{ and } \ \mathrm{V}_3^2 = .184 \tag{3}$$

and according to $Tschuprov$, $T^2 = \frac{\chi^2}{N\sqrt{(r-1)(c-1)}}$, we have:

$$T_1^2 = .127, \ \ \mathrm{T}_2^2 = .111 \ \text{ and } \ \mathrm{T}_3^2 = .106 \tag{4}$$

The expressions which appear at the denominators of V^2 and T^2 represent those theoretical maxima for χ^2 and the values they give are equal in this case for coincidence of degrees of freedom .
On the basis of a different idea we maintain that maxima are obtained for tables presenting the highest level of frequency concentration in the cells which is plausible with the marginal frequencies.
Such hypothetical tables, corresponding to the previous ones, are reported here:

	A	B	Σ
a	60	0	60
b	10	30	40
Σ	70	30	100

$Table\ 1bis$

	A	B	C	Σ
a	37	23	0	60
b	0	10	30	40
Σ	37	33	30	100

$Table\ 2bis$

	A	B	C	D	Σ
a	37	23	0	0	60
b	0	10	10	20	40
Σ	37	33	10	20	100

$Table\ 3bis$

by means of the above tables we may calculate values of χ^2 , $\chi^2_{\max}$, which will be used as normalizing values for χ^2 as referred to 1, we obtain:

$$\chi^2_{\max 1} = 64.29, \ \ \chi^2_{\max 2} = 77.64 \ \text{ and } \ \chi^2_{\max 3} = 70.96, \tag{5}$$

which are the normalized contingency coefficients we propose, $A^2 = \frac{\chi^2}{\chi^2_{\max}}$, we have:

$$A^2_1 = .198, \ \ \mathrm{A}^2_2 = .203 \ \text{ and } \ \mathrm{A}^2_3 = .259. \tag{6}$$

As we can see, $Cramèr$ and $Tschuprov$ indexes assume smaller values to those of the proposed index as for them have been choose, as units, overestimated maxima of χ^2 which, furthermore, refer to theoretical tables whose structure (dimension and marginal frequencies) cannot be compared with the one under analysis and, of course, do not belong to the $Class\ of\ Frechét$ of this table. The paper deals with the construction of a table maximizing χ^2 as applied to a table sorted from among the elements of this $Class$, both manually and by a programmed software.

Measuring the Effect of the Chronological Order of Interviews on the Probability of Passing Filter (Screening) Variables

Herbert Matschinger (Clinic and Policlinic for Psychiatry, Leipzig, Austria)

The following analysis is concerned to present a strategy and statistical models to describe and partially explain a particular behaviour of the interviewers when processing important filter variables in a survey interview. In the following we will call the effect produced by this behaviour - which will result in considerable artefacts - a *sequence effect*. Since we must assume that interviewers learn on how to shorten an interview considerably, we will focus on the order of interviews 'within' a particular interviewer, which will be suspected to affect the probability of passing a variable which serves as a filter. The analysis is carried out in two main steps, since it is assumed that the artefacts described below are produced by at least two different sources, one of which not directly observable. In the 1st step the effect of the chronological order of interviewers *within* each interviewer is analysed by means of 2-level logistic regression. Here the amount of interviews and other characteristics of the interviewer may serve as additional exogenous characteristics. In the 2nd step we attempt to model the potential latent heterogeneity of the population of interviewers with respect to both the effects mentioned above. The segmentation of the population of interviewers into optimally homogenous subgroups may serve as an additional exogenous variable in order to describe different types of interviewer behaviour.

The model adopted is a random coefficient logistic regression model (Bryk and Raudenbush, 1992; Ditton, 1998; Goldstein, 1995; Kreft and De Leeuw, 1998; Snijders and Bosker, 1999). The letter Y represents the dichotomous filter variable as mentioned in the introduction. The 1st or individual level is formed by each single interview. The 2nd level comprises the interviewers. The sequence variable (called SC) is observed at level 1, the amount of interviews (called IA) is observed at level 2. Other important characteristic of the interviewer are summarised by IX. The model reads as follows:

Level-1 Model

$$\mathrm{Prob}(Y = 1|B) = P$$

$$\log[P/(1-P)] = B_0 + B_1 * (SC)$$

Level-2 Model

$$B_0 = G_{00} + G_{01} * (IA) + G_{02} * (IX) + U_0$$

$$B_1 = G_{10} + G_{11} * (IA) + G_{12} * (IX) + U_1$$

Level-1

$$\text{variance} = sigma\_squared/[P(1-P)]$$

Since we have to assume that part of the variance of B_1 (τ_{11}) depends on B_0 (the intercept) the vector B_0 is adopted as a latent variable and will be included in the level-2 equation for B_1. Estimating both equations for B_1 provides information about the effect of IA and IX on the sequence effect controlling for the probability of the dependent variable at the first interview. The equation reads as follows:
Latent Variable Regression in equation format:

$$B_1 = G^*_{10} + G^*_{11}(IA) + G^*_{12}(IX) + G^*_{13}(B_0) + U^*_1$$

The very results for the 2-level models (particularly with respect to highly significant variances of U_1) finally leads into the attempt to identify homogenous classes of interviewers. This will be done by means of logistic regression within latent classes as has been proposed by Vermunt and other authors (Hagenaars and McCutcheon, 2002; Vermunt, 1997; Vermunt and Magidson, 1999; Vermunt and Magidson, 2000; Wedel and DeSarbo, 1994). Since the number of classes is unknown in advance, several solution have to estimated in order to gain parsimony. Log-Likelihood and BIC are employed for model selection. We estimated 6 models and choose the 4-class model due to the lowest BIC. Employing the allocation of the interviewers to each of these latent classes and estimating the interaction effect between SC and the latent class variable it can be shown by means of a random effects logit model that the conditional intraclasscorrelation drops from 0.28 to almost 0. This further more confirms the conjecture that the sequence behaviour of the interviewers can be described in a very parsimonious manner.

References

1. Bryk, A.S. and Raudenbush, S.W. (1992): *Hierarchical Linear Models; Applications and data Analysis Methods.* Newbury Park: Sage Publications.
2. Ditton, H. (1998): *Mehrebenenanalyse. Grundlagen und Anwendungen des Hierarchisch Linearen Modells*. Weinheim, Mnchen: Juventa.
3. Goldstein, H. (1995): *Multilevel Statistical Models*. London, Sydney, Auckland: Arnold.
4. Hagenaars, J.A. and McCutcheon, A. (2002): *Advanced Latent Class Analysis*. Cambridge: Cambridge University Press.
5. Kreft,I. and De Leeuw,J. (1998): *Introducing Multilevel Modelling*. London, Thousand Oaks, New Delhi: SAGE Publishers.
6. Snijders, T. and Bosker,R. (1999): *Multilevel Analysis. An Introduction to Basic and Advanced Multilevel Modeling*. London. Thousand Oaks, New Delhi: SAGE Publications.

7. Vermunt, J.K. (1997): *Log-linear Models for Event Histories*. Thousand Oaks, London, New Delhi: Sage Publications.
8. Vermunt,J.K. and Magidson,J. (1999): Exploratory Latent Class Cluster, Factor, and Regression Analysis: the Latent Gold Approach. In *Anonymous, Proceedings EMPS'99 Conference*. Lueneburg.
9. Vermunt,J.K. and Magidson,J. (2000): *Latent GOLD Users Guide*. Belmont MA: Statistical Innovations Inc.
10. Wedel, M. and DeSarbo, W.S. (1994): A Review of Recent Developments in Latent Class Regression Models. In R. P. Bagozzi (Ed.), *Advanced Methods of Marketing Research*. (352-388). Cambridge: Blackwell Publishers.

Statistical Analysis of Road Accidents in Slovenia in the Recent Years

Elvir Mujkič (Zagorje, Slovenia)
Jože Rovan (University of Ljubljana, Slovenia)

Slovenia is facing similar traffic problems as many other countries in the world. A new traffic law, compliant with the European Union Guidelines, was adopted on May 1st, 1998. The main aim of the new law was to improve road safety by means of setting lower speed limits, by stressing compulsory use of safety equipment and by reducing the number of alcoholised drivers.

The goal of this article was to find out whether the new law had contributed to safer and more effective transport system in Slovenia. We found out that new law have some positive impacts. Aggravation of penalty politics, that concentrated in particular to new drivers with no experience, proved to be the right one. We found out that new drivers did not step out as much as they did in the past, but there is still much to be done. When we had analysed the major goals of the law we found out that today people are using safety belts and other safety equipment more regularly. With alcohol some progress is shown due to smaller percentage of alcoholised provokers. Nevertheless authorities would have to focus on this problem in future and introduce some specific measures (prevention measures, imprisonment of alcoholised drivers). Alcohol problem is a cultural problem so no major improvements can be expected in short term.

Data has shown that advanced technology in automobiles has had the highest impact on decreasing the number of fatalities on the roads. We can also give some credit to the new law and policemen actions. Various preventative actions proved to be positive, since they forced drivers to prepare for difficult conditions, which they might face when driving.

The comparison of the situation in Slovenia with some other European countries has revealed that the road safety in Slovenia is far behind the level of European Union. Slovenia is in a very delicate situation. On one hand the number of automobiles is steadily growing while on the other hand we face large problems with traffic safety and law obedience. We will have to improve legal system, especially fasten penalty proceedings. Finally we will have to make our whole system work if we want to close the gap between Slovenia and European Union countries.

Cognitive Evaluation of the Hierarchical Approach for Measuring Ego-Centered Social Networks

Jana Nadoh, Petra Podreberšek, and Valentina Hlebec
(University of Ljubljana Slovenia)

The paper is exploring how respondents understand and interpret Antonucci's hierarchical approach (Antonucci, 1986) for measuring ego-centered social networks. Think aloud procedure was used to assess how respondents (egos) differentiated among people (alters) that are named in hierarchical circles and what criteria they used to select people from their global social network.

In Antonucci's approach the emotional criterion is used for selecting alters from the respondent's global network and placing them into three hierarchical circles that are graphically presented to the respondent. The respondent (ego) is in the center of the three circles. The more interior the circle is the closer and more important people (alters) within it are. The described procedure very often is used in research on ego-centered support networks. Only one 'name generator' is used for data collection, which is more convenient to the respondents and cheaper for the researcher.

Data were collected by face-to-face data collection mode on a quota sample of 84 respondents known to interviewers (gender and age of respondents were fixed). Respondents first answered questions about their ego-centered social networks. The think aloud procedure was used at the end of the interview to assess the interpretation of the emotional criterion used to elicit the members of ego-centered social network. For each circle the respondents were asked to describe what came to their mind when eliciting the names and to name the criteria used to select these people. Lastly the respondents were asked to describe the differences between people in the first, second and the third circle in their own words. Data collection and coding was done by a group of students attending Social network analysis and Questionnaire design courses in 2001/2002.

Preliminary analyses show that the respondents most often focused on alters themselves (for example 'I thought of - my family, my best friend, ...'), quality of ties (for example 'I cannot imagine my life without them.' or 'These are people that are close to me.') or formal characteristics of ties (for example 'These are people I have regular contacts with.') regardless of the circle. When focusing on the differences among people within circles respondents gave three types of answers. Some focused on hierarchical differences between people (for example 'In the first circle is my family, in the second are good friends and in the third are acquaintances.'), others focused on various characteristics of ties (for example 'In the first circle are people that are very close to me,' or 'I see most often,' or 'I have regular contacts, ...'). Some respondents used criteria at the same time. Some respondents compared the hierarchy of circles' characteristics (for example'"I distinguish between

the circles based on frequency of contacts, or closeness.'). Analyses also show that there are small differences in interpretation across gender and age of respondents.

Imputation Procedures and the Quality of Income Information in the ECHP
Cheti Nicoletti (University of Essex, United Kingdom)
Franco Peracchi (University of Rome 'Tor Vergata', Italy)

The aim of the paper is to evaluate the impact of the imputation procedures adopted in the European Community Household Panel (ECHP) on the quality of information about income variables. We focus attention on personal earnings (wages and salaries and self employment earnings), total net household income, and the incidence and dynamics of poverty.

Different types of nonresponse may affect the analysis of income variables, hence we give a detailed classification of the types of nonresponse and the methods adopted in the ECHP to correct for them.

We evaluate the imputation methods adopted by the ECHP by looking for systematic differences in the distribution of income across different types of responding units. More precisely, we compare descriptive statistics (mean, median and percentiles) of the imputed income variables (both in levels and growth rates), for the respondents and for different types of non respondents. We also repeat the same type of analysis using mean and quantile regressions in order to control for a broad set of covariates (sex, education, marital status, job duration, wave dummies, etc.). While income levels do not seem to be affected much by imputation, income dynamics is. This occurs because the imputation procedure seems to alter the tails of the distribution of income growth. The effects of imputation are reduced if we consider statistics that are robust to outliers, such as the median.

A different approach to evaluate the impact of the missing data problem on poverty can be adopted following the work of Manski (1989) and Horowitz and Manski (1998). These papers show how to derive bounds for the cumulative distribution function of a variable of interest without imposing any assumption on the missing data mechanism.

Net household income in the ECHP is affected by nonresponse in about 30Such a high nonresponse rate implies that Manskis bounds tend to be wide. In many cases, however, the information on income is not completely absent because income may be reported partially, i.e. we may know that total net household income is above a known threshold. This information may be sufficient to identify poor people. In fact, if household income is above the poverty line, then we can classify the members of a household as non poor. Further, our ability of classifying people as non poor increases as the poverty line is reduced. This lowers the nonresponse rate by a big amount, narrowing Manskis bounds. Our aim is then to evaluate if, for a suitable choice of the poverty line, it is better to combine the information from the fully respondents and the partially respondents and avoid using the imputed values.

Investigating Acquisition Patterns of Financial Product using Mokken Scale Analysis

Leo Paas (Tilburg University, The Netherlands)

The use of Social Science Research techniques in Economic Research has increased substantially over the last few years. For example, Latent Class Analysis, a technique that was originally developed in the Social Sciences is now one of the most commonly used statistical techniques in the Economic discipline of Marketing. Despite the start that has been made for cross-pollination many opportunities are still neglected.

The current paper is on a particularly useful application of Social Science Research techniques for contemporary marketing. Recently many companies are aiming to develop and maintain long-term relationships with their clients (i.e. relationship marketing), the idea being that it is more profitable to maintain and extend the relationship with existing clients than to obtain new clients. Now this approach to marketing implies that companies require insight into the developments of their customers. Most contemporary marketing research techniques, however, are aimed at predicting and explaining one of many events that occur in relationships between clients and companies. However, a single event cannot be used to gain insight into all developments that consumers go through, which is required for relationship marketing purposes.

In this contribution it is proposed that scale analysis can be used in the financial services sector, for the purpose of obtaining insight into consumer developments. That is, ownership of products can be coded as a binary item. If there are n subjects and k products, then such coding leads to an $n * k$ table of binary items. If element $(n, k) = 1$ then subject n owns item k and 0 implies that n does not own k. Scale analysis for binary items can be applied to this $n * k$ table and can be used to: "Position both services and households along a 'latent' difficulty/ability dimension. Thus, more 'difficult' services (those that require greater resources, are more complex and risky, and have lower liquidity) would require higher levels of investors 'ability' or 'maturity' (Kamakura et al., 1991).

If such a hierarchy can be found then one may assume that consumers first acquire products that are owned by a larger percentage of subjects, before acquiring less commonly owned products (Kamakura et al. 1991). Previously various scaling models have been used for investigating acquisition patterns, such as Guttman scaling, the 1- and the 2-parameter logistic model. However, in Paas (1998) and Paas and Molenaar (2002), it is shown that the mathematical properties of a scaling model known as Mokken scale analysis (Mokken, 1971; Sijtsma and Molenaar, 2002) are most consistent with the mathematical characteristics of acquisition patterns. Mokken scale analysis is suitable for ordering subjects and items (or products for our investigation) on an underlying latent trait and imposes no additional

criteria. These forms of ordering are precisely what acquisition pattern analysis requires, this mathematical consistency is discussed in Paas and Molenaar (2002).

The current study concentrates on the empirical validation of the use of survey data and of so-called transactional data for conducting acquisition pattern analysis. We use two sets of survey data and one set of transactional data. The first set of survey data was collected by CentER (Tilburg University) in 1993. The data are representative for consumers in The Netherlands aged 18 years or older. The CentER-panel is an Internet-based Telepanel. Every week, the panel members fill in a questionnaire on the Internet, while being at home. The analysed dataset contains the answers of 1763 subjects on questions regarding the financial products that they own. Questions were on ownership of the following products: (1) Checking Account, (2) Savings Account, (3) Lumpsum Policy, (4) Investment Trust and (5) Securities. The second set of survey data (EURODATA) was collected in 1990 for the Readers Digest Consumer Survey, through a survey by mail. EURODATA contains responses of 22,339 subjects from 17 European countries. All interviewed subjects were 18 years or older and living in private households (i.e. subjects living in special housing for the elderly, hospitals, prisons, etc. were excluded from the survey). Regarding the questionnaire, we may mention that in each country all interviewed subjects indicated whether their household owns the following products: (1) Bank Account, (2) Chequebook, (3) Card for Obtaining Money from Cash Dispensers, (4) Credit Card that Immediately Takes Money out of the Owners Account, (5) Credit Card Requiring the Debt to be Paid in Full when the Owner is Billed and (6) Credit Card Requiring the Debt to be Paid over a Number of Months.

The transactional data were taken from the marketing database of a large financial services provider in The Netherlands (financial services provider X) and contain information on the products owned by a selection of 6490 clients of which each client has purchased each product. Transactional data are recorded as a result of product acquisitions performed by clients of a bank. Ownership and dates of purchase with regard to each product offered by banks, are stored in large transactional databases. The data we analyse were collected in 1997 and contain information of ownership and purchase dates of the following products: (1) Checking Account, (2) Savings Account, (3) Investment Trust and (4) Securities.

Now in the paper we test whether the three following sources lead to consistent results: (1) The ownership patterns of products found in survey data; (2) The ownership patterns of products found in the transactional data of financial services provider X; (3) The actual dates at which subjects acquired financial products at of financial services provider X. We find that these sources of data lead to consistent results in terms of acquisition patterns. In the first set of survey data we find that the following four products are acquired in the mentioned order: (1) Checking Account, (2) Savings Account, (3) Investment Trust and (4) Securities (the Lumpsum Policy is not acquired in a hierarchical order with the other products in the set).

This order is also found after applying Mokken scale analysis to the transactional data and is consistent with the purchase dates in the transactional data. Moreover, acquisition patterns can also be found for other financial products. Analysis of the second set of survey data shows that most subjects in 17 European countries acquire the following six products in the mentioned order: (1) Bank Account, (2) Chequebook, (3) Card for Obtaining Money from Cash Dispensers, (4) Credit Card that Immediately Takes Money out of the Owners Account, (5) Credit Card Requiring the Debt to be Paid in Full when the Owner is Billed and (6) Credit Card Requiring the Debt to be Paid over a Number of Months.
These results are consistent with the mathematical proofs presented by Paas and Molenaar (2002), suggesting that Mokken scale analysis is suitable for conducting acquisition pattern analysis. Moreover, the found consistency also suggests that this proof applies to both survey and transactional data. After presenting the research the paper is rounded-off with conclusions and comments on applications and implications for marketing (research).

References

1. Kamakura, W.A., Ramaswami, S.N., and Srivastava, R.K. (1991): Applying latent trait analysis in the evaluation of prospects for cross-selling of financial services. *International Journal of Research in Marketing*, **8**, 329-349.
2. Mokken, R.J. (1971): *A Theory and Procedure of Scale Analysis*. Reading: Mouton.
3. Paas, L.J. (1998): Mokken scaling characteristic sets and acquisition patterns of durable and financial products. *Journal of Economic Psychology*, **19**, 353-376.
4. Paas, L.J. and Molenaar, I.W. (2002): On the use of scaling models for deducing longitudinal sequences from cross-sectional data. Submitted for publication.
5. Sijtsma, K. and Molenaar, I.W. (2002): *Introduction to Nonparametric Item Response Theory*. Thousand Oaks: Sage Publications.

Predictive Validity of High School Grade and other Characteristics on the Graduate Student University Career

Laura Pagani (University of Udine, Italy)
Chiara Seghieri (University of Florence, Italy)

The aim of this paper is to highlight the capacity of high school grade and other students characteristics (such as socio-economic status, high school of origin and so on) to predict success in the university career using multilevel models together with ROC analysis.
In fact a valid measure of predicting success in college can be used to determine students eligibility for admission (admission policy) or to guide students in their faculty choice.
To study the capacity of school grade and other students characteristics to predict success in the university career we will use data based on the records of almost 3200 freshmen who entered the University of Udine during the academic year 1992/93-1997/98.
The structure of our data-base suggest the use of the multilevel models to relate a response variable (a binary one that takes value 1 if the student gives at least one or two, or three, or four exams during the first year of his university career, 0 otherwise) to independent variables. We consider here two level: the lower level units are the students and the higher level units are their high school of origin.
After estimating the parameters of the four models we use the estimated probability to construct an evaluation test used to discriminate between two states: students that can successfully enter the university and student with lower capacities. The shape of the ROC curves and ROC plots help to quantify the accuracy of this evaluation test.

Partial Error - Free Polynomial Regression

Tibor K. Pogány, Vinko Tomas, and Mato Tudor
(University of Rijeka, Croatia)

In the papers (1,2) the polynomial regression is discussed and solved in the error - free sense when the 2D sample $U := \{(x_j, y_j) : 1 \leq j \leq n\}$ is known. Namely, assuming that

$$y = P_{m-1}(x) = p_0 + \cdots + p_{m-1}x^{m-1}, \qquad m \leq n, \tag{1}$$

we are looking for the unknown parameter array $p := (p_0, \cdots, p_{m-1})$ under the constraint

$$\sum_{k=0}^{k=m-1} \overline{\mathbf{x}_l^k} p_k = \overline{\mathbf{y}_l} \qquad 1 \leq l \leq m, \tag{2}$$

where $\overline{\mathbf{x}_l^k} := \frac{1}{\#J_l}\sum_{j\in J_l} x_j^k$, $\overline{\mathbf{y}_l} := \frac{1}{\#J_l}\sum_{j\in J_l} y_j$ and $J^\star := \{J_1, \cdots, J_m\}$ is an admissible decomposition of U (the system (2) possesses unique solution). Endly, averaging all p_j's from the set of all $J^\star$ we get the optimal *error - free* $p_j^\star$, $0 \leq j \leq m-1$, see (1, Theorem 2.1, Theorem 2.2, §3).

In the present paper our approach to the same problem is different somewhat. Assume that we know already the regression polynomial degree $m-1$, say. Then let $U_l^{(m)}$ one of the admissible subsamples of U, which size equals m, $N_U \leq \binom{n}{m}$. Consider the system in $\mathbf{p}$:

$$\sum_{k=1}^{k=m-1} p_k x_j^k = y_j, \qquad \forall (x_j, y_j) \in U_l^{(m)}. \tag{3}$$

Since for all different x's in U the system determinant in (3) is of Vandermonde type, it is an unique solution $\mathbf{p}_l := (p_0^{(l)}, \cdots, p_{m-1}^{(l)})$, say. Denote

$$\widehat{\mathbf{p}} = (\widehat{p}_0, \cdots, \widehat{p}_{m-1}) := \left(\frac{1}{N_U}\sum_{l=1}^{N_U} p_0^{(l)}, \cdots, \frac{1}{N_U}\sum_{l=1}^{N_U} p_{m-1}^{(l)}\right),$$

where N_U is the cardinality of the set of all m - sized admissible subsamples of U. All this results in

$$P_{m-1}(x) \approx \widehat{p}_0 + \cdots + \widehat{p}_{m-1}x^{m-1},$$

where the approximation is in the *partial error - free sense* used, having on mind the constraints (3).

Endly, further regression coefficient computation methods will be considered and concrete engineering problems will be discussed. Comparations with the Minimal Least Squares regression polynomial and with the error - free polynomial approach is realized, compare (2). The statistical background of the method is investigated too.

1. Pogány, T. & Tudor, M. (1993): On the error-free polynomial regression model **I.**, *Proceedings of the* **KOI'93**, Rovinj, 1993, (L. Neralić *et al. eds.*), Hrvatsko društvo za operacijska istraživanja, Zagreb, 1993, 207-216.

2. Pogány, T. (1993): On the error-free polynomial regression model **II.** (Least squares estimation under error-free condition), *Proceedings of the* **KOI'93**, Rovinj, 1993, (L. Neralić *et al. eds.*), Hrvatsko društvo za operacijska istraživanja, Zagreb, 1993, 201-206.

Measuring Gender Inequality in Income Distribution across four Industrialized Countries Using the Dagum Model

Claudio Quintano and Antonella D'Agostino (University of Napoli, Italy)

Generally gender equality means that women and men have equal conditions for realizing their full social rights and for contributing to political, economic, social and cultural development of a country. By contrast, many empirical analysis showed that this condition is not verified and generally women are disadvantaged respect to men. In fact, usually, women have lower incomes than men in almost all societies and the most heavy outcome of this evidence is that it involves higher poverty rates for women, applied analysis on self-employment confirm that self-employment is still predominated by men even if it is observed a rise of female entrepreneurship and it seems that part-time employment remains a 'womens work'.

In this paper, a particular aspect of such discrimination problem is analysed, that is the incidence of gender inequality in personal income distribution. Our unit of analysis are therefore individuals who constitute households by their own, whose incomes are not subject to any compensation effects due to the presence of other income recipients in the same household. From a methodological point of view, income distribution is modelled by a parametric approach. In fact, in recent years there has been a growing interest in the exploration of parametric models of the distribution of income, and different theoretical representations of income distribution have been used. Dagum (1977) proposed a parametric specification that proved to render a better goodness-of-fit relative to alternative existing models. In particular, there are three versions of the Dagum model, each accounting for specific assumptions about the population of income receivers. In this paper we refer to Dagum Type I distribution that contains three parameters and describes distributions starting with income recipients having positive income. In order to allow the form of the income distribution to vary with gender and other interesting characteristics, heterogeneity has been introduced in each model parameter. In other words, the conditional income distribution, i.e. the income distribution conditional on personal characteristics, is modelled directly.

It is important to outline that the covariates parameters have no interpretation in themselves; for this reason the inference about the influence of individuals characteristics can be based on differentials in some synthesis measures of the estimated income distribution as mode, median, etc.

Once the model parameters has been estimated the idea is to consider individuals with a set of benchmark characteristics and to investigate whether chancing these characteristics from to leads to an increase or a decrease in such synthesis measures.

The empirical analysis aims to measure and to compare gender inequality in Europes four biggest economies: France, Germany, Italy, and UK. Because of the

variation in the generosity of their welfare systems and the working of their labour markets, a comparison about the proposed theme, in these four countries, is particularly interesting in a perspective of the event of economic integration in Europe even if the causes of the gender inequality phenomenon are complex and deeply integrated with the degree of economic, civil and social development. In this context, Germany is regarded as a well-developed welfare state with a highly regulated labour market; Italy and France may be considered a conservative welfare state in opposite to UK that is characterized by a liberal welfare state.

The main results we obtained is that the introduction of individual heterogeneity seems to play an important role for understanding the income distribution in each country. First of all, gender can be consider a cause of income inequality, this means that male and female do not have equal conditions in term of income levels. Besides this, a gender effect there is in each country even if other observed factors have been taken into account.

The data used for the empirical comparative analysis of gender inequality derive from Wave 4 (1997) of the European Household Panel Survey (ECHP). The ECHP represents the first European panel survey characterized by a common ex-ante structure. This survey provides comparable information for many socio-economic variables, including income, at European level.

References

1. Dagum, C. (1977): A New Model for Personal Income Distribution: Specification and Estimation. *Economie Appliquee*, **3**, 413-436.

Open vs. Closed Questions in Web Surveys

Urša Reja, Katja Lozar Manfreda, Valentina Hlebec, and Vasja Vehovar
(University of Ljubljana, Slovenia)

Two quite different reasons for using open as opposed to closed questions can be distinguished. One is to discover the responses that individuals give spontaneously, the other is to avoid the bias that may result from suggesting responses to individuals. However, open questions have also their disadvantages in comparison to closed questions, such as need for extensive coding and larger item non-response. While this issue has already been well researched for traditional survey questionnaires, not much research has been devoted to it in recently used Web questionnaires. We therefore examine the differences between the two forms of survey questions by means of experiments within large-scale RIS 2001 Web survey.

Three survey questions were asked in two forms in a split-ballot experiment: question on most frequently visited Slovenian and foreign Web sites, e-shops where respondents made some purchase, and the most important, critical problem the Internet is facing today.

The results show that in all cases there were differences between question forms in univariate distributions, though no differences were found in the ranking of values. Closed questions in general yield higher percentages than open questions for the answers that are identical at both question forms. It seems that respondents restricted themselves with apparent ease to the alternatives offered on the closed form, whereas respondents on the open question produced a much more diverse set of answers. In addition, our results suggest that open questions produce more missing data than closed questions. Even more, there were more inadequate answers for open questions (e.g. respondents mention Slovenian Web pages on question on foreign Web pages or it is impossible to determine what the answer is about). This suggests that open questions should be more directive in wording (at least for Web surveys as self administered mode of data collection) as closed questions that are more specified with given response alternatives. We also found out that when respondents were asked about a more salient topic the differences between open and closed form were smaller. In this case both forms of questions were equivalent.

Empirical Comparison of Measures of Polarization: A Monte Carlo Approach
María Elena García Reyes (University of York, UK)

Recent publications have tried to acknowledge the notion of polarization, but little has been done on the empirical side. Concerning theory, there is disagreement about how to measure polarization processes. The main debate is whether to use specific measures of polarization or to use inequality measures. In this paper we attempt to use a complete set of measures including those devoted exclusively to measure polarization and those that measure inequality. We contribute to clarify this debate by comparing all the available measures of polarization and some inequality measures. Our claim is that measures of inequality may capture the main characteristics of polarization if those indices are decomposable and if we are able to create disjoint income groups. We apply this set of measures to pseudo populations.

The application of this set corresponds to a controlled environment that uses Monte Carlo experiments. Using pseudo populations enable us to emphasize polarization processes obtaining a clear picture of where the estimators are standing. In other words, Monte Carlo experiments allows us to compare the whole set of existing measures that attempt to estimate polarization. This will bring some light on the prevailing debate.

Any estimator that claims to measure polarization should capture both characteristics, intragroup homogeneity and intergroup heterogeneity. Therefore, the first set of measures used to estimate polarization is a set of decomposable measures of inequality, and those measures based on income subgroups. Measures that group the population around the mean and the median, the Decomposed Gini Index, the Maximum Equalization Percentage, and the Theil Index.

We also measure polarization using measures that have been used exclusively for this purpose such as the Measure derived by Esteban and Ray the measure derived by Wolfson, and from the later a new class of indices derived by Tsui and Wang.

We discuss a set of possible pseudo populations that will resemble different scenarios where polarization and inequality arise. We present the comparison of all the possible measures of polarization and suggest certain ranking among those measures; and the statistical evidence that supports the mentioned rank among measures.

Alternative Approaches to Measure Web Site Audiences
Matjaž Robinšak and Zenel Batagelj (CATI, Ljubljana, Slovenia)

The topic of this presentation is web site audiences measurement. In the markets where web marketing and advertising budgets spent on the Net are large enough, methods similar to peoplemeters in case of TV measurements are usually used. IN such cases Internet users - panelists - have software applications installed on their computers that are monitoring Internet traffic from their computers. The data are sent to servers located at research organizations in real-time and so also analyzed.
Alternative methods can also be used: (1) traditional surveys and (2) technological, mostly server based measurements. Both of them can be (and are widely) criticized. The reasons behind that are mostly low understanding of survey research and too high expectations that measurements are actually not needed since the Internet is highly track able.
There are also known limitations of these 'traditional' methods. Survey based methods have limitations because they are based on respondents memory what results that web pages with higher recognition (known brands) get higher scores, while non-branded or new sites get lower scores. On the other side, technical - server based methods have problems of users identifications, what is usually connected with cookie problems, multiple users on the same computer and firewall, proxy problems.
In not so rich markets (small countries and countries with low purchasing power) there is not enough money on the market to run these sophisticated panel surveys. Therefore 'traditional' methods, generally not well accepted by the Internet industry should be used. There is an additional phenomena present in Slovenia, web sites are mostly run by enthusiasts, mostly people from IT industry that even less familiar with survey research methods, media research methods in general, and therefore very suspicious.
In April 2002 an alternative approach of measuring web site audiences was used for the first time. It combined technical measurement using so called "pixels", web surveys to get demographical and other respondents data, and parallel telephone control survey. All three methods of measurement were used together for the final results: 14-days reach estimates, demographical profiles of web pages evaluated and overlap among pages.
The results are available in software application. The results with methodology explained in details (Slovenian) can be found at `http://wwwsi.cati.si`.

Development of Slovenian Municipalities

Jože Rovan and Jože Sambt (University of Ljubljana, Slovenia)

There is common belief that there are big differences in the level of development of Slovenian territory. Regional policy is needed. This is also an important subject in EU countries. In the paper we are trying to present the current situation in the development level of Slovenian territory. Slovenian regions are at present situation merely a statistical construct, so we focused on the level of municipalities. For a comprehensive analysis we used economic, demographic and social indicators and an indicator of the level of living. However we have to be careful - some variables are not appropriate for the analysis because of the small size of municipalities.
We have used cluster analysis to form groups of municipalities with respect to the level of development. In the first steps we applied Ward hierarchical method; on the basis of a dendrogram three or four groups of municipalities can be identified. In the second step we applied leaders method, using the centroids (from the first step) as group seeds. All the variables used in the analysis were standardised. the results show obvious differences between the developed west and the undeveloped east. Ljubljana, together with the adjacent municipalities and the western part of Slovenia, forms a group of the 'most developed' municipalities. Municipalities in the eastern part of Slovenia are 'developed' or 'less developed'. Some variables and Bonferroni's test support the decision for four groups of municipalities. We believe that our findings have revealed those 'underdeveloped' municipalities (problematic areas) to which regional economic policy should pay special attention. These municipalities are relatively small and also not numerous, therefore sufficient governmental support for enforcing their economic development probably should not represent a too large financial burden for Slovenian budget.

The SNAC Project: Early Results of the Social Network Analysis in a Community Health District

Patrizia Rozbowsky (University of Trieste, Italy)
Miriam Totis and Mario Casini (Udine Health District, Udine, Italy)
Tina Kogovšek (University of Ljubljana, Slovenia)
Miriam Isola and Franca Soldano (University of Udine, Italy)
Dario Gregori (University of Trieste, Italy)
Anuška Ferligoj (University of Ljubljana, Slovenia)

The growing interest in social networks analysis in the health care context is due to the beneficial effects of social relationships, particularly social support, on the daily life of patients. The present project aims at measure social support in a sample of heart failure, cancer and dementia patients belonging to Udine health district and having a home based care.
The objectives of the on going study are the identification of different typologies of support (formal, informal or hidden support), the assessment of the patients' health related quality of life, the evaluation of patients' satisfaction (or perceived quality of service) and the measurement of quality of life of patients' spouses.
In order to obtain this information we developed a questionnaire which has been administered during a telephone interview to a selected sample. We also analyzed interviewers' personality and level of burnout.
In the present paper we will present the early results (after 3 months from the start up).

Propensity Score Methods in Marketing Applications
Donald B. Rubin (Harvard University, USA)

Propensity score methods have found wide application in a variety of areas, including medical research, economics, epidemiology, and education, where randomized experiments are difficult to perform or raise questions of ethics, and thus answers often must be obtained from observational (nonrandomized) data. Thus far, there have been few applications to marketing situations, even though the effects of many interventions (e.g., advertising, promotions) are ”assessed”using generally inappropriate techniques such a regression, data mining, etc. This talk will present this new approach, and show how the more standard predictive methods can be used appropriately as a supplement. A real success story will also be presented.

Mixture Models of Missing Data

Tamas Rudas (Eötvös Loránd University and TARKI, Budapest, Hungary)

The current practice of handling the problem of missing data considers incomplete observations a deficiency of the survey. Accordingly, analysts do their best to make up for missing data and the magnitude and other characteristics of the incomplete information are often suppressed in the final reports.
This talk takes the position that missing data is an unavoidable feature of any survey of the human population, and deserves careful modelling. The population of interest is considered as consisting of an observable and unobservable part and the true distribution is a mixture of the distributions on these two parts. When the goal of the analysis is the assessment of the fit of a statistical model, this mixture is compared to that of the so-called mixture index of fit. Estimates of the magnitude of the unobserved and no-fit parts of the population are obtained under various circumstances. This model can be specified for various types of missing data and for various forms of the missing data mechanism.

Power of Chi-Square Goodness-of-Fit Tests in Structural Equation Models: The Case of non-Normal Data

Albert Satorra (Universitat Pompeu Fabra, Barcelona, Spain)

The analysis of structural equation models (SEM) has become and important tool of investigation in various disciplines, as behavioural, educational, and economic sciences. In such analysis, a chi-square goodness of fit model test is often used. When dealing with non-normal data, researchers are confronted with the choice of competing chi-square goodness-of-fit test statistics. In the context of SEM, the present paper investigate the asymptotic and finite sample distribution of various chi-square goodness-of-fit test statistics. We consider a general setting where a) the data may be non-normal, b) the estimation method is not necessarily (asymptotically) optimal one, and c) model misspecification is allowed. Power of the test is computed distinguishing whether asymptotic robustness (AR) hold or not. The power of the various test statistics is compared using both asymptotic theory and Monte Carlo simulation. The asymptotic robust and the scaled version of normal theory (NT) chi-square goodness of fit tests are compared. A scaled version of a NT goodness-of-fit test statistic for ULS analysis is included among the statistics investigated.

Rotation of Sample Units and Weighting System in the Case of Retail Trade Survey

Rudi Seljak (Statistical Office of Republic of Slovenia, Ljubljana, Slovenia)

Retail Trade Survey in Slovenia is a monthly sample survey based on a sample of 1300 units. The survey is rotated panel what means that all units selected at the beginning of a year are kept in the sample for 12 months and then one quarter of the sample is replaced with new units. Due to many demographic changes, the target population is significantly changing through years and that why some adjustments in classical selection methods and calculation of weights were introduced.
In article the introduced methods along with some comparison results will be presented. Also some aspects of variance estimation regarding the existing sample design will be introduced.

Sensitivity Analysis for Informative Censoring in Parametric Survival Models: Basic Principles and Extension in the Presence of Covariates

Fotios Siannis (Institute of Public Health, Cambridge, UK)
John Copas (University of Warwick, UK)
Gouobing Lu (MRC HSRC, University of Bristol, UK)

Methods for analyzing censored survival data are highly developed and widely used. Although many different models and approaches have been studied in the literature, they almost invariably assume that the censoring is *non-informative* or *ignorable*. Whatever the mechanism is which determines the times of censoring, it is deemed irrelevant as far as inference about the distribution of T is concerned. In many applications, however, the assumption of ignorable censoring is at best an approximation and at worst seriously misleading. Lagakos (1979) gives a number of such examples where the assumption of non-informative censoring is questionable. It is clear that if such dependence is ignored, the resulting inference will be biased.

In this work we study the bias induced by informative censoring by embedding censored survival data in a competing risks framework. For each individual we assume there is a potential random censoring time C and a potential random lifetime T. The censoring is non-informative if C and T are independent (conditional on values of covariates). We observe the time $Y = \min(T, C)$, and the censoring indicator $I = 1$ if $T \geq C$ and $I = 0$ if $T > C$. If $f_C(c, \gamma)$ is the marginal distribution of the censored times, we propose the parametric model

$$P\big(C = c | T = t\big) \quad = \quad f_C\big(c, \gamma + \delta B(t, \theta)\big) \tag{1}$$

which allows for dependence in terms of a parameter δ and a bias function $B(t, \theta)$. θ is the parameter of $f_T(t, \theta)$, and hence the parameter of interest. The initial assumption is that the conditional distribution of C given T has exactly the same functional form with the marginal distribution, with the only difference being in the parameter.

Given that we are unable to estimate the level of dependence between lifetime and censoring mechanisms, Tsiatis(1975), we argue that the next best thing is to develop a sensitivity analysis which will enable us to see how robust our estimates and conclusions are to different degrees of dependence which may be present in our data. We develop a relatively simple sensitivity analysis using linear approximations to parameter estimates for small values of δ. Consequently, the maximum likelihood estimate (MLE) $\hat{\theta}_\delta$ equals

$$\hat{\theta_\delta} \quad \simeq \quad \hat{\theta_0} + \delta S \tag{2}$$

where $\hat{\theta}_0$ is the independence MLE and S is essentially the *Sensitivity Index*, which includes functions from both processes.

Bias function $B(t, \theta)$ plays an important role in the sensitivity index. Under certain proportional hazard assumption we can see that the choice $B(t, \theta) = 1 - H_T(t, \theta)$, where $H_T(t, \theta)$ is the cumulative (integrated) hazard function of the failure process, produces a simple and symmetric form for the above expression. Furthermore, although parameter δ is probably the most important parameter in the model, the fact that it is assumed known allows us to choose values arbitrarily. Nevertheless, an attempt to link δ to the correlation between any two functions of the two processes results to

$$Corr(t, c) \quad \leq \quad \delta, \tag{3}$$

which immediately gives a clear idea of the order of magnitude of δ, as well as bounds for it.

Extension of the model to include covariates is a straight forward application. We have a vector of covariates x_i for each patient i, and the sensitivity analysis is performed on the vector of parameters θ. Similarly, if we consider a different quantity to be of main interest, then we can perform the sensitivity analysis directly on this quantity.

Examples are presented to illustrate the theory.

Reliability in Communication Systems under non-Gaussian Disturbances

Adnan Al-Smadi and Mahmoud Smadi
(Jordan University of Science & Technology, Irbid, Jordan)

The purpose of communication system is to carry information-bearing baseband signals from one place to another over a communication channel. The objective is to design communication systems that are both efficient and reliable. While efficient communication from the transmitter to the receiver is attained through source coding, reliable communication over a noisy channel is attained through error-control coding. In this paper, the reliability of some communication systems is studied under non-Gaussian noise distributions. We consider a signaling system with different heavy tailed noise distributions. In particular, Cauchy, Laplace, and logistic distributions are considered. The reliability of the communication systems is studied by calculating the error probabilities and compared with the case of Gaussian noise. Using the error probabilities, it is found that the reliability of the signaling system under non-Gaussian noise distributions is low. Simulations indicate significant performance in achieving high reliability if we double the signal-to-noise ratio and increase the number of repetitions when sending the same signal different times.

A Comparative Evaluation of Minimum Rank Factor Analysis, Minres and Maximum Likelihood Factor Analysis

Gregor Sočan (University of Groningen, The Netherlands)

Minimum Rank Factor Analysis (MRFA) is a method of common factor analysis which yields, for a given covariance matrix C, a diagonal matrix U of unique variances which are nonnegative and which entail a reduced covariance matrix C-U which is positive semidefinite. Subject to these constraints, MRFA minimizes the amount of common variance left unexplained when we extract any fixed small number of common factors. Another unique feature of MRFA is that it makes possible a separate estimation of the amount of the explained common variance, the unexplained common variance and the unique variance.The aim of this research was to evaluate the practical utility of MRFA by means of simulated datasets. The criteria included the accuracy of retrieval of the population factor loadings, communality estimation, reproduction of correlations and the frequency of so called weak Heywood cases (boundary solutions). The influence of sample size, number of extracted factors and standardization of the sample covariance matrix was also investigated. The performance of MRFA was finally compared to the performance of two popular methods of exploratory factor analysis: MinRes and Maximum Likelihood. Based on the results, some suggestions about the choice of a factor analysis method will be given.Topic area: Data Analysis Techniques

On the Use of Frailties in Proportional Hazards Models

Janez Stare (University of Ljubljana, Slovenia)
John O'Quigley (University of California at San Diego, USA)

We discuss the concept of a frailty model where each subject has his or her own disposition to failure, their so called frailty, additional to any effects we wish to quantify via regression. Although the concept of individual frailty can be of value when thinking about how data arise or when interpreting parameter estimates in the context of a fitted model, we argue that the concept is of limited practical value. Individual random effects (frailties) can be made to disappear by elementary model transformation. In consequence, unless we are to take some model form as carved in stone, and if we are to understand the term 'frailty' as referring to individual random effects then frailty models can be misleading.
However, using the equivalence between proportional hazards models with frailties and non proportional hazards models, frailty models can be used to address the question of fit. A goodness of fit test of the proportional hazards assumption against an alternative of declining regression effect is equivalent to a test for the presence of frailties. Such tests are now widely available in standard software. Although a number of tests of the proportional hazards assumption have been developed there is no test that directly formulates the alternative in terms of a non-specified monotonic decline in regression effect and that enables a quantification of this in terms of a simple index. The index we obtain lies between zero and one such that, for any given set of covariates, values of the index close to one indicate that the fit cannot essentially be improved by allowing the possibility of regression effects to decline. Values closer to zero and away from one indicate that the fit can be improved by relaxing the proportional hazards constraint in this particular direction.

Application of the Latent Growth Curve Model: Methodological Issues

Reinoud Stoel and Godfried L.H. van den Wittenboer
(University of Amsterdam, The Netherlands)
Joop J. Hox (Utrecht University, The Netherlands)

Growth curve models are well suited to analyze systematic change in longitudinal data collected from a panel design. They represent outcome variables explicitly as a function of time and other measures. This paper focuses on three advanced methodological issues that can pop up in the application of latent growth curve methodology within the framework of structural equation modeling. The issues discussed are: (1) estimating the shape of growth curves, (2) multiple indicators of the latent construct at each time point: the issue of measurement invariance, (3) time-invariant covariates.

The presentation will focus on the first issue, concerning an approach for modeling non-linear growth curves originating from the work of McArdle (1986) and Meredith and Tisak (1990). In contrast to higher order polynomial growth curve models, the model introduced by these authors allows the nonlinearity in the growth curves by estimating the basis coefficients (i.e. the factor loadings) for the growth factor, instead of including higher-order polynomials. While this model presents a challenging and elegant way of modeling growth, it contains some inherent pitfalls, which, to our knowledge, have not been addressed explicitly enough in the literature thus far. Yet, the pitfalls of this growth curve model should be recognized at some point in the application to fully understand the opportunities.

Explaining systematic differences in growth between subjects is one of the main objectives of growth curve modeling. It seems logical, therefore, to incorporate the non-linear growth model into larger structural models. Care should be taken with this procedure since it may lead to biased estimates of some parameters of this growth curve model. The presentation discusses latent growth curve models with estimated basis coefficients and it proposes ways to deal with the problem.

References

1. McArdle, J.J. (1986): Latent variable growth within behavior genetic models. *Behavior Genetics*, **16**, 163-200.
2. Meredith, W.M. and Tisak, J. (1990): Latent curve analysis. *Psychometrika*, **55**, 107-122.

Correspondence Analysis and Categorical Conjoint Measurement

Anna Torres-Lacomba (Universidad Carlos III de Madrid, Spain)
Michael Greenacre (Universitat Pompeu Fabra, Barcelona, Spain)

We show the equivalence between the use of correspondence analysis of concatenated tables and the applications of a particular version of conjoint analysis called categorical conjoint measurement. The connection is established using canonical correlation. The second part introduces the interaction effects in all three variants of the analysis and shows how to pass between the results of each analysis.

We start by showing how correspondence analysis can be applied to tables concatenated in a certain way in order to emulate a particular algorithm of conjoint analysis applied to categorical data. Conjoint analysis can be understood as a technique which predicts what products or services people will prefer and assesses the weight people give to various factors that underlie their decisions. There exist different conjoint algorithms for analyzing such data, depending on the type of conjoint measurement: in our case we are interested in conjoint measurement on a categorical scale. For this case there exists an algorithm due to Carroll (Carroll, 1969) known as categorical conjoint measurement. We use canonical correlation analysis applied to dummy variables as a bridge in order to show the equivalence in results between categorical conjoint analysis and correspondence analysis. Previous literature has already demonstrated the equivalence between simple correspondence analysis and canonical correlation analysis for two categorical variables. Carroll used canonical correlation analysis in his categorical conjoint algorithm, which could handle any number of categorical design variables, or attributes. Our first innovation came with the demonstration of the equivalence between correspondence analysis and canonical correlation analysis for more than two categorical variables. The key point is in the way of coding the data prior to correspondence analysis. Data are coded as a concatenated table of frequencies, where rows are the levels for the different attributes and columns are the levels for the categorical response variable. This table was recovered from two indicator matrices, one of them composed of dummy variables for the levels of the different attributes, and the other one composed of dummy variables for the levels of the categorical variable.

We illustrate the equivalence in an application using a data set given by Rao (1977). It comes from a situation of an apartment-dweller planning to purchase a house that is already built in a college town. The decision-maker has isolated the attributes of the house considered most important in the decision. The attributes are: size of the house, price of the house, and general conditions, each one with several levels. The response variable is composed of four levels, which are: very high worth, just high worth, just low worth and very low worth.

An issue in the conjoint analysis literature for future research is the study of interaction effects. We incorporated interactions in canonical correlation analysis for

the usual way of coding data as well as for a new one, establishing the connection with correspondence analysis and categorical conjoint measurement in the presence of interactions. In canonical correlation analysis, the usual way of coding an interaction is as the products of all the main effect dummy variables, having omitted one dummy variable for each attribute. A new way is to include all products of the main effect dummies, dropping just one of these products. It is this latter way which is easier to relate to the correspondence analysis approach. In correspondence analysis no dummy variables have to be omitted, neither from main effect attributes nor their products.
An application of interactions in conjoint analysis came from an airline company study and involves the preferences of a single respondent in the context of transatlantic flights. The objective was to know the trade-off value between the different attributes offered as well as possible interaction effects between them. The attributes were: airline company, price, service and timetable, with their particular levels, and a categorical response variable expressing preferences, with four levels. We started with the analysis without interaction effects and then introduced one interaction effect to see the improvement. Based on inertia values, which are measures of variance in correspondence analysis, we decided to create an interaction effect composed of the variables price and timetable. We could then capture different effects, for example that the timetable does not matter at the lowest price level, while as price level increases, this attribute becomes important.

Bibliometric Research on Statistical Methods in Biomedicine as an Aid to Curriculum Development

Gaj Vidmar (University of Ljubljana, Slovenia)

Several courses in (bio)statistics are taught at the Institute of Biomedical Informatics of the University of Ljubljana as part of undergraduate and graduate study at the Faculty of Medicine, undergraduate study at the School of Public Health and in collaboration with the Institute of Public Health. The topics covered range from fundamentals to multivariate methods, and course objectives and contents are constantly adjusted to students' profile.

One of our efforts aimed at keeping the curriculum up-to-date and ensuring quality of education is regular bibliometric research. Results of two surveys of statistical methods applied in scientific papers are presented: one in the New England Journal of Medicine and one in four journals from the field of public health, available in full-text electronic form at the University of Ljubljana.

A taxonomy of statistical methods, based on experience rather then conceptual rigor, is introduced as a useful starting-point for such research. A summary of findings is presented, including data presentation issues and software usage. Implications are focused on our graduate course in contemporary statistical methods in medicine and possible topics for a summer school on data analysis in public health are proposed.

Qualitative Technique - Group Dynamics Simulation Technique
Miha Vogelnik (CATI, Ljubljana, Slovenia)

In the world of business decisions are mainly based on numbers and in-depth analytical statistics. However, the modern mans world is complex and often cannot be so easily revealed by simply applying quantitative methods to data gathering. Qualitative techniques are gaining greater and greater importance in providing insight into the meaning of hard data. Following different disciplines and various techniques that emerge within them, we see that conducting qualitative research is not an easy task. In order to achieve accepted level of validity data gathering often has to be repeated which requires time and money. In the world of business both being the prime obstacles in achieving good results, while the path to results is often determined by pragmatic approach.

Such pragmatic approach was one of the reasons to search for a qualitative method that would take the time and the financial component of data gathering into account and would at the same time provide valid qualitative-quantitative mixture of data. This mixture was found to be important because clients many times wanted (representative) hard data within the qualitative survey in order to be able to estimate profit, sales, market potential, etc... A two-step approach - a qualitative survey followed by a quantitative survey or vice versa - was often out of the question (neither time nor money allowed it). During 1999-2001 we evaluated different techniques and modified them in order to find such a method. Based on various research projects, which we did for our clients, and approaches to data collection, which were included in those projects, we created a technique that met our criteria. We call it group dynamics simulation technique (GDST).

In its essence the GDST is a qualitative technique that can be used alone or in combination with other qualitative as well as quantitative techniques. If used in combination with other techniques, the GDST uses the other technique as a vehicle upon which it is applied. From that aspect the GDST can be seen as a multi-method technique. As such vehicles in-depth interviews (un- or semi-structured) or a quantitative questionnaire are used. Following the logic of the NGT technique or the Delphi method, the respondents answers are shared among the respondents upon which they can express their opinion. However, the GDST is not a group technique but an individual technique. The distribution of opinion makes it possible to simulate group dynamics known from group techniques. However, in the case of the GDST the researcher has control over the contents in much the same way as when using the Delphi method. The difference being that in the GDST our goal is not reaching the consensus among respondents in regards to the researching object but rather to gather information in the same way as we do with in-depth interviews, but with additional responses to this information as would be done in focus groups. With the GDST we get depth as well as width in information

gathering. Additionally, since quantitative questionnaire can be used as a vehicle mentioned, the GDST can also provide quantitative information that is so important to the world of business.

New Ways of Survey Data Collection: Touchscreens
Martin Weichbold (University of Salzburg, Austria)

Introduction

Empirical research in the social sciences is unthinkable today without computers. From a historic view, they were firstly used for data analysis, but over the years also all other phases of the research process were carried out with the support of computers: project planning as well as sampling, organisational things as well as the presentation of the results. The 'last' computer-free part for a long time was the interview situation itself: even in Computer Assisted Telephone Interviewing - CATI - the answers are 'produced' in a dialogue between the interviewer and the respondent, who himself is not aware of the computer. Although some forms of computer-based self-administered questionnaires were presented already in the seventies (e.g. disk-by-mail), they didn't gain any significance in empirical social research. But again it was a question of time: with the increasing number of computers in private use and their connection via the Internet, World-Wide-Web-surveys were the first data collection method to become popular with the respondent answering a questionnaire directly facing the computer. But there are further forms of computer-based interviewing that might play a major role in future: For a few years I am working on surveys using touchscreen terminals. The results of various studies are very encouraging, although some methodological questions are still open.

The presentation firstly gives a short description of touchscreen interviewing and its fields of application; in a second step some characteristics of touchscreen-surveys will be discussed. At last a comparison with WWW-Surveys will lead us to the question, whether it is possible to work out some characteristics for computer-based forms of data collection.

Touchscreen Surveys

Touch-screens are touch-sensitive monitors; the computer treats a finger-tip as a mouse-click. The technology itself is not new and already in use for several years in various fields like information terminals or for the operation of machines in industry. We use this technology to conduct surveys, which - on some aspects - are similar to WWW-surveys: a question is faded in, the interviewee touches applicable answer on the screen, the next question is called up and so on. Like the well-known WWW-surveys this provides a row of possibilities: the questionnaire is dynamic, as the order and the selection of the questions depends on the previous answers and is not rigidly pre-set. This makes touchscreen surveys highly appropriate even for very heterogeneous target groups or multilingual surveys. Of course it's possible to integrate pictures, video- or soundfiles into the questionnaire, the questionnaire can be adapted at short notice, it's a very flexible instrument.

But there are some import differences to Internet surveys: data are stored on the

local terminal or a local area computer network (LAN); the terminals are set up without keyboard and mouse, the only possible input is touching the screen (which makes it impossible to crack the system and enables even people with little or no computer experience to fill out the questionnaire). But the main difference is the interview situation: touchscreen-terminals are located in public space, accessible to everybody passing by. Therefore touchscreen surveys are only appropriate for certain kinds of studies: The crucial point is that the interviewees have to come to the interviewer (= the terminal), not the other way round. In order to meet this point, a terminal has to be located in a frequented place in order to address enough people. On the other hand, people should be in a situation that makes them likely to take part in the survey, they should have some time and not be only passing by in a hurry. From this point of view, the evaluation of customer or visitor satisfaction in museums and exhibitions or passenger surveys at airports (while waiting for boarding) are highly appropriate for touchscreen interviewing. Additionally, there is a strong economic argument: the costs of a touch-screen survey are independent of the sample size. Apart from the costs for the terminal the only costs incurring are those for programming the questionnaire. There are no additional costs for printing the questionnaires, for paying an interviewer and collecting the data, as the respondents enter their data themselves.

Methodological characteristics of touchscreen surveys

The findings are based on about 10 projects with samples ranging from a few hundreds to tens of thousands (in a long-term visitor satisfaction monitoring running since June 2001). In three projects the same questionnaire was additionally used in face-to-face- and/or paper&pencil interviews in order to evaluate the comparability of the data collection methods. Furthermore, a observation of respondents and interviews with drop-outs in two of the surveys shall bring us to a better understanding of the interview situation and the motivation of people to take part in a survey or to break off.

Analyses have shown that three - interdependent - major areas can be distinguished; the first central methodological issue is the self-selection of the interviewees: in conventional surveys the researcher selects a probability sample out of the population; in touch-screen surveys neither the individual elements of the population are known, nor is the sample chosen by the researcher: the terminals just stand there with an inviting start screen and maybe a poster pointing out the survey. Basically, they are addressed to everybody passing by. The respondents decide whether they want to participate or not, i.e. they select themselves. However: as practice shows, self-selection is 'successful' not only relating to the number of respondents, but also to the demographic structure of the samples; of course this fact can't eliminate the basic problem. Solutions like online panels, which can help to avoid self-selection in WWW-Surveys, cannot be implemented for touchscreen surveys.

A second area, where again some similarities with WWW-Surveys can be ob-

served, is answering drop out, i.e. interviewees who start the interview but do not finish it. This kind of non-response may also happen in conventional interview situations, but is - quantitatively - of no significance there. The situation is completely different with the collection of data in computer environments: the probability of an interview being answered completely is under some circumstances lower than of interviewees dropping out during the interview. Additionally a typical drop-out 'behaviour' can be observed in the course of the interview: most people break off after the first or the first couple of questions, a phenomenon that is also known from questionnaires carried out via the Internet. A crucial point is how to deal with the drop-outs, as their answers differ significantly from complete responders (as far as we have any information on them anyhow).
The third point is the question, whether data collected by touchscreens are comparable with those from conventional data collection methods. As it can be seen so far, touchscreens yield similar but not identical results, with the deviations obviously following certain regularities. For example, social desirability of answers seems to have less impact as answers to rating questions show a more critical tendency. Obviously this is a positive effect of the absence of an interviewer.

Discussion

Although we are far from being able to tell how to deal with all the open methodological questions or even being able to give a final evaluation, it's possible to show some interim results. Touchscreen interviews are both economic and methodical good ways of survey data collection, at least for specific purposes.
It is striking that in some regards strong similarities between touchscreen- and WWW-surveys can be observed, despite of the completely different interview situations. This give rise to the question, whether the fact of a computer being present and a interviewer being absent in the interview situation provides some specific effects: Are there 'common characteristics' of data collection in a computer environment and how can we deal with? Is it possible to work out a 'methodology of computer-based data collection', combing the findings both of WWW- and touchscreen surveys? Of course, only preliminary answers can be given for now and some questions will have to stay unanswered, but further research should help us to improve touchscreen surveys and make them an established data collection technique.

What Can We Say about the Case?

Malcolm Williams (University of Plymouth, United Kingdom)
Wendy Dyer (University of Durham, United Kingdom)

Variable analysis of survey data utilises the standard or frequency theory of probability to measure the relative frequency of an event A in a sequence defined by conditions B. Thus the objective probability of an event A occurring is conditional upon B - symbolised as $p(A/B)$. However the odds of A occurring in an actual individual will not be equal to those predicted for an 'ideal' individual. To say, for example, that a 20 year old Dutch young offender is only half as likely to re-offend as his British counterpart can say nothing about the actual probability of a given Dutch 20 year old offender to re-offend.

Yet the frequency from which we can predict the odds of an event for an ideal individual is made up of the dispositional properties of actual individuals, the single cases. It must follow from this that actual events themselves have a probability of occurrence, yet despite these individual probabilities producing the frequency, the individual probabilities cannot be known from the frequency. This paradox has been recognised since the time of Von Mises, who believed it to be irresolvable and was taken up by Popper, in the 1950s, who believed a resolution was possible (Popper 1957; 1959).

The resolution proposed by Popper was at a theoretical level and its practical applications never explored. However the social survey has always suffered from a weakness in respect in individual analysis, a weakness often noted and exploited by interpretivists. In recent years a new family of methods, cluster analysis, initially used in biology have begun to find favour. In this paper we argue that case based analysis, using cluster techniques, both vindicates Poppers solution and offers a potentially powerful tool for the analysis of individual careers over time. The paper begins by setting out the argument for single case probabilities, as proposed by Popper, then describes the method of cluster analysis and illustrates its use with a short case study of Custody Diversion Teams in the North East of England.

Analysis of Collaboration Networks

Matjaž Zaveršnik and Vladimir Batagelj (University of Ljubljana, Slovenia)

On the Internet we can find many databases of scientific papers from different areas. From such databases we can build networks of collaboration between scientists (collaboration networks). In these networks two scientists are considered connected if they have co-authored one or more papers together.
Many bibliographic data can be found in BibTeX format [1,2]. Usually we need special programs to convert BibTeX data into collaboration network. Because of errors in original data we also have to verify the network and correct the errors manually.
We shall present two approaches to analysis of collaboration networks (cores and decomposition according to short cycles). Each approach will be demonstrated on real life examples.

Links:

1. B. Jones, Computational Geometry Database (GeomBib):
   http://compgeom.cs.uiuc.edu/~jeffe/compgeom/biblios.html
2. Nelson H.F. Beebe's Bibliographies Page:
   http://www.math.utah.edu/~beebe/bibliographies.html

Slovenian National Readership Survey - Methodological Issues

Andraž Zorko, Andrej Kveder, Tomaž Hohkraut, and Zenel Batagelj
(CATI, Ljubljana, Slovenia)

In July 2001 a new tender for National Readership Survey (NRS) was released. The new survey consists of much more indicators for print media than previous one. Because of these new indicators new survey design was needed.
The survey design includes interviewing in two steps, for some indicators a telephone survey is needed in the first step, which acts also as a recruitment survey for the second step - the face-to-face survey. Because of this two-step survey design and because of cost optimisation an interesting sampling design is used.
Whole surveying process is computerized: starting with sampling administration, CATI, face-to-face respondents selection, very complex CAPI and data gathering at the end.
One of the most demanding parts of this project is the weighting process. The telephone sample is weighted on basic demographical variables where each day of the week needs to be as 'representative' as possible. On the other side the face-to-face sample should be representative on same variables. Besides that the F2F sample is the subsample of the telephone sample where also variables of more 'dependent' character are available: print media consumption in general and 'day-before reading'. In turned out (what was expected) that respondents in F2F sample tend to read more. To avoid that the F2F sample is weighted also regarding print media data from telephone survey. Raking (RIM weighting) method is used with some new approaches for marginals selection.

Comparison of some Multivariate Statistical Methods of Grouping Customers into Segments

Aleš Žiberna and Vesna Žabkar (University of Ljubljana, Slovenia)

Market segmentation involves identifying homogeneous groups of consumers who behave differently in response to a given marketing strategy. For segmentation we can apply different statistical multivariate methods that make it possible to group customers into segments according to similarity in a set of demographic, socio-economic, psychographic, behavioural and other variables. The purpose of this article is to compare usage of different methods of clustering (Wards method, Complete linkage method) combined with selected variables with highest variability, Principal axes factoring or Principal components method for preparing data. Data from a simple random sample of 241 consumers that participated in a survey on usage of health-strengthening food supplements was applied for this purpose. The survey included demographic, socio-economic, psychographic variables and reported behaviour data.