The Japanese Journal of Behaviormetrics, 1987, 14, 38-40.

行動計量学 14 巻 2 号　通巻 27 号）
1987 年．38－57

第 14 回　日本行動計量学会大会発表一覧　座長報告

1986 年 8 月 25 日（月）～27（水）
於　東京大学経済学部

特別講演　『認知統計学の勧め』

マッギル大学　高根　芳雄
司会　大学入試センター　柳井　晴夫

Development of cognitive statistics and statistical expert systems

Yoshio Takane, McGill University

1. Background and Motivation

The rapid progress in computer technology in recent years has had a tremendous impact on statistical practice. With hardware cost down and user friendly statistical software packages, it looks as if statistical data analysis (SDA) was at everyone's disposal. Indeed these technological advances have greatly facilitated some aspects of SDA. For example, it has released statistical practitioners from the heavy computational burdens often involved in SDA. Furthermore, new methodological developments (e.g., iterative model fitting, nonparametric density estimation, resampling methods, interactive computer graphics etc.) are underway which would never have been envisaged without the massive computational power of modern day computers. However, this does not necessarily imply that the quality of SDA has improved in general, or guarantee that SDA practioners have a better chance of better SDA. In contrast, as an undesirable side effect, there is a greater chance of malpractice of SDA, and there is some indication that the number of "abuses" of SDA has increased considerably in recent years.

The problem is that most of the existing statistical packaged programs implement only technical (computational) aspects of statistical methods with minimal emphasis on strategic aspects of SDA. Theory, models and associated numerical/computational techniques are certainly essential ingredients of statistical expertise, but at least equally important are such "human factors" as practical know-how of SDA (which quite often must be gained through practical experiences with SDA), strategies, heuristics and other human cognitive skills. Any serious SDA practitioners will readily realize that a good SDA requires at least 1) translating one's research questions into statistical terms, 2) planning the course of analysis (e.g., selecting appropriate techniques according to aims of analysis and constraints from the data), 3) diagnosing and interpreting results, and 4) selecting what to report and how to report. Unfortunately none of these aspects are within the proper framework of current statistical packages (Goto, 1986).

What is crucially needed is more intelligent statistical software and basic research on strategic aspects of SDA, upon which the software is built.

2. Cognitive Statistics : A proposal

Cognitive statistics (coined by myself) is a new branch of statistics that emphasizes the importance of

various human cognitive factors in SDA. More specifically it consists of:

1) Epistemology of scientific knowledge and methodology; philosophical foundations of inductive and deductive reasoning.

2) Re-evaluation of conventional statistical theory; possible development of a new normative theory for information integration; examination of new methodologies in statistics (e.g., robust techniques, exploratory data analysis, Bayesian statistics, interactive computer graphics, etc.)

3) Studies of human cognitive processes, such as those underlying human pattern recognition, causal reasoning, heuristic reasoning, problem solving, etc., more focussed studies of human information processing capabilities pertaining to SDA (e.g., how humans acquire competence in such complex cognitive tasks as SDA, how expert researchers formulate their research questions, how expert statisticians use their statistical expertise in actual data analysis, how expert statistical consultants give statistical advice, etc.).

4) Applications of the knowledge obtained above to improve the quality of SDA, and of statistics training and education; more concretely, development of statistical expert systems.

Cognitive statistics is thus interdisciplinary, encompassing a wide range of areas, including philosophy (epistemology), conventional and new statistical theories, cognitive psychology, and computer science (in particular, data management systems, artificial intelligence and expert systems).

3. Statistical Expert Systems

The development of a statistical expert systems (SES) plays a central role in the development of cognitive statistics. An SES is a computer implementation of statistical expertise in such a form that "it can provide intelligent statistical advice, make intelligent decisions, and explain its lines of reasoning for such decisions." A machine realization of the statistical expertise is a corner stone for the development of cognitive statistics, because how competent a statistical expert we can build into a computer serves as an ultimate test of how successful we are in understanding the nature of statistical expertise, and more fundamentally, the nature of human cognitive skills underlying the complex cognitive tasks. At the same time such a system would be invaluable from users' viewpoints; for the statistically naive it offers appropriate statistical guidance, and for statistics students who intend to become expert statisticians it serves as a teacher from whom to learn the statistical expertise, always an integral part of professional competence.

We plan to build an expert system capable of 1) "discussing" a subject matter of a study with the researcher, 2) recommending appropriate statistical techniques, 3) carrying out the analysis, and possibly, 4) interpreting major results. In carrying out the analysis the system checks assumptions underlying the analysis, tells the researcher what is wrong (if there is anything wrong), takes appropriate corrective actions if necessary, asks for guidance where appropriate, and tells the researcher what it is doing and why it is doing it. It should be as user friendly as possible in the sense that it has a tutorial capability (e.g., manual-free), caters for user errors, and can explain technical terms when asked. It should allow maximum interactions with the user, when desired (e.g., user interventions, user control of the course of analysis, analysis recalling features, etc.) Ultimately a self-evolutionary system should be developed; that is, the system that learns from experience and improves its performance. At the minimum the system should accomodate easy modifications while maintaining its consistency. (These as well as other desirable attributes of a statistical expert system are discussed in Hand, 1985a).

4. Where to Start

Some related work is currently underway. For example, Novick (1980) has been working on CADA (Computer Assisted Data Analysis) for quite some time; this is an extremely user friendly environment for interactive data analysis in the Bayesian framework. Gale and Pregibon (1984; Pregibon & Gale, 1984) have built a semi-expert system called REX (Regression Expert) which focussed on a particular analysis

technique (simple regression analysis). REX is built on "S" (Becker & Chambers, 1985) for interactive computer graphics for data analysis. Gale (1986) is now working on STUDENT that learns strategies for SDA through examples worked out by human experts. Hand (1985b), on the other hand, focussed on a particular aspect of SES, and constructed a rule based system for selecting appropriate statistical techniques. (See also Poitier & Lai, 1983; STATPATH). The system called RX (Blum, 1982) translates a research question posed in medical terms into a description of the statistical study that needs to be done to answer the question. These developments indicate that people recognize the need for work in this area, and with the development of artificial intelligence (AI) techniques and accumulation of knowledge in cognitive psychology, more sophisticated, full-range SES will be constructed (Hand, 1985b; Gale, 1986).

SES seems interesting in the following analysis domains: experimental design, analysis of data from controlled experiments (ANOVA/MANOVA, nonparametric tests), regression analysis, path analysis, discriminant analysis (DA), analysis of frequency tables by the log-linear model, factor analysis (covariance structure analysis, linear structural relationships), multidimensional scaling (MDS), quantification method III (dual scaling), cluster analysis, etc. Realistically, however, one should start with what one knows best or where the need is most pertinent. In this regard analysis of experimental data and DA are of particular interest. I have been teaching experimental design courses for quite some time, and I have always felt that students in those courses should be given opportunities to observe expert statisticians working out their problems. Takane and Shibayama (1984) developed what they call ideal point DA. Over the years a number of data sets have been analyzed by the method, and knowledge has been accumulated for intelligent use of this method. A computer program should be written incorporating this knowledge. At the minimum this program should have the following capabilities: various diagnostic capabilities (outlier detection, error analysis, sensitivity analysis), stepwise selection of predictors, optimal data transformations, automatic interaction detection (AID), proper handling of missing data, and bootstrapping for estimating error rates.

5. References

Becker & Chambers ("S" Wadsworth, 1984). Blum (*The RX project*, Springer, 1982). Gale (*AI and stat.*, Addison-Wesley, 1986). Gale & Pregibon (*Compstat*, 1984). Goto (*Jap. J. of Behaviormetrics*, 1986). Hahn (*Amer. Statistician*, 1985). Hand (Proc. of 45th ISI, 1985a; *J. of Applied Stat.*, 1985b). Novick (*Psychometrika*, 1980). Poitier & Lai (*Proc. of ASA Stat. Comp. Section*, 1983). Pregibon & Gale (*Compstat*, 1984). Takane & Shibayama (*Proc. of 12th Behaviormetric Soc. of Japan*, 1984).