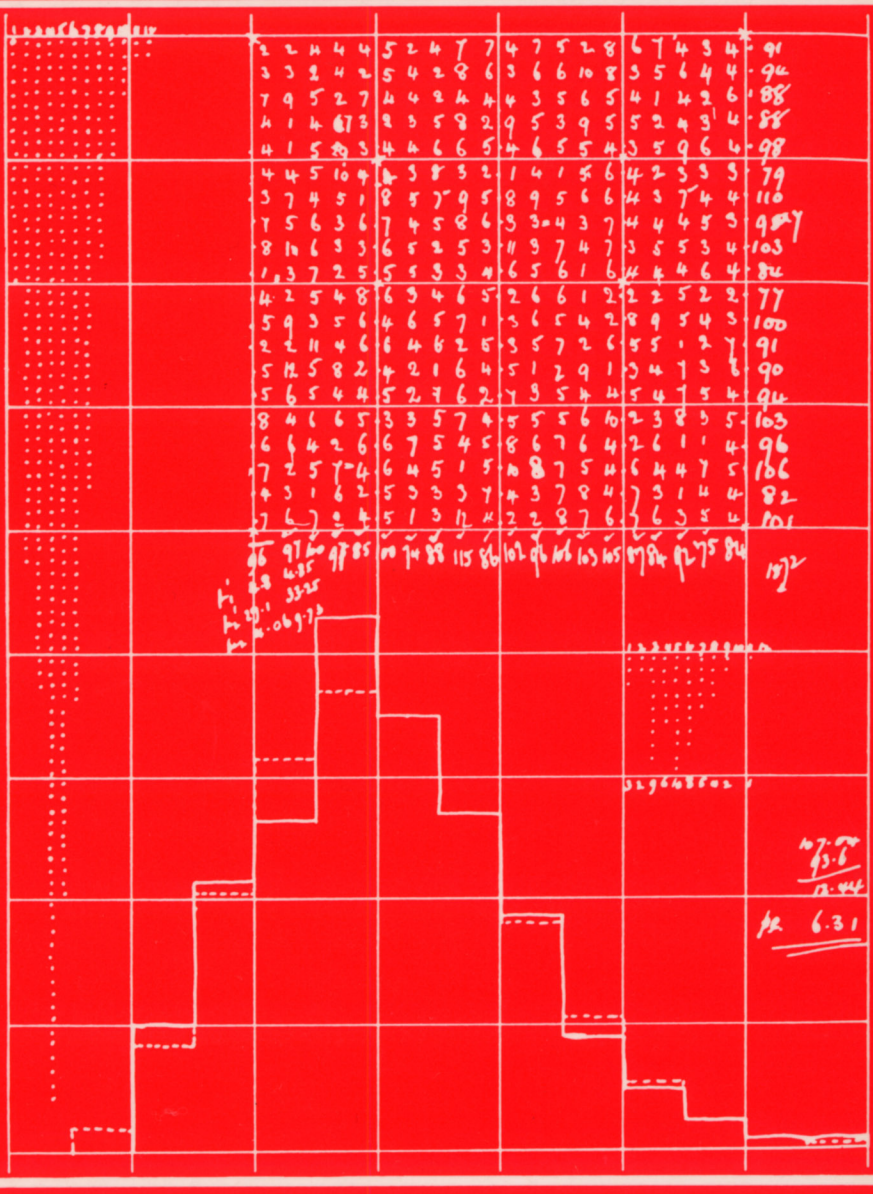
Student

Volume 2 Number 1

incorporating

Data & Statistics



March 1997

PAN

Presses Académiques Neuchâtel

The aim of ***Student*** is to present the statistical thoughts of graduate students to the community of researchers, teachers and practitioners of statistics. The journal welcomes the submission of articles in all areas of the theory, methodology and applications of statistics and probability, as well as papers on the history of statistics. Articles of an expository or tutorial nature are very much welcomed.

Submissions should be sent in two copies to the address below. It is assumed that the manuscripts have not previously been published and are not simultaneously being considered for publication elsewhere.

The language used must be either English or French. Manuscripts should be typed on one side of the paper in double spacing with wide margins. They should generally not exceed 20 pages including figures and tables. Papers must be accompanied by two summaries, in English and French respectively, each one at most 200 words, followed by about five key words.

In addition, if possible, a LaTeX input file should be sent by electronic mail to **student@seco.unine.ch**. If electronic mail is not feasible, a standard floppy disk may be submitted. Manuscripts submitted in the LaTeX format should use the "article" style and should not use any special macros. All papers are refereed.

Student is published twice a year. The annual subscription price of ***Student*** is SF 25.00 (postage included).

For advertising, subscription and general enquiries, please contact:

Groupe de Statistique
Pierre-à-Mazel 7
CH-2000 Neuchâtel
Switzerland
Phone: +41-32-718 13 80
Fax: +41-32-718 13 81

Edited by Statistics Group and published by the Presses Académiques Neuchâtel

---

Cover: Page from Gosset's (Student) notebook containing the analysis of Haemacytometer counts (page 370 of "Studies in the History of Statistics and Probability" (1970) edited by E.S. Pearson and M.G. Kendall). Reproduced by permission of Hodder and Stoughton publishers Ltd, London, England.

Student

Volume 2, Number 1 **March 1997**

Data & Statistics

Student (ISSN 1420-1011) incorporating **Data & Statistics** (ISSN 1420-3308), Volume 2, Number 1, March 1997. Published bi-annually by Statistics Group, University of Neuchâtel, Pierre-à-Mazel 7, CH-2000 Neuchâtel, Switzerland.

Student (1997), Vol.2, No.1, 44-46

Chikio Hayashi and Data Science

Chikio Hayashi was born on June 7, 1918, in Tokyo, Japan. He received a B.A. in Mathematics from the University of Tokyo in 1942, and a D.Sc. from the same University in 1954. From 1955 he was a Professor at the Institute of Statistical Mathematics (National Research Institute) where he became Director-General in 1974 and Professor Emeritus in 1986. From 1986 to 1991 he was Professor at the University of the Air. Since then he is Visiting Professor in 12 Japanese universities. In 1965, Chikio Hayashi received the NHK Broadcasting Culture Prize for his research in social survey and in theory of quantification. In 1977, he was named Honorary Fellow of the Royal Statistical Society. In 1982 he has been awarded the Purple Ribbon Medal from the Japanese Government.

Doctor Hayashi is the author or coauthor of 28 books in Japanese language in different fields of Statistics. He especially wrote books which treat on the Japanese National Character such as *Measurement of Japanese Mind* (1988) and *Structure of Japaneseness* (1996). He is also the author of books in English such as *Data Analysis for Comparative Social Research* (1992), *Treatise on Behaviormetrics* (1993) and *Quantification of Qualitative Data - Theory and Method* (1993).

Chikio Hayashi was the student of Professor Z. Suetuna of the Department of Mathematics of the Tokyo University. He was a senior colleague of two other famous Japanese statisticians G. Taguchi and H. Akaike who had the same teacher as him.

G. Taguchi is famous for his method for off-line quality control which brought a lot of admiration from practical industry including the United States (Taguchi, G. (1976), *An Introduction to Quality Control*, Central Japan Quality Control Association, Nagoya, Japan).

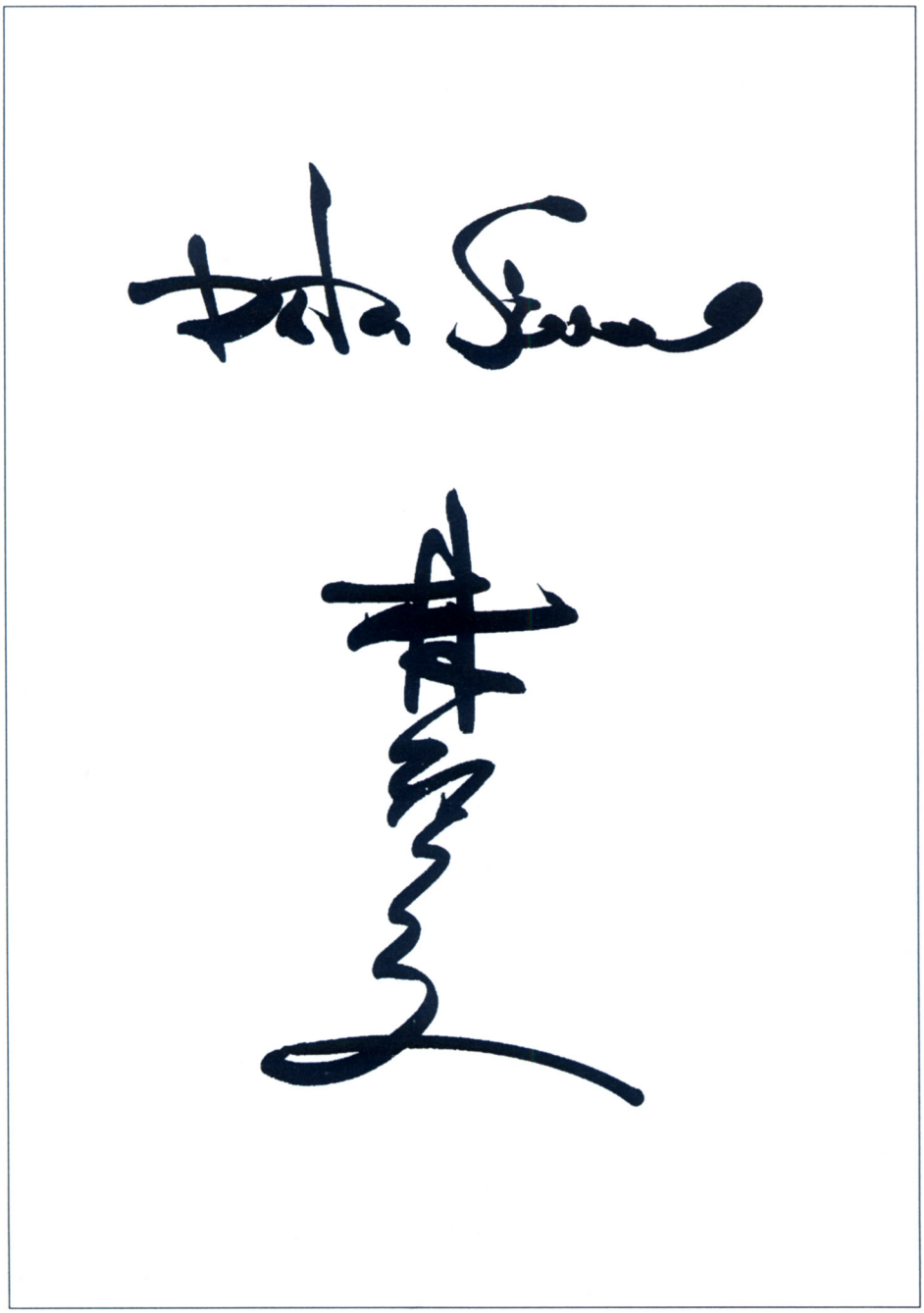
Hirotugu Akaike is the founder of a criterion which is used to decide on the order of a regression, where there is a natural sequence for introduction of successive predictor values, for example, ARIMA. An Information Criterion is defined by the relation

$$AIC = -2 \cdot \text{maximum log likelihood} + 2 \cdot \text{number of parameters}$$

(Akaike, H. (1973), "Information Theory and an Extension of the Maximum Likelihhod Principle" in *2nd International Symposium on Information Theory,* B.N. Petrov and F. Csaki eds., Akademiai Kiado, Budapest).

In what follows we provide the reader with Professor Hayashi's point of view on what is *Data Science* and how it is developed in Japan, a country for which we may not be familiar with.

Chikio Hayashi giving lecture during the IFC 1996 Conference, Kobe, Japan.



Chikio Hayashi writes on *Data Science*.

Student (1997), Vol.2, No.1, 47-51

What is Data Science ?

- Fundamental Concept -

By Chikio Hayashi

Institute of Statistical Mathematics

1. Introduction

Statistics and data analysis have developed in their realms separately. They contributed to the development of science, showing their properties. The idea and various methods of statistics were very useful as well-known and solved many problems. Mathematical statistics succeeded to it and improved new frontiers by the idea of statistical inference. Thus the elevation of view point brought us many useful results. On the other hand, the method of data analysis has developed in the fields disregarded by mathematical statistics and given useful results to solve complicated problems based on mathematico-statistical methods (which are not always on statistical inference but rather descriptive). As the differenciation proceeds with specialization, the innovation of useful methods of statistics and data analysis seems to disappear and a sign of stagnation appears. The reason is that the essential aim of analysis of phenomena by data has been forgotten. For extensive and profound development of intrisically useful methods of statistics and data analysis beyond present state, the necessity of unification of statistics and data analysis is felt acutely. For this purpose, the construction of a new point of view or a new concept is a crucial problem. So, I will present *Data Science* as a new concept.

2 Fundamental concept of data science

Data science is not only a synthetic concept to unify statistics, data analysis and their related methods on data, but also comprises its results. Data science intends to analyze and understand actual phenomena by *data*. In another expression, the aim of data science is to reveal the features or the

hidden structure of complicated natural, human and social phenomena by data from the different point of view from the established or traditional theory and method. This point of view implies multidimensional, dynamic and flexible ways of thinking.

Data Science consists of three phases: *design for data*, *collection of data* and *analysis on data*. It is important that three phases are treated in the concept of unification based on the fundamental philosophy of science explained as below. At these phases, the methods, which are fitted for the object and valid, must be studied in a good perspective. The strategy for research in data science through three phases is summarized in Figure 1.

Generally speaking, phenomena are multifarious. First, these phenomena are formulated and the planning of survey or experiment is devised, based on the idea of data science (phase of design for data). Thus phenomena are expressed as multidimensional and, frequently, time-series data. The characteristics or properties of these data are necessarily made clear (phase of collection of data). The obtained data are too complicated to draw clear conclusion. So, by methods of classification and multidimensional data analysis, and other mathematico-statistical methods, the data structure is revealed, in other words, simplification and conceptualization are carried out. However, this information generally turns out to be incomplete and unsatisfactory even though the structure finding was realized. At this stage, by finding and reconsidering the deviation of "individuals", which gives vivid account of roughness of conceptualization or simplification, from the mean values or class-belonging (classification) and structure, diversification of data is made. Based on this multifariousness, structure finding or conceptualization is attained, in an advanced sense, on the progressive stage. Such a circular movement of research always continues. Dynamics of both simplification or conceptualization and diversification begin in turn. Further, having been able to solve a problem, it is expected to find out another new problem to be solved in an advanced sense. The developmental process,

$$\text{design} \rightarrow \text{collection} \rightarrow \text{analysis} \rightarrow \text{design} \rightarrow \text{collection} \rightarrow \text{analysis} \rightarrow \text{design} \rightarrow \text{collection} \rightarrow \text{analysis} \rightarrow \text{design} \rightarrow \ldots.$$

in phase, and dynamic process mentioned above, that is to say, progress and regress are indispensable in data science. It is expressed that methodology of data science develops, as it were, in the ascending-spiral-process and research proceeds as seen in spiral stairs. The main point is schematically depicted in Figure 1.

Thus we can say that data science comprises not only the results themselves of theory and method but also all methodological results related to

various processes which are necessary to work out the results mentioned above. The former is called *hard results* and the latter is called *soft results.* Data science includes simultaneously hard and soft results. It goes without saying that an useful solution emerges in coping with the complicated problem in question by data science. It is repeatedly emphasized that the coherent idea through all items shown in Figure 1 flows in data science for the purpose of analysis of phenomena by data.

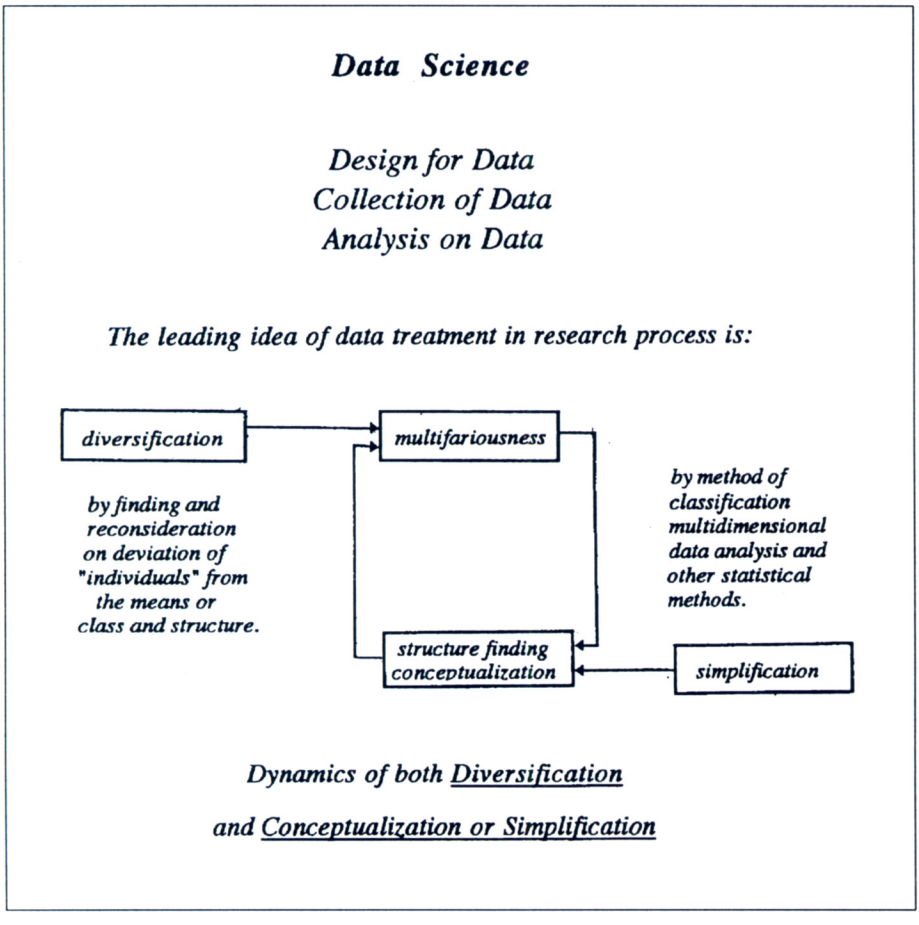


Figure 1: Strategy for research.

3 Content of data science

Some concrete examples in social and medical surveys for three phases are shown as below. Before everything, it is stressed that the relevant methods are always treated with validity.

3.1 How to design

The theory and method concerning this phase are considered. Particularly, theoretical and systematic construction of a questionnaire is a very important problem with the methods of observation and experiment. The problems in this phase are frequently solved using various kinds of methods of data analysis. For example,

- Sampling survey methods,
- Design of experiment,
- Evaluation of bias in quota sampling,
- New systematic idea of survey planning for the solution of difficult and complicated problem,
- Construction of questionnaire,

- Theory-driven (which is on the extension of hypothesis testing), Gutlman's Facet Theory,

- Data-driven (exploratory approach), Hayashi's cultural Link Analysis in comparative study,

- Device of various types of questions, for example, dynamic use of closed and open ended questions,
- Use of various projective methods,
- Evaluation of data quality and data characteristics,
- Randomized response method,
- Problem of translation in international comparative study, and etc.

3.2 How to collect

Collection of data is not only a problem of practice, but must be theoretically and concretely studied. The problems in this phase can not be solved without any information of design for data and any use of data analysis.

- Evaluation of survey bias and evaluation of experimental bias including question bias, interview bias, interviewer bias, observation bias and etc.,
- Evaluation of non-response error,
- Evaluation of measurement error,
- Evaluation of response error, inevitably variable response data, for example, live data,
- Method of diminution of the relevant bias and error, and etc.

3.3 How to analyze data

The problems in this phase are, of course, closely related to the previous two phases. The main point is to obtain useful and instrumental information without any distortion or with validity. For this purpose, clear and lucid methods of analysis without unnecessary mathematical conditions only for

model building and too sophisticated style are desirable. For example,

- Various methods of scaling, quantification methods, correspondence analysis (analyse des données), multidimensional scaling, exploratory data analysis, categorical data analysis and various methods of classification and clustering,
- Useful data analysis suitable for the purpose,
- Useful coding of questions and their synthesis,
- Valid analysis of data including various errors,
- Evaluation of data quality and data analysis depending on data quality,
- Analysis on probabilistic response,
- Exploratory approach by data analysis,
- Method of simultaneous realization of classification and structure finding,
- Treatment of open answers in an open ended question, for example, exploratory approach for coding or automatic processing of textual data,
- Probabilistic approach,
- Computer experiments, and etc.

These three phase must be synthetically treated or taken into consideration on the consistent idea in order to understand phenomena by data. This is the fundamental concept of data science. Of course, each subject will be studied separately. However, each subject must be studied in the content of data science. This idea will lead to development of statistics and data analysis in a new direction. Thus the stand points of them heighten and a new horizon will appear. Innovative method and theory will be created in three phases.

Chikio Hayashi

Tokyo, Japan
April, 1996.