Mem. Fac. Sci. Shimane Univ.,
27, pp. 81–90 Dec. 27, 1993

Design of a Statistical Expert System: STATEX

Yasuyuki KOBAYASHI

Department of Information Science, Faculty of Science,
Shimane University Matsue, 690 JAPAN
(Received September 6, 1993)

This paper describes the design of a statistical expert system STATEX. Statistical program in STATEX is considered as a sequence of capsules calculating statistics. The insides of the capsules are not open to statistical analysts but the definitions and the meanings of the statistics calculated by the capsules are open to them. STATEX can automatically generate statistical programs based on the numerical definition for each capsule and a sequence of capsules with loops and branches. STATEX has various statistical knowledges; purposes for statistical analysis, statistics, statistical methods, conditions for loops and branches, massages from statistical programs, responses of statistical analysts, and examples of statistical analyses. Statistical method developers store most of the statistical knowledge in order to develop statistical programs. Hence STATEX makes easy statistical knowledge acquisition. STATEX supports statistically naive users to analyze statistical data by the functions; selection of the optimum statistical method, explanation of statistics and statistical methods including the flow of a statistical program, and display of examples of statistical analyses. The function of automatic program generation in STATEX supports non-programers to develop/modify statistical programs without any programing tequnique. We have already developed the functions of method selection and automatic program generation.

1. Introduction

The development of statistical packages enables non-programers to analyze statistical data without much knowledge about statistics. But there are some problems in statistical packages. Statistical analysts often use statistical packages without enough understanding about statistical methods. Therefore they feel to be difficult to select the optimum statistical method for their purposes from many available statistical programs. They also mistake to understand the results of statistical analyses. Moreover, it is difficult for statistical method developers to include new statistical programs into statistical packages because they do not have much knowledge about computers.

In order to solve these problems, some researchers developed various statistical expert systems. Gale [2] developed REX which is an expert system for regression analysis. EXPLORA [3] is an expert system to roughly analyze large data sets and to extract interesting results from data. WAMASTEX [1], which is integrated into the SAS package, gives heuristic statistical guidance to clinical physicians in hospital departments. But these expert systems partially solved the above problems. There is

no discussion about statistical knowledge acquisition which is one of the most important feature in expert system.

In order to solve the above all problems, we design a statistical expert system STATEX. In statistical packages, statistical programs are considered as black boxes for statistical analysts. Statistical analysts only give few parameters for executing statistical programs. Then statistical programs calculate various statistics. In STATEX, statistical programs are considered as sequences of capsules calculating statistics. The insides of the capsules are not open to statistical analysts but the definitions and the meanings of statistics calculated by the capsules are open to them. Statistical analysts can execute the sequence of capsules defined by statistical method developers. They can also specify the next excuted capsule according to the result of each capsule.

STATEX manages various knowledges about statistical analysis; correspondence of statistical methods and purposes of statistical analysis, the sequence of statistics constructing statistical methods, the definition and the meaning of each statistic, and so on. These knowledges about statistical analysis are given by statistical method developers. Statistical programs realizing statistical methods are automatically generated by the sequence of statistics and the definition of each statistic.

Therefore STATEX can support statistical analysts not having much knowledge about statistics. Automatic statistical program generating function mekes easy to acquire statistical knowledge from statistical method developers.

2. Problems in statistical packages

In this section we discuss some problems for statistical analysts and statistical method developers in statistical packages.

2.1 Problems in statistical analysis

In the following, we analyze processes in statistical analysis and statistical knowledge used in statistical analysis.

(1) Clarification of purpose

We understand the real world which we try to analyze by using statistical methods and clarify our purpose. In this step, we need the knowledge about (a) effective statistical methods for our purposes and (b) examples of statistical analyses applied to similar purposes.

(2) Data collection

We collect data for the attainment of the clarified purpose. In this step, we need the knowledge about (c) the method of data collection for applicable statistical methods.

(3) Data input into computer

We input the collected data into computer and browse the summary of the data. In this step we need the knowledge about (d) the method for finding outliers and the

standard of judgement of them (e) variable transformations and (f) the precondition for statistical methods and the method for checking the precondition.

(4) Selection of statistical methods

We select effective statistical methods for the clarified purpose in (1). In this step we need the knowledge about (g) the definitions and the meanings of statistics, and results of testing statistical hypothesis, and (h) the models assumed in statistical methods, and (i) the correspondence of statistical methods and statistical programs.

(5) Execution of statistical programs

We execute statistical programs corresponding to the statistical methods selected in (4). In this step we need the knowledge about (j) the flow of processes in statistical programs, (k) the meanings of parameters for executing statistical programs, and (l) the methods to deal with the messages from statistical programs.

(6) Interpretation of results

We apprehend the statistical meanings of the results of statistical programs and judge whether the results are consistent with our experience in the real world. In this step we need the knowledge about (m) the statistical meanings of statistics and the results of testing statistical hypothesis.

Users of statistical packages usually must get the above mentioned knowledge from manuals and text books of statistics. Statistical packages output suitable results based on suitable parameters and data even if the selected method is not appropriate for user's purpose of the specified parameters are statistically wrong. Therefore users not having enough statistical knowledge often use unapplicable statistical methods for their purposes and mistake to understand the results of statistical analyses.

2.2 Problems in statistical method development

Statistical method developers can use various sub programs calculating eigen values and inverse matrix, which are prepared in statistical packages, in order to develop main programs corresponding to new statistical methods. However they must develop sub programs calculating new statistics. In order to develop these new programs, they must study the knowledge about computer system and numerical calculation. Even if a new statistical program is made up by the modification of a conventional statistical program, it is difficult to edit the conventional program without any knowledge about computer and numerical calculation.

3. Design of STATEX

In statistical packages, most of the statistical knowledges mentioned in 2.1 are embedded in statistical programs. Statistical analysts can get the result of statistical analysis by a few parameters for executing a statistical program. But they cannot know the flow of the process in the program without manuals.

In STATEX, programs calculating statistics or testing statistical hypothesis are

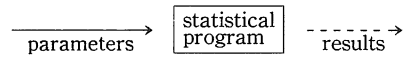

Figure 1. Statistical program in statistical packages.

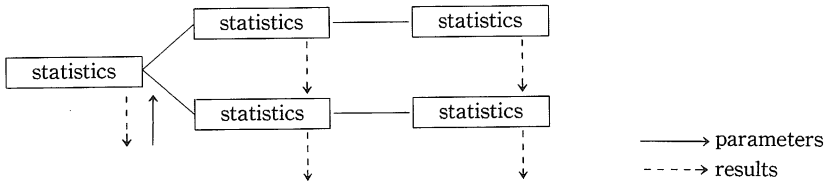

Figure 2. Statistical program in STATEX.

considered as capsules and a statistical program corresponding to a statistical method is defined as a sequence of capsules. So statistical analysts cannot see how the programs calculate statistics, which are insides of capsules, but they can see the numerical definition of statistics calculated in each capsule and the sequence of capsules realizing the statistical method. Moreover statistical analysts can also decide next executed capsule according to the results of previously executed capsules.

3.1 Design of files in STATEX

STATEX has the following files.

〈Purpose〉

Stevens [9] classified scales in statistical data into four types; nominal scale, ordinal scale, interval scale, and ratio scale. Nominal scale is used for identification of groups, such as kind and brand. There is only equal relation between groups. Ordinal scale gives ordinal relation between groups, such as rank of quality and level of taste. Nominal scale and ordinal scale are called qualitative scale. Interval scale defines interval or distance, such as temperature (Centigrade or Fahrenheit). Ratio scale is an interval scale having the absolute origin, such as weight and height. Interval scale and ratio scale are called quantitative scale. We can select a desirable statistical method corresponding to user's purpose and types of scale presenting data by using the hiarachy of purposes of statistical analysis in Appendix.

Purpose file manages the hierarchy of purposes of statistical analysis and statistical methods corresponding to the purposes.

〈Examples of statistical analysis〉

This file manages the flow of statistical analysis, the purpose of the statistical analysis, the used statistical methods, and the field of the real world such as agriculture, economics, and medicine.

〈Method〉

Method file has the fields, the purpose of statistical analysis achievable by the statistical method, the assumed statistical model and precondition, the desirable data

sampling method, the name of statistical program corresponding to the statistical method, and the sequence of statistics with some loops and branches.

〈Statistic〉

Statistic file manages statistics. Main fields in the file are the definition by numerical equation with predefined statistics and the meaning of statistic.

〈Program〉

We can define almost statistics by numerical equations. However it is difficult to define some complex statistics by numerical equations, such as eigen values and inverse matrix. We may directly develop programs calculating these statistics without the function of automatic program generation. Program file manages these programs by the fields, program name, purpose, input/output parameters and their meanings.

〈Condition〉

Sequences of capsules in STATEX have some loops and branches which are controlled by statistical analysts and the results of statistical calculations. Condition file manages these loops and branches with the fields, condition (COND) and next process (PROC). In COND field, we may record conditions for loops and branches with numerical equations. In PROC field, we may specify "message number" (display the specified message), EXIT (continue next process defined in Method file), and RETURN "process number" (return to the specified process in Method file).

〈Message〉

Message file manages messages to statistical analysts which are the meanings of the values of statistics and the results of testing statistical hypothesis, and questions for deciding next process, such as "Data are skewed. Do you want any transformation?".

〈Response〉

Response file manages statistical analyst's responses (RESPONSE), its meanings (MEAN), and next processes (PROCESS). For example, RESPONSE='Y', MEAN ='transformation of data' and PROCESS='trans (X): EXIT' imply that; if you want to transform data then you may input 'Y'. Statistical program continues next process defined in statistical method after transformation of data X. In PROCESS field, we may record "program name" (execute the specified program), EXIT (continue next process defined in Method file), "message number" (display the specified message), and RUTURN "process number" (return to specified process in Method file).

3.2 Functions of STATEX

We show the summary of STATEX in Figure 3. Statistical data and statistical knowledge are managed by a database management system (DBMS). The functions in the right side are usually used by statistical method developers. Statistical analysts usually use the functions in the left side. But they sometimes use the functions in the right side in order to modify statistical methods. We show the functions of STATEX in the following.

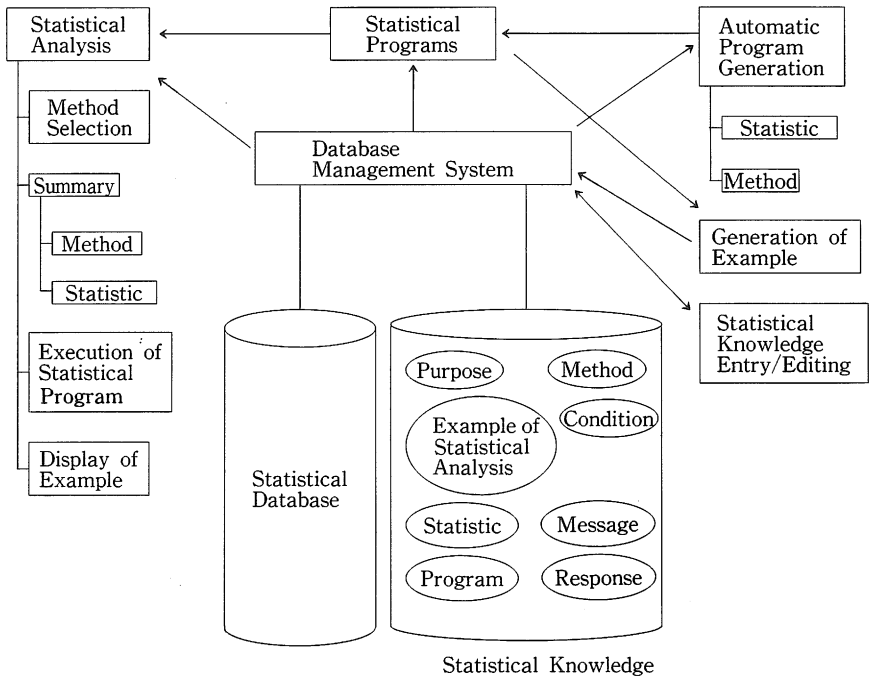

Figure 3. Summary of STATEX.

〈Method selection〉

STATEX suggests a desirable statistical method on the basis of statistical analyst's purpose and some characteristics of data. We show an example in Figure 4. Each underlined string of characters means a parameter specified by a user.

1. Prediction of observed variables (Y) from observed variables (X).
2. Prediction of unobserved variables (Y) from observed variables (X).
3. Comprehension of relation between two groups of variables, X and Y.
4. Representation of data X with fewer variables Y.
5. Comprehension of the information about distribution of variables X.

What purpose do you have? 1

1 MATHT	2 MATWT	3 FATHT	4 FATWT
5 MATAGE	6 FATAGE	7 BABYHT	8 BABYWT

Specify X……1, 2, 3, 4, 5, 6

Specify Y……7, 8

1. Predictor equation is linear
2. Predictor equation is not linear

Which do you select? 1

Multiple regression analysis is effective for your purpose. Do you execute? Y

Figure 4. Example of Method selection

STATEX searches Purpose file with the parameters specified by a user and the types of scale of the used variables.

〈Summary of statistical method/statistic〉

STATEX displays numerical equation and the meaning of statistic, and the meanings of the values of statistic. STATEX also displays the flow of process, model, achievable purpose, and data sampling methods effective for statistical methods.

〈Execution of statistical program〉

While a statistical program is executing, STATEX also supports a statistical analyst to interpret the meanings of the values of statistics and to decide the next process by Condition file, Message file, and Response file.

〈Generation and display of examples of statistical analysis〉

STATEX can record the logging of statistical analysis and display it as an example of statistical analysis according to the specified statistical method or purpose.

〈Statistical knowledge entry/editing〉

This function supports statistical method developers to entry and edit statistical knowledge except for examples of statistical analysis.

〈Automatic generation of statistical programs〉

STATEX automatically generates programs which can calculate statistics defined by numerical equations. STATEX also automatically generates statistical programs corrospoinding to the statistical methods defined by sequences of statistics.

4. Features of STATEX

STATEX has the following features.

〈Implementation of a statistical expert system〉

STATEX has much statistical knowledge in manuals of statistical packages and text books of statistics. STATEX also has the flow of process in statistical programs as records in files. So statistical analysts can see the statistical knowledge at any time.

STATEX can easily acquire the statistical knowledge because statistical method developers store it to generate statistical programs.

〈Support for developing statistical programs〉

Statistical method developers may give numerical equation of statistic, sequence of statistics, message to statistical analyst and so on, in order to develop statistical programs. Therefore when they develop a new statistical program by editing the partial of a conventional statistical program, they may edit the stored information corresponding to the conventional statistical program. STATEX automaticlly generates a new statistical program by using the editted information. Moreover if a statistical analyst want to calculate some other statistics in a statistical program, he may add numerical equations for the statistics to Statistic file.

〈Easiness in changing messages〉

In a statistical package we must directly edit a statistical program to change some

messages from the statistical program. But we can freely change any message from statistical program in STATEX by editing Message file. Therefore it is easy to display messages according to the level of statistical knowledge of statistical analysts and the fields of the real world to which data belong, such as agriculture, economics, and medicine.

5. Conclusion

We have designed the statistical expert system STATEX supporting statistical analysts not having much statistical knowledge. The function automatically generating statistical programs makes easy to acquire statistical knowledge from statistical method developers. We have already developed the function selecting statistical method for statistical analyst's purpose. We also have developed the function automatically generating statistical programs [5] and have generated the similar program to regression analysis program by Tarumi and Tanaka [10]. We will try to generate more complex statistical programs and implement STATEX based on our design.

References

[1] DORAN, W., FROESCHL, K. A. and GROSSMAN, W. WAMASTEX-Heuristic Guidance for Statistical Analysis, COMPSTAT, (1990), 93–98.
[2] GALE, W. A. Artificial Intelligence & Statistics, Addison Wesley (1986).
[3] GEBHARDT, F. An Expert System Strategy for Selecting Interesting Results, COMPSTAT, (1990), 81–85.
[4] KOBAYASHI, Y. Data Dictionary/Directory System of a Statistical Database, J. Inf. Process., **7**, 4 (1984), 233–239.
[5] KOBAYASHI, Y. and KAJITANI, Y. An Automatic Generator System of Statistical Analysis Programs, Bulletin of the Computational Statistics of Japan, **4**, 2 (1991), 3–14.
[6] OKUNO, C. et al. Multivariate Analysis, Nikkagiren (1971).
[7] SAWA, T. Regression Analysis, Asakura Shoten (1979).
[8] SNEDECOR, G. W. and COCHRAN, W. G. Statistical Methods, The Iowa State University Press (1967).
[9] STEVENS, S. S. Mathematics, Measurement, and Psychophysics, in S. S. Stevens (ed.), Handbook Experimental Psychology, John Wiley (1951).
[10] TANAKA, Y., TARUMI, T. AND WAKIMOTO, K. Handbook of Statistical Analysis with Programs for Personal Computers, 2 (1984), Multivariate Analysis, Kyoritsu Shuppan.

Appendix

1. Prediction
1.1 All variables are observable.
1.1.1 Prediction of criterion variables with quantitative scale from predictor variables with qualitative scale

(We denote "Qualitative scale⇒Quantitative scale").

1.1.1.1 The number of criterion variables is one.

1.1.1.1.1 The number of predictor variables is one.

〈One way layout〉

1.1.1.1.2 The others

〈Analysis of variance〉

1.1.1.2 The number of criterion variables is more than one.

〈Multivariate analysis of variance〉

1.1.2 Qualitative scale⇒Qualitative scale

〈Multiple contingency table〉

1.1.3 Qualitative scale and Quantitative scale

⇒Quantitative scale

〈Analysis of covariance, Multiple regression analysis〉

1.1.4 Qualitative scale and Quantitative scale

⇒Qualitative scale

〈Multiple contingency talbe〉

1.1.5 Quantiative scale⇒Quantiative scale

〈Discriminant analysis〉

1.1.6 Quantiative scale⇒Quantiative scale

1.1.6.1 Predictor is linear.

1.1.6.1.1 The number of criterion variables is one.

〈Simple regression analysis〉

1.1.6.1.2 The number of criterion variables is more than one.

〈Multiple regression analysis〉

1.1.6.2 Predictor is not linear.

1.1.6.2.1 Predictor is polynomial.

〈Polynomial regression analysis〉

1.1.6.2.2 Predictor is not polynomial.

〈Curve regression analysis〉

1.2 Criterion variables are not observable.

1.2.1 Quantitative scale⇒Quantitative scale

〈Factor analysis〉

1.2.2 Quantitative scale⇒Qualitative scale

〈Cluster analysis〉

1.2.3 Quantitative scale⇒Qualitative scale

〈Cluster analysis〉

1.2.4 Qualitative scale⇒Quantitative scale

〈Latent structure analysis〉

2. Comprehension of the present state

2.1 Comprehension of relation between two groups

2.1.1 Relation between a group of variables with quantitative scale and a group of

variables with qualitative scale
(We denote "(Quantitative scale, Qualitative scale)")
2.1.1.1 Each variable with qualitative scale has only two values.
2.1.1.1.1 Difference between means of variables with quantitative scale
〈T-test〉
2.1.1.1.2 Difference between variances of variables with quantitative scale
〈F-test〉
2.1.1.2 Some variables with qualitative scale have more than two values.
2.1.1.2.1 Difference between means of variables with quantitative scale
〈Analysis of variance〉
2.1.2 (Quantitative scale, Quantitative scale)
〈Correlation analysis〉
2.1.3 (Qualitative scale, Qualitative scale)
2.1.3.1 (Ordinal scale, Ordinal scale)
〈Coefficient of rank correlation (Spearman/Kendall)〉
2.1.3.2 One group has ordinal scale and each variable in the other group has only two values.
〈Sign test〉
2.1.3.3 The others
〈Test of independency in Multiple contingency table〉
2.2 Presentation of data with fewer variables
2.2.1 Analysis as one group
2.2.1.1 All variables have quantitative scale.
〈Principal component analysis〉
2.2.1.2 All variables have qualitative scale.
〈Guttman scale analysis〉
2.2.2 Analysis as two groups
2.2.2.1 All variables have quantitative scale.
〈Canonical correlation analysis〉
2.3 Comprehension of the information about distributions of a variable
2.3.1 Comparison with theoretical distributions
2.3.1.1 Analysis of the variable with qualitative scale.
〈Chi square test in One way classification〉
2.3.2 Estimation of unknown parameters of distribution function
2.3.2.1 One point
〈Point estimation〉
2.3.2.2 Interval
〈Interval estimation〉