Data-oriented Learning System of Statistics based on Analysis Scenario/Story (DoLStat)

Yuichi Mori
Department of Socio-Information, Okayama University of Science
1-1 Ridai-cho, Okayama 700-0005, Japan
mori@soci.ous.ac.jp
Yoshiro Yamamoto
School of Management and Information Sciences, Tama University
4-1-1 Hijirigaoka, Tama 206-0022, Japan
yama@tama.ac.jp
Hiroshi Yadohisa
Department of Mathematics and Computer Science, Kagoshima University
Kagoshima 890-0065, Japan
yado@sci.kagoshima-u.ac.jp

1. Introduction

Statistics are widely used in a variety of research and business areas, and the advent of the computer has helped to simplify the application of statistical techniques to real data sets. Recently, the number of individuals who are interested in learning how to use data analysis to solve real-world problems has increased at a rapid pace. Therefore, a quality statistics education is essential. Although there are many good textbooks, statistical classes, and e-learning systems, we often meet the following problems. First, there are often not enough examples, and the existing examples rarely relate to the students' interests. Also, traditional textbooks present the material in a systematic order, building upon previous statistical knowledge and methods. The students learn what is the method well through the material, but often don't know how to apply the statistical techniques and methods appropriately in real situations. In order to mitigate these method-oriented education problems, a databank should be established to house a large number of data sets. Furthermore, it is desirable that the databank provides guidance on how to practically analyze the data, how to interpret the results, and how to use a general package to obtain the desired information. This type of approach is referred to as a "data-oriented" education.

Currently, there is an abundance of web sites that provide free or semi-free access to collected data and related information. Chance Database (Dartmouth College, [1]), Data and Story Library (DASL, Cornell University, [2]), Statlib (Carnegie Mellon University, [3]), and Data Representation System (DRS, Inoue et al., 2002), among others, could prove useful in providing statistical education. DASL and DRS, in particular, are valuable sites for data-oriented education. DASL contains both a large number of real data sets and the information on the analysis of the data (analysis stories), and DRS includes an interactive system that performs an immediate online analysis of any subset of data in the database. These types of examples are useful for educational purposes because they provide an analysis scenario/story and an interactive analysis system over the Internet. The analysis scenario/story, which includes a document or record that describes the actual process used in the original analysis, is especially necessary for data-oriented education because this information takes on the role of the manual or chart typical in a similar situation to original one.

Ideas from these two approaches are being used to develop a data-oriented learning system over the Internet called DoLStat@*d* (Data-oriented Learning system of Statistics based on analysis scenario/story). This system provides multiple courses. Each course includes real data sets and their analysis scenarios/stories. The data sets used in a course are selected from a database (DoDStat@*d*, Data-oriented Database of Statistics based on analysis scenario/story) and are educationally ordered according to the purpose of the course. An online analysis, based on the story, can be performed on the data set (DoAStat@*d*, Data-oriented Analysis system of Statistics based on analysis scenario/story). This is a part of DoSS@*d* (Data-oriented Statistical System based on analysis scenario/story) mentioned in the next section.

2. DoSS@*d* (Data oriented Statistical System based on analysis scenario/story)

DoSS@*d* is located at `http://mo161.soci.ous.ac.jp/@d/index.html` and consists of three modules: DoDStat@*d*, DoAStat@*d* and DoLStat@*d*, as shown in Figure 1. DoDStat@*d* is a database of real data sets. Each data set includes a body, in csv format, attributes (e.g. case names, variable names and variable types) and analysis stories. The users are able to select an interesting or appropriate data set using a retrieval key, such as a subject, method or keyword, on the built-in search interface. DoAStat@*d* provides a computational environment with a server-side statistical engine (R or XploRe Quantlet Server). The users are able to analyze any data set stored in DoDStat@*d* as well as their own data sets. DoLStat@*d* is a learning system that utilizes the other two modules. Figure 2 depicts the structure of DoSS@*d* and the relationship between the modules.

Data oriented Statistical System based on Analysis Scenario/Story - Microsoft Internet Explorer

DoSS@*d*

Data oriented Statistical System based on Analysis Scenario/Story

English | Japanese

Welcome to DoSS@d !

DoDStat@*d* Data-oriented Database of Statistics based on Analysis Scenario/Story

DoAStat@*d* Data-oriented Analysis System of Statistics based on Analysis Scenario/Story

DoLStat@*d* Data-oriented Learning System of Statistics based on Analysis Scenario/Story

Figure 1. Top page of DoSS@d

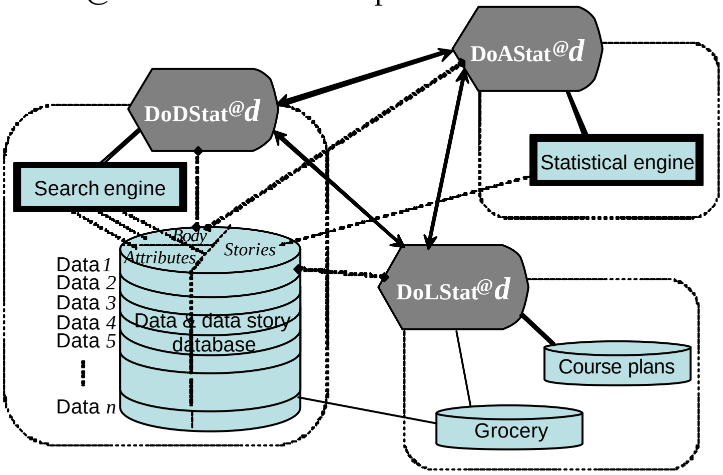


Figure 2. Structure of DoSS@d

3. DoLStat@*d* (Data-oriented Learning system of Statistics based on analysis scenario/story)

DoLStat@*d* contains several different courses that utilize the data sets and analysis stories in DoDStat@*d* and employ the computational functions in DoAStat@*d*. Each course has a specific educational purpose and contains five to ten analysis stories arranged in a suitable educational order, according to the lesson's purpose. The DoLStat@*d* courses are classified into four major categories, as shown in Figure 3:

- A general statistics course (e.g. introduction to statistics, advanced statistics, introduction to multivariate analysis, etc.)
- Statistics appropriate to a specific field (e.g. marketing, economics, biometrics, etc.)
- Statistics techniques (e.g. summarization, visualization, prediction, dimension reduction, etc.)
- Statistical methods (e.g. regression analysis, cluster analysis, principal component analysis/correspondence analysis, factor analysis, etc.)

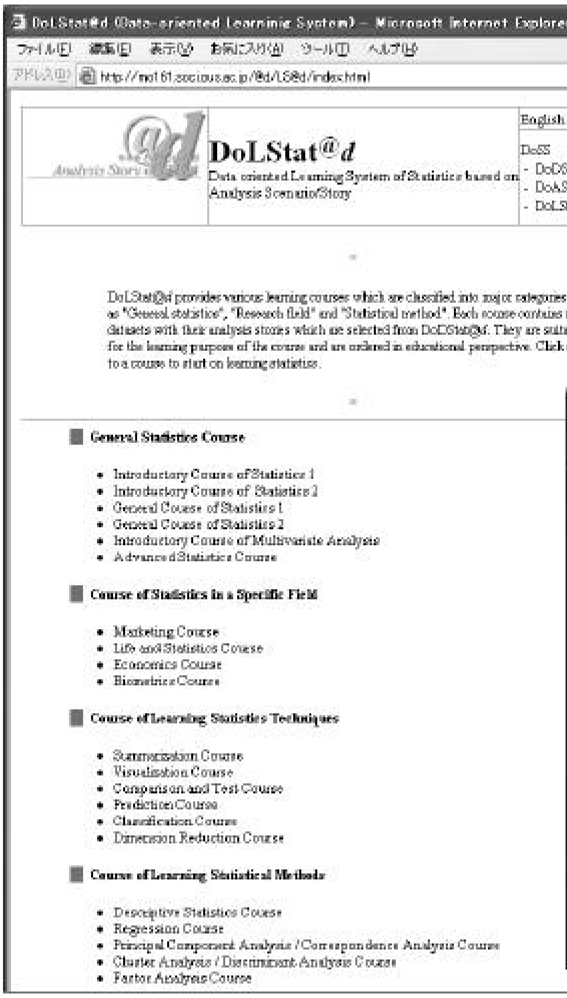
Several courses may utilize the same data set. For example, the "Physical measurement of alate adelges" data set was analyzed with principal component analysis (PCA), so it can be used in "Introductory course of multivariate analysis," "Principal component analysis course," and "Dimension reduction course." Then, within each course the teacher/student can select lessons appropriate for his educational purpose.

When a course is selected, the course contents page appears. Figure 4 shows the course contents page for "Introductory Course of Multivariate Analysis." The students learn the units of material starting from the top of the page. When a unit is selected, the data set analysis story page is displayed, which includes the following items:

- Title: The title of the analysis story
- Goal: The goal and aim of the data analysis
- Data: The name of the data set that is linked to the description page, a link to the data body in csv format, and the case and variable details
- Research subject: The field in which the data was collected and analyzed
- Statistical method: The method that was applied to the data
- Source: The source or reference of the story, if it exists
- Analysis story: The document that describes how the data was analyzed in practice.

This page is automatically generated from an XML file of the story.

Figure 5 shows the story page for "Physical measurement of alate adelges." From this information, the students can learn that the aim of the "Alate adelges" data analysis was to identify sub-groups in which individuals have similar characteristics based on their global scores of measured values. They will also discover that PCA was applied to the data. The students will learn how the data was analyzed and the results were interpreted during the actual process, which is described in the Analysis story item. There are two types of link buttons at the bottom of the page. The [Analysis] button allows the students to perform an online data analysis using the same statistical method and parameters as the original analysis. When the [Analysis] button is clicked in the "Alate adelges" example, a PCA interface appears, as shown in Figure 6. This is the same interface as in DoAStat@d, but the initial analysis parameters, such as the matrix type and the number of components, have been automatically input from the XML story file (see Figure 6). This system in this example was created using JAVA technology with the XploRe Qunatlet Server though MD*Crypt (MD*Tech, [4]). Using the interface, the students can immediately analyze the data according to the original story and can observe the results when the analysis parameters are changed. The other link buttons [SPSS],[R] and [XploRe] link to description pages that explain how to use the corresponding package to analyze the data. If the students desire to use SPSS, R or XploRe, they can obtain macros or functions with documentation for the package by selecting the link.


DoLStat@d (Data-oriented Learninig System) – Microsoft Internet Explorer
ファイル(F) 編集(E) 表示(V) お気に入り(A) ツール(T) ヘルプ(H)
アドレス(D) http://mo161.socious.ac.jp/@d/LS@d/index.html
移動
English | Japanese
DoSS
- DoDStat@d
- DoAStat@d
- DoLStat@d

DoLStat@d

Data oriented Learning System of Statistics based on Analysis Scenario/Story

DoLStat@d provides various learning courses which are classified into major categories such as "General statistics", "Research field" and "Statistical method". Each course contains real datasets with their analysis stories which are selected from DoDStat@d. They are suitable for the learning purpose of the course and are ordered in educational perspective. Click a link to a course to start on learning statistics.

General Statistics Course

- Introductory Course of Statistics 1
- Introductory Course of Statistics 2
- General Course of Statistics 1
- General Course of Statistics 2
- Introductory Course of Multivariate Analysis
- Advanced Statistics Course

Course of Statistics in a Specific Field

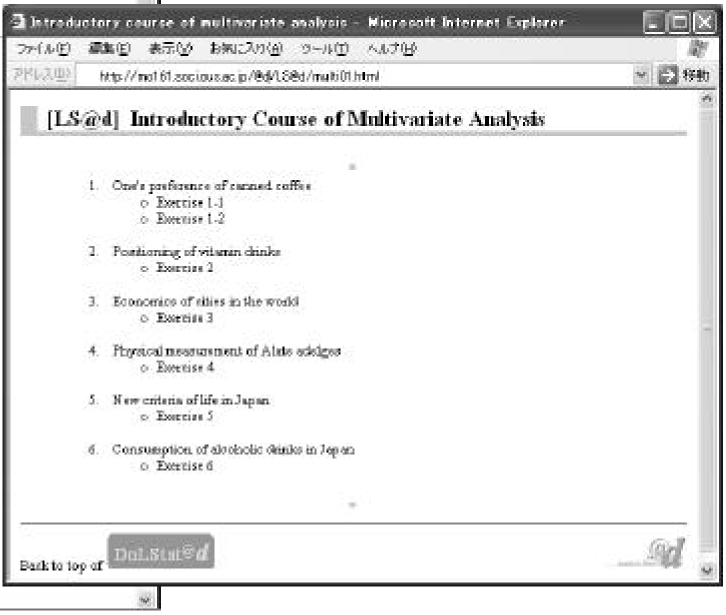
- Marketing Course
- Life and Statistics Course
- Economics Course
- Biometrics Course

Course of Learning Statistics Techniques

- Summarization Course
- Visualization Course
- Comparison and Test Course
- Prediction Course
- Classification Course
- Dimension Reduction Course

Course of Learning Statistical Methods

- Descriptive Statistics Course
- Regression Course
- Principal Component Analysis / Correspondence Analysis Course
- Cluster Analysis / Discriminant Analysis Course
- Factor Analysis Course


Introductory course of multivariate analysis – Microsoft Internet Explorer
ファイル(F) 編集(E) 表示(V) お気に入り(A) ツール(T) ヘルプ(H)
アドレス(D) http://mo161.socious.ac.jp/@d/LS@d/multi01.html
移動

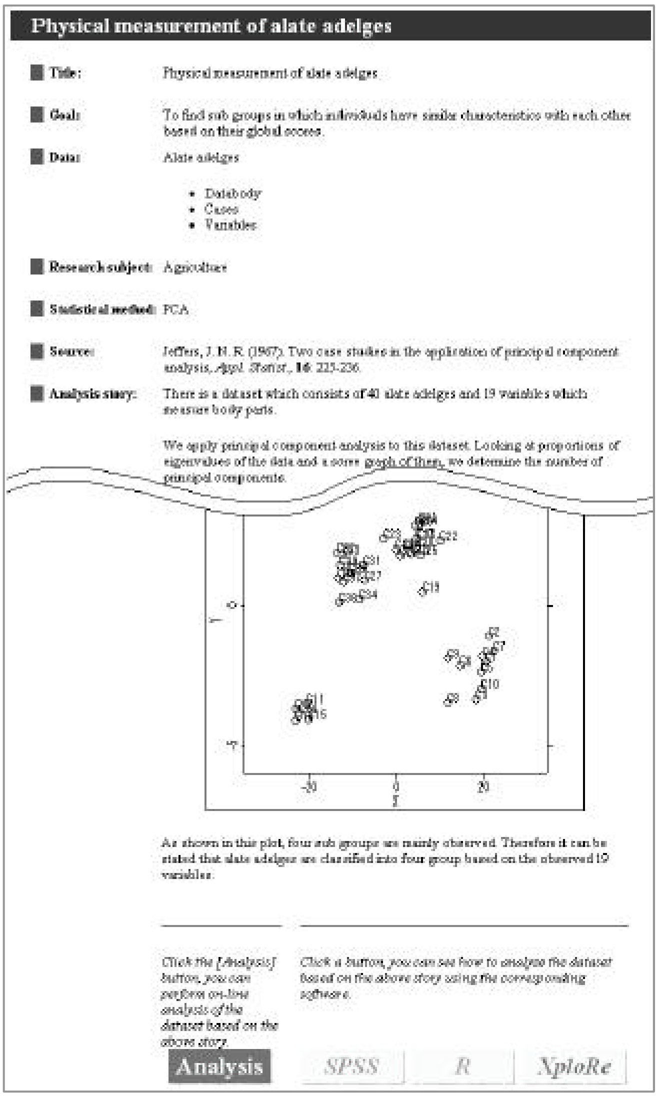
[LS@d] Introductory Course of Multivariate Analysis

1. One's preference of canned coffee
   - ○ Exercise 1-1
   - ○ Exercise 1-2
2. Positioning of vitamin drinks
   - ○ Exercise 2
3. Economics of cities in the world
   - ○ Exercise 3
4. Physical measurement of Alate adelges
   - ○ Exercise 4
5. New criteria of life in Japan
   - ○ Exercise 5
6. Consumption of alcoholic drinks in Japan
   - ○ Exercise 6

Back to top of DoLStat@d

Figure 3. Top page of DoLStat@d (Current courses classified into appropriate categories)

Figure 4. Contents page of "Introductory Course of Multivariate Analysis"



Physical measurement of alate adelges

- Title: Physical measurement of alate adelges
- Goal: To find sub groups in which individuals have similar characteristics with each other based on their global access.
- Data: Alate adelges
  - Databody
  - Cases
  - Variables
- Research subject: Agriculture
- Statistical method: PCA
- Source: Jeffers, J. N. R. (1967). Two case studies in the application of principal component analysis, *Appl. Statist.*, **16**: 225-236.
- Analysis story: There is a dataset which consists of 40 alate adelges and 19 variables which measure body parts.

We apply principal component analysis to this dataset. Looking at proportions of eigenvalues of the data and a scree graph of them, we determine the number of principal components.

As shown in this plot, four sub groups are mainly observed. Therefore it can be stated that alate adelges are classified into four group based on the observed 19 variables.

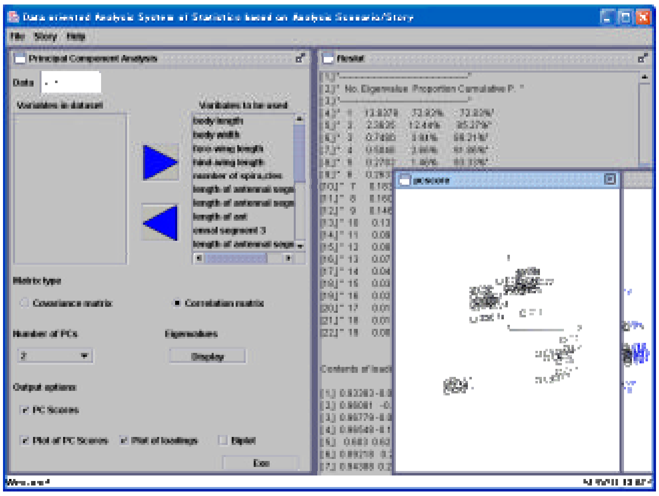
Click the [Analysis] button, you can perform on-line analysis of the dataset based on the above story.

Click a button, you can see how to analyze the dataset based on the above story using the corresponding software.

Figure 5. Analysis story page (Physical measurement of alate adelges)

4. Concluding Remarks

DoLStat@*d* is a web-based statistical educational system that combines real-world data and its analysis story with an online interactive analysis function. There are a number of advantages in constructing an Internet-based system, such as universal accessibility, easy maintenance, and instant updates. This system was developed to educate people in the area of statistical analysis using a data-oriented approach, however, method-based education also continues to provide a important role in statistical education. This system can be implemented either as a self-contained course or as supplementary material in a standard statistics course.



Data-oriented Analysis System of Statistics based on Analysis Scenario/Story
File
Story
Help
Principal Component Analysis
Result
pcscore

Data
Variables in dataset
Variables to be used: body length, body width, fore wing length, hind wing length, number of spiracles, length of antennal segment, length of antennal segment, length of ant, antennal segment 3, length of antennal segment
Matrix type: ○ Covariance matrix ● Correlation matrix
Number of PCs: 2
Eigenvalues: Display
Output options: ☑ PC Scores
☑ Plot of PC Scores ☑ Plot of loadings □ Biplot
Exe

```
[2]" No. Eigenvalue Proportion Cumulative P. "
[4]"  1  13.8278   72.82%   72.82%"
[5]"  2   2.3635   12.44%   85.27%"
[6]"  3   0.7490    3.91%   89.21%"
[7]"  4   0.5848    2.90%   91.90%"
[8]"  5   0.2702    1.40%   93.32%"
```

Figure 6. Online analysis interface (PCA)

REFERENCES

Inoue, T., Asahi, Y., Yadohisa, H. and Yamamoto, Y. (2002). A statistical data representation system on the Web. Comp. Statist. Special Issue, Springer.

[1] Chance database Home Page, http://www.dartmouth.edu/~chance/

[2] The Data and Story Library Home Page, http://lib.stat.cmu.edu/DASL/

[3] The StatLib Home Page, http://lib.stat.cmu.edu/

[4] MD*Tech, http://www.mdtech.de

RÉSUMÉ

DoLStat@d(Data-oriented Learning system of Statistics base on analysis scenario/story) est un web-base système d'étude basé sur des données. Il fournit plusieurs cours d'étude qui contient des données avec des histoires d'analyse. Il a également le système interactif d'analyse. On s'attend à ce que ce système rende beaucoup de peuples forts dans l'analyse de données.