放送大学研究年報　第6号（1988）113-119頁
Journal of the University of the Air, No. 6 (1988) pp. 113-119

Principles and Strategy of Data Analysis*[1)]

Chikio HAYASHI

データ解析の「基本概念と戦略」

林　知己夫

要　旨

本論文はデータ解析の根本的なあり方を述べたものである．ここに，データ解析というとき，これは単なるデータ処理を意味するものではなく，「データによる現象解析」を志向するものである．したがって，その根本理念と戦略について考察する必要がある．本論文はこの点について議論を進めている．

まず，データによる以上，データをどのように計画してとるか（design of data），どのように実際にデータを獲得するか（collection of data），どのようにデータを分析するか（analysis of data）の三相について考察する必要があり，この三相がどのように繰返されて現象が解明されて行くかの過程が説明されている．繰返し原点に戻りつつ進むという上昇螺旋的な行き方が論じられている．また，データの分析にあたっては，特に重要な三つのトピック，(i) 調査誤差の評価と誤差あるデータの分析方法，(ii) 多次元データ解析の方法，(iii) 質的データを数量化する方法についての基本的思想が論じられている．

As many authors point out, the definitions of "data analysis" are diversified, because the concepts and contents of data analysis are multifaceted. In the present paper, I should like to show my definition and mention about data analysis from a slightly different point of view from the others [1, 2, 3, 4, 5, 6, 7].

My definition is that data analysis means to analyze the multifarious and complex phenomena by data. It comprises not only the methodological results of data analysis but also the process of analyzing data, including : both the ideas and other all activities to produce the results of data analysis. The methods of data analysis are not a simply statistical instrument. Methods of useful data analysis are grounded in the logic, methodology and philosophy of science. Particulaly, I should like to emphasize that the system of data analysis, which, I call data science, will evolve out of experience with data itself. The way of thinking, that there exist some theories and then they are

* 1) This paper was read at the invited session "Principles and Strategy of Date Analysis" of the 46-th Session (Tokyo) of International Statistical Institute, September, 1987.

applied in order to analyze the data in hand, is not recognizable. Theories and applications must be unified in data analysis.

Here, the process of data analysis will be explained. Data analysis has three phases, i. e. design of data, collection of data and analysis of data. First, I mention about the problem of the former two phases as Table 1.

Table 1

A. Design of Data and Collection of Data
  1. In case where we conduct a survey,
    a. What is our universe?
      What is our population?
      Design of sampling
      Collection of data
    b. Numerical evaluation of the properties of the data,
      Sampling error
      Non-sampling error
        Non-response error
        Errors based on survey methods
        Other errors in survey practice
        Response error or variability — in methods of measurement : in sample reactions by telling a lie or by inevitable fluctuation of response inherent in an individual
  2. In case of existing data in hand,
    a. What is the universe of the data?
      What is the population of the data?
       (Is the universe reasonable for our object of analysis?)
      What kind of sample are the data?
      What characteristics and properties have the data?

Data analysis is a continuous series, for a goal, of the process shown in Table 2. I call this flow an "ascending spiral process of research" shown in Fig. 1.

The evaluation of the data (including errors or fluctuation) may be always incomplete. Then it is indispensable in data analysis to treat the data carefully from diversified aspects, to approach a problem with flexible attitude and to avoid any decisive attitude. The properties inherent in the data are always to be taken into consideration in those three phases together with the background surrounding the data. Fundamental conception of data analysis (logic, methodology and philosophy of data analysis) plays a particulary important role and provides the guiding concepts for both the development of method, or theory, and the design of computer software or sometimes hardware for data processing.

Table 2

Data analysis is a continuous series, for a goal, of
design of data
collection of data
analysis of data
to analyze the data, in the most suitable ways for the object, and in various multifaceted and dynamic askects, depending on the properties of the data.
to solve a problem
to find another new problems
to obtain some informations
to set new hypotheses

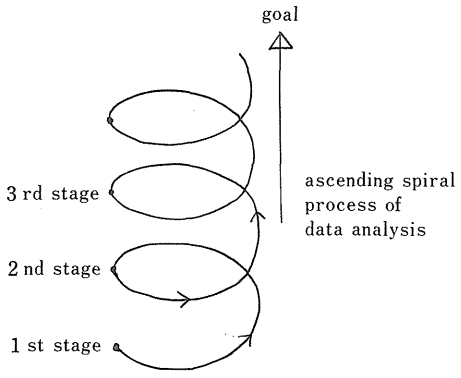

Fig. 1

It is difficult to discuss, in a short space, the details of main points of concrete process of data analysis in my sense. The wisest way may be to explain heuristically some methodological problems in the process of data analysis we meet frequently. So, I should like to take up the following three methodological topics of data analysis,

I. Quality of data,
II. Multidimensional statistical data analysis,
III. Treatment of qualitative data.

I : Quality of data

I have already touched on the various types of errors or fluctuations of data. So, I talk about the necessity of evoluation of them (quality of data) except sampling error, see Table 3.

Table 3 Quality of data (except sampling error)

evaluation of errors in data	* systematic errors	* Deviation of zero-point in measurement → bias
		* Other biases
	* variability or fluctuation of response being usually expressed in probability → bias and random error	

If the measurement is numerical value X in one dimension, it can be expressed as a random variable,

$$X = Xo + a + \varepsilon$$

(true, bias, error or variability or fluctuation; random variable)

$$E(\varepsilon) = \bar{\varepsilon} : \text{bias}$$

$$E(\varepsilon - \bar{\varepsilon})^2 = \sigma_\varepsilon^2.$$

Daviation of zero-point and other biases must be of course evaluated in fundamental research. Even though there is not any bias (including $\bar{\varepsilon}=0$), the existence of ε results a bias. Generally speaking, $Ef(\mathrm{X}) \neq f(EX)$, so far as $f(\mathrm{X})$ is not a linear function even when $X=Xo+\varepsilon$ and $E(\varepsilon)=0$. For example, if $f(X)=X^2$, we have $E(X^2)=Xo^2+E(\varepsilon^2)=Xo^2+\sigma_\varepsilon^2=\neq X_0^2$, where σ_ε^2 is variance of random error or fluctuation variable ε. Even in one dimensional case, the expectation of non-linear function of X results generally bias. In the multidimensional case, we assume that,

$X_1, X_2, \cdots\cdots\cdots\cdots X_R$

$Xi = Xoi + \varepsilon_i$

where $\{Xoi\}$ has a distribution function, and $\{\varepsilon_i\}$ has a conditional distribution function with the property depending on the realization of $\{Xoi\}$. A compound distribution function must be taken into consideration. It is well-known that an interesting problem in regression analysis is found, even in a simple case where $R=2$ and X_1 and X_2 have a linear relation. In the more complicated and multidimensional case, misleading results may be brought about if, at least, the random errors are disregarded. When a bias is added to the data, the sesults of data analysis must be carefully cheked.

An example is shown in the case where X is qualitative and expressed in categorical response. This example of dichotomous response is extremely simple as in Table 4. Now are assume the n_+, n_- and P_{ij} $(i, j: +, -)$ as an example in Table 5.

Then we have the data in the mean as $m_+=4750$ and $m_-=5250$. Frequency of + is smaller than that of − in the data whereas flequency of + is larger than that of − in the true. If we evaluate P_{ij} for all i, j, we can easily obtain the correct result by

Table 4 Qualitative Case (Categorical Data)
An example of simple case

data / true	+	−	Total
n_+ +	P_{++}	P_{+-}	1
n_- −	P_{-+}	P_{--}	1
n	m_+	m_-	n

Where $P_{--}+P_{+-}=1$
$P_{-+}+P_{--}=1$

P_{ij} : response probability; if i ≠ j, gives response error

n_+, n_- : true frequency for true +, −,

m_+, m_- : frequency of data for appearent +, −.

Table 5

data / true	+	−	Total
5500+	0.7	0.3	1
4500−	0.2	0.8	1
Total	4750	5250	10,000

P_{ij} for all i, j and the data (m_+, m_-). In the more complicated and multidimensional case, we can find the more unexpected results which are remarkably misleading.

I want to stress that data analysis is often useless and misleading if errors found in data are disregarded or if such evaluations of errors or fluctuations are not enough. The numerical evaluation of the properties of errors or fluctuations of the data is a crucial problem in data analysis. The data can be analyzed in a valid sense, only based on these informations. The method of analysis must have a good harmony with the quality of data.

II : Multidimensional statistical data analysis.

It is needless to say that the methods of multidimensional statistical data analysis are useful tools to reveal the data structure hidden in the raw data, and further, that

Table 6

Multivariate analysis
Quantification of qualitative data
Analyse des données (correspondences analysis)
Multidimensional scalogram analysis
Multidimensional scaling
Cluster analysis and classification

we make unmesurable mistakes when the multidimensional data are analyzed by simple methods of rigorous mathematical statistics. Recently, many interesting methods have been developed as above (see Table 6).

III : Treatment of qualitative data

In the process of data analysis, we frequently meet the qualitative data. In the case, we find two different approaches as the following. One is categorical data analysis in mathematical statistics, i. e. inferential data analysis, and another is descriptive data analysis of qualitative data. The former includes, contingency table, logistic model, logit model, loglinear model and so on. The latter case is mainly based on the idea of scaling. In this case, the methods of coverting qualitative data into

Table 7 Guiding conception of quantification as a method of data analysis

1.	To investigate how to express qualitative events as categorical data [categorized] *2)3)
2.	To quantify qualitative data for our specific purpose
3.	Numerical value is not immanent in a thing itself or an event itself but is given by the researcher only for the purpose of scientific research (Numerical value is a tool for the researcher's scientific purpose and variable depending on the purpose.)
4. **4)	To quantify based on the distinction between two approaches outside variable exists.……for example, the idea of regression analysis, discriminant analysis, and etc. no outside variable exists.……for example, the idea of principal component analysis linear factor analysis, and etc.

2) What idea is used in coverting qualitative events into categorical data? This is a crucial problem to be discussed.

3) • Expression in categorical data

$\delta_i(j,k_j) = 1$, i responses on the k_j-th category of the j-th item,

$= 0$, otherwise,

$i = 1, 2, \cdots\cdots, I,\ j = 1, 2, \cdots\cdots, J,\ k_j = 1, 2, \cdots\cdots, K_j.$

These form a data set of a survey or an experiment.

• Probabilistic expression in categorical data

In the case of probablistic response, the following expression is adopted.

$$\delta(j,k_j) = {}^hp(j,k_j)$$
$$i \epsilon h$$
$$h = 1, 2, \cdots\cdots, H \ (\ll I),$$
$$\sum_{k_j}^{K_j} {}^hp(j,k_j) = 1, \text{ for all } j, h.$$

4) Outside variable is "what we intend to know and predict or estimate by factors". The two approaches should be regarded as an operation in the series of processes which makes all the analyses valid, and whether or not the problem should be treated as having an outside valiable, is to be determined by the effectiveness of analyses.

categorical data are studied too. This comprises quantification of qualitative data, analyse des données (analyse des correspondances), multidimensional scalogram analysis, and multidimensional scaling. For example, I should like to mention the outline of the idea of quantification of qualitative data along the guiding conception of data analysis, as shown in Table 7.

Finally, the importance of graphical representation of data is to be refered to. This graphical representation includes graphs of raw data in the total under careful investigation of the conditions of data, or of raw data after clustering or dividing the whole into sub-groups by relevant informations, and of the results revealed only by appropriate multidimensional statistical data analysis. And, also it is to be remarked that computing environment is a matter of primary concern for data analysts.

References

1) Benzécri, J.-P. (1973) : L' Analyse des Données, Dunod.
2) Hayashi, C. (1973) : Methodological Problems in Mass Communication Research—From a statistico mathematical standpoint……, Studies of Broadcasting, N.H.K..
3) Hayashi, C. (1988) : New Developments in Multidimensional Data Analysis, Recent Developments in Clustering and Data Analysis, eds. Hayashi, C., Diday, E., Jambu, M. and Ohsumi, N., Academic Press.
4) Mallow, C.L. (1987) Some Principles of Data Analysis, Invited paper in the 46-th Session of ISI, Tokyo.
5) Rao, C.R. (1987) : Strategies of Data Analysis, Invited paper in the 46-th Session of ISI, Tokyo.
6) Tukey, J.W. (1962) : The Future of Data Analysis, AMS, vol.33.
7) Tukey, J.W. (1977) : Exploratory Data Analysis, Addison Wesley.

(昭和 63 年 11 月 9 日受理)