Presented at the XXXth International ASTIN Colloquium,
23 August 1999, Tokyo, Japan.

ON THE STRATEGY FOR EFFICIENT REALIZATION OF STATISTICAL REASONING

Akaike Hirotugu
Toride 1-7-14-204, Toride-Shi, 302-0004, Japan
Tel & Fax: +81(0297)73 6204

Abstract: Historical development of the logic of statistical reasoning is reviewed briefly and the informational view is discussed. The role of "true distribution" in the conventional statistics is explained and the necessity of considering alternative models is pointed out. The importance of the activity of constructing statistical models in the process of statistical reasoning is stressed and the use of the informational data set, composed of established knowledge, empirical findings, and observational data, is discussed. Utility of this constructive view is demonstrated by a realistic example of the golf swing analysis. The example shows that some strategic consideration is necessary for the construction of a practically useful model or image for the control of a complex autonomous system.

Key words: statistical model, information criterion, informational data set, golf swing.

1. INTRODUCTION

It is not an overstatement to say that uncertainties provide a significant drive to the intelligent activities in the human society. Personal and social activities are almost all concerned with the preparation for the future and scientific activities are aimed at reducing some kind of uncertainties.

Systematic approach to the handling of uncertainties is realized by the adoption of statistical reasoning. It is the purpose of the present paper to discuss the strategy for effective implementation of statistical reasoning by paying attention to the non-mathematical verbal aspect of the process of reasoning.

The discussion in this paper is developed as follows. Historical development of the concept of statistics is reviewed briefly to clarify the structure of the process of statistical reasoning based on the use of a stochastic structure. The necessity of considering alternative structures or models is pointed out and the use of the information criterion to measure the relative closeness of a statistical model to the "truth" is discussed. The importance of the process of constructing statistical models is pointed out and the use of informational data set that extends the traditional concept of observational data is discussed . The evolving nature of the informational data set is demonstrated by the discussion of the golf swing analysis as a realistic example. Some insight is obtained for the strategy of proper construction and use of informational data set. The whole discussion is directed to the clarification of the nature of statistical reasoning as an intellectual human activity.

2. HISTORICAL DEVELOPMENT OF THE LOGIC OF STATISTICAL REASONING

The standard view of statistical reasoning at present is to base the reasoning on the assumption of a probability distribution with some unknown parameter. In this section an explanation is given for the reason why such a use of stochastic structure has been developed and the limitation of this standard framework is pointed out.

Early developments

As is well-known statistics started with the description of the state of a country or nation for the purpose of economic and political planning. It is basically concerned with the use of observational data to help the decision to be made by the leader of the country.

Historically randomizers were used in making decisions under uncertainties. The classical example is the use of the divination by tortoiseshells in the ancient China. This was realized by reading the sign represented by the cracks of burned shells. A later refinement is the teaching of the I-Ching (the Book of Change) that reads the future through the random array of divining sticks.

The use of a randomizer was contemplated to cope with the lack of knowledge for the handling of the uncertainty about the future. It would be natural to expect that in such a case every effort was made to study the nature of the situation before consulting a randomizer. Thus the use of randomizers is inherently connected with the proper use of available knowledge.

Contact with mathematics

The first well-known serious use of mathematics in relation to social statistics is the analysis of the life-table of the city of Breslau which was done by Edmund Halley (Pearson,1978). This was to estimate the mortality and apply the result "to ascertain the price of annuities upon lives". The use was related to the prediction based on the analysis of the dynamics of the society.

Explicit application of probability distribution to the description of the distribution of social data was made by Adolphe Quetelet (Porter,1986). Probability distributions were also used for the handling of errors in the astronomical and geophysical observations.

The progress of the statistical method in this line may be viewed as the development of *the method of adjusting the structure of a randomizer to cope with a particular uncertainty, by using the knowledge provided by the observational data.* This leads to the view that the stochastic structure used in statistical reasoning is an artifact designed for the extraction of useful information from observational data.

Limitation of the conventional view of mathematical statistics

The systematic use of the parametric representation of a stochastic structure was developed by R. A. Fisher. In the theory of estimation, it is assumed that the functional form of the distribution is known but the numerical value of the parameter is unknown. Here, the most important aspect of statistical reasoning, the choice of the functional form, is ignored as the problem of the subject area.

3. INFORMATIONAL VIEW OF STATISTICS

A significant improvement of the framework for the discussion of statistical reasoning is realized by the introduction of the information criterion for the comparison of statistical models. Philosophical implication of this criterion is discussed in this section.

Fisher's view of likelihood

In the conventional theory of estimation, interest is focused on the expected accuracy of the estimated parameter *under the assumption of the truth of the model.* Thus the theory discusses the method of adjusting a model by the information supplied by observational data.

The classical method of maximum likelihood is realized by adjusting the parameter to the value that maximizes the likelihood which is defined by the value of the probability (density) of the event identical to the present observation. Fisher treated likelihood as "a measure of rational belief", without explication of the rationale (Fisher, 1973).

Fisher's theory of estimation is directed towards the estimation of the numerical values of "theoretical quantities in the specification of the causal system operating". Thus, if the assumed structure of the system is very far away from the truth, the theory loses its logical basis. This is often the case with the handling of a real system.

The representation of observational data as the output of a causal system gives the appearance of a scientific procedure to the method of estimation. However, in the case of an exploratory research, greatest uncertainty lies in the choice of the structure of the system. Further, the causal system could simply be an artifact for the extraction of necessary information. In this case the basis of the theory is in jeopardy. A solution of this difficulty is obtained by the adoption of the information criterion for the evaluation of statistical models.

Information criterion

The informational view explicitly recognizes the status of the probability distribution in the statistical reasoning as a model or an artifact for the extraction of useful information from observational data and accepts log likelihood as the basic measure for the comparison of statistical models.

The justification for the use of log likelihood in this case is given by the relation

$$\begin{aligned} I(p;q) - I(p;r) &= Ep\{\log p(x) - \log q(x)\} - Ep\{(\log p(x) - \log r(x)\} \\ &= Ep\{-\log q(x) + \log r(x)\}, \end{aligned}$$

where I(p;q) denotes the Kullback information, which is a measure of the deviation of the "true" distribution p(x) from the assumed distribution q(x), and Ep denotes the expectation with respect to p(x). This relation shows that the difference of minus log likelihood provides an unbiased estimate of the difference of the badness of the two models, q and r, as measured by the Kullback information.

The characterization of the difference of minus log likelihood as an unbiased estimate of the corresponding difference of the Kullback information remains valid for any choice of the "true distribution" p(x), i.e., *the above relation holds even if the "true distribution" p(x) is conceived differently by each observer.* This ensures the intersubjectivity of log likelihood as a measure of relative goodness of a model.

The meaning of this observation may be best illustrated by an analogy which the present author already discussed in several places. This is the analogy to *the measurement of the closeness of a mountain to the sky.* The concept of the sky may be

taken differently by each observer, yet the elevation above the sea level can serve as a common relative measure of closeness of a mountain to the sky.

A serious implication of this analogy is the clarification of the nature of the "truth" or the "true distribution" as an ideal of an observer. The activity of comparing models by log likelihood is a realization of the search for the best available knowledge at a particular situation.

Necessary modification of log likelihood for a model with parameters determined by the method of maximum likelihood was realized by the introduction of the information criterion AIC (Akaike, 1973,1974) which was defined by the relation

AIC = (−2) (log maximum likelihood)+2 (number of estimated parameters).

The second term on the right-hand side represents the necessary discount of the log likelihood to compensate for the effect of fitting the model to the data by adjusting the parameters by the method of maximum likelihood. The quick acceptance of AIC in various fields of scientific research provides a proof of the effectiveness of the informational view.

Use of Bayesian models

The introduction of AIC illuminated the necessity of considering the situation where the model can be refined indefinitely by increasing the number of adjustable parameters. A direct solution to this problem is obtained by introducing a Bayesian model where the overall variability of the parameters is controlled by assuming a probability distribution, the prior distribution of the parameters.

Since the resulting structure takes the form of a stochastic structure that generated the observational data, the log likelihood of the resulting structure can serve as a relative measure of goodness in the search for better models. Here, the likelihood of a Bayesian model is defined by the integral of the likelihood function of the original data distribution with respect to the measure defined by the prior distribution.

The conflict between orthodox and Bayesian statisticians has been notorious. Serious confusion with this conflict was that it was taken to represent the division between objectivists and subjectivists. Creative activities are always realized through serious subjective efforts. However, the results must be checked objectively to be useful for scientific applications. The informational view makes it possible to compare the results of subjective efforts represented by the models and eliminates the conflict and confusion.

4. CONSTRUCTIVE VIEW

In the conventional statistics emphasis has been placed on the development of proper procedures under the assumption of the knowledge of the structure of a "true distribution". Once the feasibility of comparing statistical models is established the main concern is shifted to the construction of good models. This naturally leads to the concept of constructive statistical reasoning which demands the discussion of the method or strategy for the construction of good models.

Use of informational data set

It is now evident that any information that can be useful for the conception of appropriate models must be included into the data set on which the reasoning is to be based. According to this idea the present author introduced the concept of IDS (Informational Data Set) which is defined by

IDS = (factual knowledge, hypothetical knowledge, observations),

where factual knowledge represents objectively confirmed knowledge that will be useful for the construction of models, and hypothetical knowledge represents subjective or personal knowledge not yet objectified, including empirical findings, tentative models, and other intermediate results of reasoning (Akaike,1997).

Importance of verbal analysis

The structure of IDS suggests the interdisciplinary character of constructive statistical reasoning. The activity cannot be limited within the realm of mathematical statistics. This fact shows the importance of the verbal analysis of the problem.

The stochastic structure in statistical reasoning focuses on a particular aspect of the object by properly ignoring unnecessary details of the information supplied by the observational data. Similarly, verbal expressions are used to focus on particular aspects of the situation. Thus the verbal analysis is essentially a statistical analysis with very natural method of representation which is easy to handle and convenient for communication.

The usual difficulty is the ambiguity of verbal representation. Nevertheless, the success of statistical reasoning depends critically on the performance of the verbal analysis of the problem and related IDS. This connection with the verbal analysis demonstrates highly intellectual nature of statistical reasoning.

5. ANALYSIS OF THE GOLF SWING: An Example of Constructive Statistical Reasoning

The use of verbal analysis may best be explained by the example of the golf swing analysis. This subject is concerned with an extremely complex system that does not allow simple mathematical handling.

Mutual dependence between IDS and model

The golf swing has following characteristics:

1) it is concerned with a system that has an extremely complex structure, including the system of muscles,

2) the parts of the body are mutually connected and the overall function cannot be confirmed easily by the conventional process of decomposition into components,

3) significant amount of anatomical knowledge is required for the understanding of the structure of the motion, and

4) the system that performs the function of the golf swing is generated by the intention of the golfer and is controlled by his image of the motion.

These characteristics suggest that the golf swing analysis is a challenging subject of statistical reasoning that requires explicit consideration of the structure of informational data set. The following explanatory example that represents a successive development of the model of the golfer and IDS demonstrates the evolutionary character of the statistical reasoning:

IDS(1) = (literature on swing technique, empirical findings, observations)
model(1) = (stick figure, for graphical or numerical analysis),
IDS(2) = (literature on swing technique and anatomy, empirical findings, observations)
model(2) = (skeleton covered with massive muscles,for the understanding of the

movement),

and

IDS(3) = (literature on swing technique and anatomy, increased empirical findings, observations)

model(3) = (image of the swing motion, for the control of the motion).

The dependence of the content of the informational data set on the intention of the analyst can be seen clearly by this example.

The first model, model(1), is a natural choice when conventional model fitting is contemplated. The intention of the analyst would be to find out an objective description of the swing motion which could be used for the analysis of the swing motion. A typical example is given by Jorgensen (1994). The model lacks anatomical description and cannot provide direct guidance for the swing motion of a golfer.

The second model, model(2), represents an effort to rectify the fault by including the anatomical knowledge in the informational data set. This expansion of the informational data set leads to the recognition of the fact that the structure and function of the muscle system is so complex that only the knowledge of dominant movements can and should be pursued. This shows the essentially statistical nature of the swing motion analysis.

The third model, model(3), takes the role of the brain into account. The understanding of the movement by the function of muscles is not sufficient for the control of an actual swing motion. The motion of a golfer is controlled by the "image" of the swing motion. The construction of the image is connected with the function of the brain. This shows the deep connection of statistical reasoning with the activity of prediction and control.

Necessity of the strategy for efficient reasoning

Above discussion has shown that the golf swing analysis is best suited for the study of the process of statistical reasoning. In particular, the inherent connection of the golf swing with various aspects of human intelligence makes it mandatory to develop some strategic consideration for the construction of IDS.

To get some idea of the strategy the result of the golf swing analysis by the present author is presented in the following. An interim report has been presented elsewhere (Akaike,1997), but that is now obsolete and the latest result will be presented. However, since the subject is extremely complex, only the gist of the finding will be reported.

An image of swing motion

The final purpose of statistical reasoning is the reduction of uncertainty of the knowledge about an object. This suggests a general strategy to focus attention on those aspects with maximum uncertainties or ambiguities.

In the case of the golf swing motion, text books of anatomy and biomechanics provide necessary background knowledge. The parts of the body with maximum freedom of the motion are the shoulder joints. The movements of hip joints also show similar freedom but the range is restricted. These joints are ball and socket (spheroidal) joints. The joint of this type can support a wide variety of motion.

A distinguished golfer suggests that ordinary golfers should focus on only one or at most two key swing thoughts (Nicklaus,1974). Thus it seems appropriate to consider the stabilization of shoulder joints as the primary objective. The basic image for the realization of the control is thus given by the verbal expression, "*Swing the shoulders parallel to the line of flight, particularly at the beginning of the backswing and at the*

hitting area."

Anatomical knowledge suggests following realization of this image. The back and downswing are respectively started by the action of the big deep muscles of the back of the leading side. The muscles are connected to a big muscle which goes down through the pelvis to the inner side of the knee and supports the swing action by locking the knee. The swing motion pulls the shoulders back and forth parallel to the line of flight and positions the shoulders and arms for the swing motion of the arms. The motion also realizes the movement of the wrists parallel to the line of flight, particularly in the hitting area. *A key to proper realization of the basic image is the uninterrupted natural lateral rotation of the leading upper arm.*

The swing motion of the arms around the shoulder joints is realized by the action of the superficial muscles of the back, in coordination with the action of other muscles. This motion is accompanied with the motion of the shoulder blades, which produces the turn of the shoulders around the neck. However, the trunk must be kept stationary to realize the basic image.

Complexity of an autonomous system

It might be considered that a detailed instruction such as "Start backswing by swinging the shoulders backward and then swing the arms freely back by the back muscles. Start downswing by pulling the trunk forward and swing the shoulders and arms forward to hit through the ball." would provide better image of the motion. However, there is a serious problem with this type of description.

An autonomous system like a human body is always equipped with some kind of built-in feedback or coordination between various parts of the system to realize the overall function. The conventional approach by the principle of "divide and rule" often ignores the availability and necessity of the use of the inherent structure of this type.

In the case of the motion of human body, various types of synergetic and antagonistic actions between the groups of muscles are required to realize a desired motion effectively. For example, for effective use of the muscles in the back, coordinated actions of the muscles around the waist and hips are required. There is a danger in giving a detailed description of a particular aspect of the motion, as this often distracts the attention of the golfer from necessary consideration of other parts.

This is the reason why an image which is a global description of the motion often functions better than a collection of detailed descriptions. *One further particularly important characteristic required of an image is that it must be intelligible to others,* as in the case of the connection of ideas discussed by Whorf (1956). In the case of the simple basic image of swinging shoulders parallel to the line of flight, it worked well to help the improvement of the swing motion not only of the present author but also of some fellow long hitters.

Incidentally, the above description of the swing motion shows that the locus of the club head in the downswing will resemble to a circular motion combined with the shift of the center parallel to the line of flight. This fits to the result of the model fitting by Jorgensen and suggests the possibility of going into the next stage of mathematical modeling of the swing motion by taking the anatomical structure into account.

6. CONCLUSION

The mathematical analysis based on the use of stochastic models occupied the central place in the history of the development of the method of statistical reasoning. The introduction of the information criterion emphasized the importance of the activity of proposing and comparing statistical models for each particular application. The

consideration of the structure of informational data set has shown that the construction of a statistical model is an interdisciplinary activity.

The analysis of the golf swing motion demonstrated the complexity and essentially statistical character of the problem. The example also suggested the necessity and importance of verbal analysis. The experience of the construction of an image for the control of the swing motion suggested some general idea of the strategy for efficient realization of statistical reasoning.

It is hoped that the whole discussion in this paper will be accepted as an effort for the illustration of the structure of statistical reasoning as a basic intellectual activity.

REFERENCES

Akaike, H. (1973) Information theory and an extension of the maximum likelihood principle, *2nd International Symposium on Information Theory,* B.N. Petrov and F. Csaki, eds., Akademiai Kiado, Budapest, 267-281. Included in *Breakthroughs in Statistics, Vol.I, Foundations and Basic Theory,* S. Kotz and N.L. Johnson, eds., Springer-Verlag, New York, 1992, 610-624.

Akaike, H. (1974) A new look at the statistical model identification, *IEEE Transactions on Automatic Control,* Vol. 19, 716-723. Included in *Selected Papers of Hirotugu Akaike,* E. Parzen, K. Tanabe and G. Kitagawa, eds., Springer-Verlag, New York, 215-222.

Akaike, H. (1997) On the role of statistical reasoning in the process of identification, *Preprints: SYSID'97 11th IFAC Symposium on System Identification,* Y. Sawaragi and S. Sagara, eds., International Federation of Automatic Control, 1-8.

Fisher, R. A. (1973) *Statistical Methods and Scientific Inference, 3rd edition,* Hafner Press, New York.

Jorgensen, T. P. (1994) *The Physics of Golf,* Springer-Verlag, New York.

Nicklaus, J. (1974) *Golf My Way,* Simon and Shuster, New York.

Pearson, K. (1978) *The History of Statistics in the 17th & 18th Centuries,* E.S. Pearson, ed., Charles Griffin, London.

Porter, T. M. (1986) *The Rise of Statistical Thinking 1820-1900,* Princeton University Press.

Whorf, B. L. (1956) *Language, Thought, and Reality,* J.B. Carroll, ed., MIT Press.