JYVÄSKYLÄ STUDIES IN COMPUTER SCIENCE, ECONOMICS AND STATISTICS 4

ERKKI PAHKINEN

THE METHOD OF SUPPORT AS STATISTICAL INFERENCE MODEL FOR INSTANT SAMPLE

JYVÄSKYLÄN YLIOPISTO, JYVÄSKYLÄ 1981

EDITOR: Lauri Saretsalo, Professor

Department of Statistics
University of Jyväskylä

URN:ISBN:978-951-39-8233-1
ISBN 978-951-39-8233-1 (PDF)
ISSN 0357-9921

ISBN 951-678-591-3
ISNN 0357-9921 C

Jakaja *Jyväskylän yliopiston kirjasto* *40100 Jyväskylä 10*
Distribution *Jyväskylä University Library* *SF-40100 Jyväskylä 10 Finland*

JYVÄSKYLÄ STUDIES IN COMPUTER SCIENCE, ECONOMICS AND STATISTICS 4

ERKKI PAHKINEN

THE METHOD OF SUPPORT AS STATISTICAL INFERENCE MODEL FOR INSTANT SAMPLE

ACADEMIC DISSERTATION TO BE PUBLICLY DISCUSSED, BY PERMISSION
OF THE FACULTY OF SOCIAL SCIENCES OF THE UNIVERSITY
OF JYVÄSKYLÄ,
IN AUDITORIUM PAULAHARJU
ON OCTOBER 17, 1981, AT 12 O'CLOCK NOON

JYVÄSKYLÄN YLIOPISTO, JYVÄSKYLÄ 1981

THE METHOD OF SUPPORT AS STATISTICAL INFERENCE MODEL FOR INSTANT SAMPLE

JYVÄSKYLÄ STUDIES IN COMPUTER SCIENCE, ECONOMICS AND STATISTICS 4

ERKKI PAHKINEN

THE METHOD OF SUPPORT AS STATISTICAL INFERENCE MODEL FOR INSTANT SAMPLE

JYVÄSKYLÄN YLIOPISTO, JYVÄSKYLÄ 1981

The Method of Support as Statistical Inference Model for Instant Sample/Erkki Pahkinen. - Jyväskylä: University of Jyväskylä, 1981.- 99 pages. - (Jyväskylä Studies in Computer Science, Economics and Statistics, ISSN 0357-9921, 4)
ISBN 951-678-591-3

Diss.

The aim of this study is to introduce a statistical inference model applicable to the analysis of an instant sample. In this connection instant sample means an observed probability sample or designed experiment which is not repeated or aimed to be repeated. We have demonstrated that the method of support put forward by Edwards, is an appropriate statistical inference model for such an inference situation. The central inference concept of the method of support is logarithm of the likelihood ratio named support $S(\Theta)$. Because the method of support, as such is not able to measure inference uncertainty, we have proven a new theorem in order to show how the support $S(\Theta)$ measures local uncertainty. This theorem shows that the method of support is an ordinary inference model. The proof is based on Rényi's incomplete probability distribution and Rényi's local uncertainty concept defined for it. In the case of instant sample we observe only one event whose probability is the joint probability $P_{\Theta}(y)$ of the sample, which is accordingly an incomplete probability distribution. In practice, applications of the method of support to the estimation and statistical test theory lead to the least local uncertainty (LLU) estimators and tests. As an empirical application we have analyzed an instant sample from Finnish pupils in 1970 with their learning achievement (Finnish IEA data). The normal linear regression analysis is used as a statistical model. Parameter estimation and diagnostics are performed using the method of support as inference model.

Inference Model; Method of Support; Local Uncertainty; Instant Sample.

Acknowledgements

A number of people have contributed to the progress of this study. I would like to mention them with greatest gratitude.

I am heavily indebted to Professor Hannu Niemi, the supervisor of this study, who encouraged me to look for the solution to the problem in information theory. This hint led to the verification of Theorem 1, one of the central findings here. I am also grateful to Professor Leif Nordberg for reading an earlier version of the manuscript and for his valuable suggestions. The numerous discussions I have had with Mr. Antti Penttinen have contributed to the shaping of the content of instant sample, another central concept used.

Professor Lauri Saretsalo has revised the statistical terminology of the English version translated by Miss Helena Valtanen and Mr Ilkka Utriainen from the orginal Finnish manuscript. The strenuous typing was carried out by Mrs.Ritva Riepponen and the typing of the earlier Finnish version by Mrs. Liisa Misukka-Niemi.

My wife Tuula has given me her spontaneous, mentally stimulating support which has helped me to overcome the crises arisen in the course of this study. My daughters Pyry and Metsälintu have taken a positive attitude in my devotion to science.

I have received financial assistance from the Ellen and Artturi Nyyssönen Foundation.

To all those mentioned above my sincere thanks are due.

Jyväskylä
September 1981 *E.J. PAHKINEN*

CONTENTS

I INTRODUCTION

The present study deals with statistical inference on a theoretical level, with a practical research situation as the frame of reference. The research situation considered here is a typical statistical study in social sciences which can be described in terms of statistical concepts on the basis of the data producing system. We are dealing with a probability sample drawn at random from an existing human population at a given moment. Thus, the target population is finite and the sample obtained from the population is large, calculated as sampling units. Because of the nature of the population, sampling procedure cannot be repeated as such and often it is not even intended to be repeated. In short, the data {y} is obtained by means of instant sampling. There is in statistical literature criticism the methods commonly used in this type of research, as can be seen in the report of Henkel and Morrison (1969) on the incorrect use of statistical inference. Nevertheless, scientific and other institutes on different fields of research produce statistical surveys based on the instant sampling procedure described above. In the present study, we are primarily interested in how the inference situation described above can be analyzed in such a way that statistical inference can be seen as an entity of its own and also in how the properties of the data influence the choice of the inference model and the expression of inference uncertainty.

In Chapters 2-1 and 2-2 preliminary definitions are presented, which concern the decomposition of a statistical inference situation into a model triplet [M,P,I] where M denotes the substantive problem examined as a mathematical model and P the randomness of observations due either to the phenomenon in question or to a data producing system like the probability sampling. The symbol I denotes the inference model the central concept of which is the measurement of inference uncertainty. By inference uncertainty we mean uncertainty occurring in inductive inference where generalizations are made on the basis of the particular data.

Inference uncertainty is one of the main objects of interest considered in the present study. It is examined in Chapter 2-3 where a synthesis of the inference models in use is presented, including the Neyman-Pearson test theory and Bayesian inference. Inference uncertainty is considered in terms of the definition of those events for which the inference uncertainty, usually some concept of probability, used in the model is determined. Our frame of reference is the concept of total evidence, put forward by Suppes (1966), which refers to the observed events at a statistician's disposal and to such events which he is able to obtain (unobserved events). The synthesis of the inference models reveals that most of these models presuppose the use of unobserved events in the measurement of inference uncertainty. The clearest case of unobserved events is found in the frequentist inference which presupposes the multiple repetition of the sample.

In the case of an instant sample only one point is observed from the sample space. The realization of this point is data $\{y\}$ and the probability assigned to it is the joint probability $P_\Theta(y)$ of the sample. The probability measure P_Θ is assumed to be discrete and it includes in some mathematical form the substantial problem to be studied and the stochastic element due to the sampling fluctuations. In a situation like this, it is appropriate to use the class of inference models operating with a limited total evidence where the central concept is usually likelihood and the application of a likelihood function. If the parameter Θ belongs to the parameter space $\boldsymbol{\Theta}$, then the likelihood for Θ is $L_y(\Theta) \propto P_\Theta(y)$. In its most reduced case, an inference model based on likelihood uses data alone as the total evidence. This class of inference models includes the method of support, put forward by Edwards (1972), whose central concept of inference, the support $S(\Theta)$, is defined as the natural logarithm of the likelihood ratio.

The problem that there does not exist any clear measure of inference uncertainty in the method of support is considered in Chapter 3-1. The support is compared with both Kullback's and Shannon's information but it is to be discovered that they do not provide any interpretation for it with regard to the measurement of inference uncertainty. Instead, we arrive at the central finding of the present study, according to which the support is the difference between two Rényi's local uncertainties (Theorem 1 in Chapter 3-1). It is proved on the basis of Rényi's incomplete probability distribution and the concept of local uncertainty determined for it by him. It is only this interpretation which makes the method of support a statistical inference model with a concept of inference uncertainty of its own. In a statistical reasearch situation, for example, the incomplete probability distribution is represented by the joint probability of the observed sample $P_\Theta(y)$ for which the local uncertainty is thus determined. In the case of statistical instant samples, the statistician has only one event at his disposal and in Chapter 3-6 the method of support is shown to be an inference model for this kind of samples.

The properties of the method of support, like norming, are considered in the present study from an information theoretical point of view. An interesting interpretation for support is provided by a statistical model by means of which the least local uncertainty can be achieved. With regard to this, we present in Chapters 3-3 and 3-5 a new interpretation for estimation and the testing of hypotheses. According to this interpretation, the method of support produces least local uncertainty estimates (LLU estimates) and least local uncertainty tests (LLU tests). They correspond to Edwards's concepts of evaluation and support tests, respectively.

In Chapter 4, we go back to the original starting point of our study and apply the method of support to a statistical inference situation where the data is an instant sample. Obviously, the ability of an inference model to function is best found out in a real research situation. The statistical method used is the linear regression analysis, a well-known tool of scientific communities, by means of which a substantive problem in the field of educational sciences is analyzed.

Preparatory considerations in Chapters 4-1 and 4-2 deal with statistical inference in connection with the regression analysis to which the method of support is applied as the inference model. These considerations are more comprehensive than the empirical applications on this study require. They concern the least local unccrtainty estimates of the regression coefficients and statistical tests, the latter of which can be used in constructing a hierarchic least local uncertainty test of the regression analysis. The hierarchic test is analogous to Kullback's (1959) information theoretical test.

The empirical application of the method of support is presented in Chapter 4-3. As a substantive problem it belongs to the field of educational sciences. Our aim is to provide an explanation for pupils' school achievement with home type and school type variables and to evaluate what the relative explanation of each group of these variables is. Several international studies of this problem have been carried out, one of which, by Noonan and Wold (1977), has here been used as the reference study. The data used consists of 1310 primary school pupils drawn in 1970 by instant sample from Finnish primary schools.

2 MODELS IN A STATISTICAL INFERENCE SITUATION

A statistical inference situation is a stage of a scientific or practical research in which inferences are made with the aid of information derived from the data by estimating the unknowns of a model, theory or preconception under study or their fittings with the observations. Our aim is thus the reduction of informational uncertainty in the phenomenon studied by performing statistical experiments and sampling procedures.

The inductive nature of statistical inference is due to the property that we make a generalization either to the direction of a hypothetical population (the space of all experiments), or to the direction of an existing population (a finite existing population) on the basis of an observed experimental design or sample. The proper treatment of a statistical inference situation presupposes the representation of its stages as models. The definition of the models can be derived from the phases of empirical research: specification of the substantive problem, planning and implementation of data producing and, as the last phase, statistical inference.

The starting point is a substantive problem, derived from the phenomenon studied, which we try to specify as a mathematical model M{ }. The mathematical model M{ } implies the structure of a causal relationship, dependence,or some other property deduced from the theory of the field. It is often linked with the definition of the connections between the parameters of the model. Our aim is a model the interpretation of which corresponds to a substantive problem and which is as careful a description of real world as possible. Such a model is also called a substantive model.

A substantive problem is made statistical by adding a random component to the mathematical model. The model of the random component is a probability model P{ }. Its introduction is based on the randomness of separate measurements, of data producing, or on the substantive randomness of the phenomenon under study. Each of these alone brings about that the data is interpreted as one event or realization among all the events and realizations of a random phenomenon. Statistical inference cannot even be considered in any other system as that described above, as has been emphasized by Fraser (1979).

In a statistical inference situation, the mathematical model M{ } and the probability model P{ } are formally treated as their combination or as the statistical model [M,P]. A theoretically important statistical model is the joint probability of an observed simple random sample, determined for a sample space,

$$P_\Theta(y) = P_\Theta(Y_1 = y_1, Y_2 = y_2, \ldots, Y_n = y_n)$$

where P_Θ is a probability measure dependent on the mathematical model M{ } and the probability model P{ }, and y is an observation of an n-dimensional random variable $(Y_1, \ldots, Y_n)$. Parameter Θ represents the unknown component which is specified in the statistical inference model. Thus, the joint probability $P_\Theta(y)$ is the probability of a single event, which plays a central role in statistical inference.

The statistical inference model I{ } is either a collection of inference rules or an inference function by means of which the researcher draws conclusions about the unknown parts of the statistical model. The properties of the inference model include an expression of uncertainty of inference results or an inference situation. In this respect, the statistical inference models in use offer different alternatives depending on what is meant by a total evidence at the statistician's disposal in an inference situation. Probability is often used as a measure of uncertainty in a frequentist sense, in which case the total evidence is the sample space of an observed sample and its multiple repetition. In the present study, we concentrate on an inference model I{ } in which inference uncertainty is measured with information theoretical concepts.

From the definition of the model triplet above, we can deduce Requirement 1 concerning a statistical inference situation.

Requirement 1. A statistical inference situation should be organized into a model triplet [M,P,I].

The organization of a statistical inference situation into a model triplet clarifies the empirical research strategy and the measurement of uncertainty in inference results.

2-1 Substantive Model

In a statistical inference situation, we are studying a real world state or system the realization of which in a research situation is the data. For example,according to Fraser (1979), we should deduce from the background information or from the theory based on earlier studies a mathematical model in which our research problem is formulated in a statistically solvable form. In the present study, this model is termed a substantive model M{ }. It contains the listing of the variables, their relationships,which are of interest to us, and parameters.

Technically, the substantive model is usually given as an explicit formula. For example, linear regression analysis is based on the folloving model for the relationship between the variables X and Y:

$M \leftrightarrow \mathbf{y} = \mathbf{X}\Theta$, where

$\mathbf{y}$ is a vector of n observations on the dependent variable

$\mathbf{X}$ is an n x p matrix of n observations on the independent variables

Θ is an p-vector of parameters.

The mathematical model is often a very scarce one, due to the nature of the phenomenon. This alternative regularly leads from the substantive model directly to the testing of statistical hypotheses. For example, in behavioural sciences the object of interest may be the comparison of two populations with regard to some common attribute, in which case the mathematical model is reduced as follows:

$$M \leftrightarrow \Theta_A = \Theta_B, \text{ where } \begin{array}{l} \Theta_A \text{ attribute in population A,} \\ \Theta_B \text{ attribute in population B.} \end{array}$$

The importance of the substantive model is pointed out many times in the discussions of the fallacious use of statistical inference see Pearson (1962) and Sterling (1959). Erroneous interpretations of inference results arise from the fact that researchers have not been able to return to the substantive model at the end of an inference situation. A statistical inference situation ought to start from and end in the consideration of the substantive model.

The consideration and deduction of the substantive model is the task of the researchers in the substantive field. Statistical knowledge is required in checking, for example,whether it is possible to measure the variables of the substantive model with the aid of data planned to be collected. The substantive model is not examined separately in the present study.

2-2 Probability Model

The second member in the model triplet [M,P,I] of a statistical inference situation is the probability model P{ }. It is used, in the first place, for defining the joint probability $P_\Theta(y)$. On the other hand, it is also used in statistical inference models as a measure of uncertainty, as in the Neyman-Pearson theory for the size and power of the test. Seidenfeld (1979) describes this dualism of the probability model in the following manner: probability_1 describes the random nature of a phenomenon and probability_2 measures the inference uncertainty. As we in the present study also use other concepts than probability for the measurement of uncertainty in an inference situation, we will later define a more comprehensive concept which will also include the probability model in the sense of probability_2.

The probability model here means the same as the probability P(A) which is defined in an algebra of events as follows: If P(A) is the mathematical probability of an event A and events A and B are subsets of a sure event E, then

$0 \leq P(A) \leq 1$, for every event A

$P(E) = 1$, for sure event E and

$P(A \cup B) = P(A) + P(B)$, for every pair of mutually exclusive events A and B.

The probability model is linked with the substantive model as a distribution assumption. A normal regression analysis would be as follows if the distribution of its random component is assumed to be known:

$$M \leftrightarrow y = X\Theta + \varepsilon$$

$$P \leftrightarrow \varepsilon \sim NID(0, I\sigma^2)$$

If we consider the X matix as given, the density $f_\Theta(y)$ of the joint probability $P_\Theta(y)$ of a sample $y = \{Y_1,\ldots,Y_n\}$ is

$$f_\Theta(y) = (2\pi\sigma^2)^{-n/2} \exp\{-\frac{1}{2\sigma^2}(y - X\Theta)'(y - X\Theta)\}.$$

(The consinuity of the probability distributions is unessential, when dealt with likelihood rations (cf. Chapter 3)).

What makes the use of the probability model as inference uncertainty appealing is the fact that the probability measure P_Θ can be defined by using various philosophical principles as the starting point. Good (1950) lists six concepts of probability: classical, empirical, frequentist, axiomatic, subjective and fiducial. One problem in the use of these probabilities in the measurement of inference uncertainty is how to find those events for which the corresponding probability is determined. This problem is not associated with the statistical model $P_\Theta(y)$.

2-3 Inference Model

The third member of the model triplet [M,P,I] of a statistical inference situation is the statistical inference model I{ }. It is either a collection of inference rules or an inference function. By using the inference model we produce the estimates or statistical tests concerning the unknown constants of the statistical model and express the degree of uncertainty included in our inferences. The central problem is thus how to measure uncertainty in general. We will approach this problem from two points of view: from the theory of statistical inference and from the information theory. Another essential question is for which events we define the inference uncertainty. Our starting point is the data together with the various sets of events added to it by the statistical model and the inference situation. The collection of these sets is here regarded as the concept of the statistician's total evidence.

The status of probability as a measure of uncertainty is emphasized by the fact that uncertainty is often expressed as various risks which are probabilities. This has also been stressed in inductive logic where, according to Hacking (1965) and Seidenfeld (1979), it has been concluded that the confirmation function c(|) measuring the degree of uncertainty or acceptance of a hypothesis must follow some axioms of probability with the difference, however, that there is no interpretation for the unconditional event c( ). This view of the measurement of uncertainty is found in the most common statistical inference models, as can be deduced from their synthesis.

The use of the probability concept as a measure for inference uncertainty is restricted by that algebra of events for which the probability in question has been defined. Here we borrow from inductive logic the concept of total evidence, put forward by Suppes (1966). By total evidence we mean all the information at the statistician's disposal at the moment of inference. It consists of: 1)one observed event as the data {y}, 2) unobserved events which the statistician is able (in principle) to obtain like all the events of the sample space or the whole of sample space {Y}, and 3) those parts of the statistical model which are connected with the measurement of uncertainty as, for example, the parameter space.

The collection of these events forms the class of all events for which the measure of uncertainty is defined. For example, if uncertainty is defined as a mean over the whole sample space, the statistician introduces events which he is not normally able to observe in practice. The observation of all the events of the sample space is often quite impossible. Total evidence is denoted as an unordered sequence { }.

Let us consider how inference uncertainty is treated in some inference models and how large their required total evidence is. The inference models are the following:

1. Significance test (Fisher - Karl Pearson - Student)
2. Neyman-Pearson test theory (Neyman - Egon Pearson)
3. Fiducial inference (Fisher)
4. Bayesian inference (Savage et al.)
5. Likelihood inference (Fisher et al.)
   5.1. Likelihood test (Hacking)
   5.2. Relative likelihood (Kalbfleisch - Sprott)
   5.3. Method of support (Edwards)
6. Kullback's information (Kullback)

The inference function of the significance test is the statistical model $P_{\Theta}(y)$ in which the data is a realization in the sample space Y. The sample space is divided into two subsets C and $\bar{C}$ so that the size of the region $\bar{C}$ (the critical region) is

$\alpha \approx 0$. The size of the acceptance region is $1 - \alpha$ and the inference rule consists of determing whether the observation y belongs into the rejection or the acceptance region. Inference uncertainty is measured as a risk level α for which a frequentist interpretation is given in terms of repeated experiments or samples. Criticism against the significance test has been directed, according to Hagood (1969), at the interpretation of the inference result. We are dealing with a simple logical disjunction: if the realization is located in the critical region $\bar{C}$, then P_Θ is either true or untrue but, in the former case we have a rare event. Inference uncertainty is determined from two different spaces of events connected with the data. They are the sample space and the infinite repetition of the sample space, the latter being $Y^{(\infty)} = \{Y^{(1)}, Y^{(2)}, Y^{(3)},\ldots\}$. The total evidence of the significance test is $\{y,Y,Y^{(\infty)}\}$.

In the Neyman-Pearson test theory also an unambigous division into acceptance and critical region of the sample space is needed. This is made by applying the likelihood ratio $\lambda = L_y(\Theta_o)/L_y(\Theta_1)$, which is used as the inference function in this inference model, and which in fact thus presupposes the addition of a parameter space into the total evidence in order that the alternative models could be specified. Inference uncertainty consists of two risk levels, α and β, which determine the division of the sample space Y, and which are usually interpreted in a frequentist manner. Egon Pearson (1962) points out that the frequentist use was not originally intended for the Neyman-Pearson test theory. The total evidence of Neyman-Pearson inference is thus $\{y,Y,Y^{(\infty)},\theta\}$.

In fiducial inference, the inference function is a fiducial probability $P_f(\Theta)$ which is determined by the pivotal function $g(y;\Theta)$ from the statistical model $P_\Theta(y)$. The fiducial probability represents both inference uncertainty and a form of output. The user of the results must continue his own inference from the observed fiducial distribution. Wilkinson (1977) recommends the fiducial distribution as a general solution in the measurement of inference uncertainty. The use of a parameter as a random variable has been criticized. Usually it is known for sure in advance that there exists only a single correct value for the parameter Θ. In fiducial inference, the required total evidence is $\{y,\theta\}$,where θ represents the whole parameter space for which the probability measure P_f has been deduced.

In Bayesian inference, the inference function is a posteriori distribution $P_B(\Theta)$ which is determined by the Bayesian postulate from the a priori distribution $P_A(\Theta)$. When applying this inference model the user of inference results must make two independent decisions. First, he has to make a subjective choice of the a priori distribution and second, to decide which concept of probability to apply to the distribution in question. The alternatives $\{d_1,d_2,...\}$ connected with the choice form an addition to the statistician's required total evidence which is denoted by the set of choices D. The required total evidence of Bayesian inference is $\{y,Y,\theta,D\}$.

Bayesian inference is a very interesting inference model used by scientific communities. New scientific knowledge produced by means of observations is a link in the chain which starts from the level of earlier accumulated knowledge and culminates in a new level corrected by observations. In Bayesian inference these phases are represented by a priori distribution, statistical model and a posteriori distribution as inference result. Lindley (1956) has presented an interesting application of this. He treats the a priori distribution as an expected a priori inference uncertainty which is corrected by expected information carried by the data. Technically speaking, one has here passed on from the concept of probability to the concept of information.

The difficulty with Bayesian inference is the choice of distribution that together with its concept of probability determines the measurement of uncertainty. Here Savage (1954) clearly recognizes the existence of metacriteria in a statistical inference situation, because the statistician has to make subjective decisions when choosing the statistical model. These decisions imply the existence of metacriteria. A clear expression of the statistician's choices for example for the a priori distribution is organized subjectivism which is superior to a choice lost behind a more "objective"inference model. Bayesian inference has been criticized because of its use of subjective probabilities. For example, in Hoggarth's(1975) extensive literature synthesis, the core of criticism is against the poor communication of subjective probabilities from one person to another and the weak consistency from one inference situation to another. If we want to be free from the statistician's own subjective, previously fixed probabilities, one possibility is to produce the a priori distribution as a nomogram, as Dickey (1973) and together with Freeman (1975) recommends.In nomograms, some parameters are left open and determined case by case in the way which the user of inference results considers the best.

In likelihood inference, the inference function is a likelihood function.

Definition 1. The likelihood of a statistical model is $L_y(\Theta) = P_\Theta(y)$.

Likelihood is often considered as a likelihood function $L_y(\Theta)$ which is represented in the form of c $P_\Theta(y)$, where c is an arbitrary multiplicative constant and the value of y is fixed. Inferences made by means of the likelihood function are based on various likelihood principles. These principles regulate the expression of inference uncertainty in likelihood inference. In our study, we consider three different principles which lead to different inference models. The total required evidence of likelihood inference is $\{y,\theta\}$.

Of the various applications of the likelihood function, we will consider here three inference models: likelihood test (Hacking), relative likelihood (Kalbfleisch-Sprott) and the method of support (Edwards). The common property of these three models is the fact that uncertainty is measured by a rank measure which thus cannot realize the axioms of the function c( ) measuring the acceptance order in inductive logic, at least not in a probabilistic sense. It is a question of the order of statistical models with regard to acceptance or uncertainty. Maybe because of this,likelihood inference is generally regarded as a weak inference model.

According to Hacking (1965), the acceptance order of statistical models can be expressed by comparing observed likelihoods with each other. The inference function is an operator that ranks observed likelihoods in the order of magnitude and arranges the corresponding hypothesis according to the order of acceptance or uncertainty. Nothing is normed in this inference model.

In relative likelihood, Kalbfleisch and Sprott (1970) norm the likelihood function in relation to its own maximum and thus get as the inference function the ratio

$$R(\theta) = L_y(\theta)/L_y(\hat{\theta}),\ R(\theta)\ \varepsilon\ [0,1].$$

In this case, too, the acceptance order of statistical models is produced by means of relative values R(θ) which are strikingly similar to propabilities. We are dealing with an acceptance order which at the same time indicates the order of inference uncertainty.

The comparison of different models produces families in the acceptance order. Difficulties in interpretation may come up when the acceptance orders between families and inside families are compared.

In the method of support, the inference function is the natural logarithm of likelihood ratio $S(\Theta) = \log_e R(\Theta)$. Edwards (1972) holds the view that the support function, when used for in the comparison of scientific hypotheses, is an adequate inference model which produces the acceptance order. The method of support is considered in more detail in the following chapters of this study. In the empirical application presented in Chapter 4-3 we use the method of support as an inference model and its computational results are compared with those obtained by fiducial and rclative likelihood inference.

It is also possible to measure uncertainty in a statistical inference situation with some other concepts than that of probability, in connection with likelihood inference. The measurement of uncertainty is one of the central concepts of information theory. Kullback (1959) has considered the relation of statistical inference and information theory.

Kullback information is derived from the likelihood axiom, according to which all information obtained from the data which concern the comparison of two statistical models is contained in the ratio $P_{\Theta_i}(y) / P_{\Theta_j}(y)$. Its logarithm

$$\log P_{\Theta_i}(y) - \log P_{\Theta_j}(y)$$

is the discrimination between the models at the point $Y = y$. Kullback information is its mean value

$$I(i:j) = \sum_{y \varepsilon Y} P_{\Theta_i}(y) \log \frac{P_{\Theta_i}(y)}{P_{\Theta_j}(y)} .$$

Although Kullback information measures expected uncertainty of choosing between two statistical models, it is not, as such, used as inference uncertainty in a statistical inference situation. In an inference situation, Kullback information is put into the Neyman-Pearson inference function as a statistic and after that uncertainty is expressed, according to the inference theory in question, as risk levels α and β, which are probabilities and commonly interpreted. Thus Kullback information measures substantive uncertainty between statistical models, and its required total evidence is $\{y, Y, Y^{(\infty)}, \boldsymbol{\theta}\}$.

The synthesis of inference models reveals three common properties which should be checked and displayed in a statistical inference situation. They are a) inference function, b) total evidence {y,...} and c) inference uncertainty of the statistical model. The statistician has to consider these things prior to inference and thus to set himself the metacriteria presupposed by a genuine inference situation. The consideration of these questions in an inference situation gets clearer, if the statistical inference model I{ } is decomposed into the model triplet as a formula which expresses both the inference function and the property measuring inference uncertainty. In order to illustrate this, we have chosen the linear regression analysis based on normal distribution, where the statistical model remains the same [M,P] but the inference models is a) Neyman-Pearson inference, b) Bayesian inference and c) relative likelihood R(Θ). Three different inference models in connection with the same substantive and probability model.

$M \leftrightarrow y = X\Theta + \varepsilon$

$P \leftrightarrow \varepsilon \sim \text{NID}\ (0, I\sigma^2)$

(a) $I \leftrightarrow$ N - P theory α_o, β_o, where α_o and β_o are risk levels

(b) $I \leftrightarrow$ Bayes $\{P_A(\Theta), P_B(\Theta)\}$, where $P_A(\Theta)$ is a priori and $P_B(\Theta)$ a posteriori distribution

(c) $I \leftrightarrow R(\Theta)$ and $100\{R(\Theta)\}\%$, where $100\{R(\Theta)\}\%$ is percent order of acceptance

An important factor with regard to the realization of a statistical inference situation is Requirement 2, derived from the above, which concerns the definition of the inference model.

Requirement 2. In a statistical inference situation, the inference function and the concept of inference uncertainty must be displayed.

The display of the inference models does not include a visible statement of the total evidence. It is examined as a criterion of choice of the inference models included in the synthesis. The total evidence observed in an inference situation and the adapted inference model should correspond to each other because the measurement of uncertainty is linked with it. The structure of the total evidence is presented as a figure below, in which the data y{ } is in the middle and the other components surround it as suppelements.

$$\begin{tikzcd}
 & Y & \\
\theta & y \arrow[u] \arrow[l] \arrow[r] \arrow[d] & Y^{(\infty)} \\
 & D &
\end{tikzcd}$$

Explanation of the components of the total evidence

y Data, which corresponds to one realized point of the sample space Y. The probability of this point is $P_\Theta(y)$.

Y Sample space, which depends of the sampling procedure or design of experiment.

$Y^{(\infty)}$ Infinite repetition of the sample space, necessary for the realization of the frequency principle in defining the risk levels.

θ Parameter space which is needed for the construction of the inference measures for a parameter.

D Space of decision alternatives, necessary in choosing the a priori distribution in the Bayesian inference.

The measurement of the total evidence and inference uncertainty can be combined as Requirement 3 concerning the choice of the statistical inference model I{ } .

Requirement 3. The statistical inference model must be chosen in such a way that the total evidence needed for the measurement of inference uncertainty matches with the total evidence observed or the total evidence at the statistician's disposal.

An interesting dualism is connected with the measurement of inference uncertainty. If we use a concept of probability, we always have to add to the data a class of events in order to make the total evidence as extensive as possible. In this sense, the richest total evidence is that of the Bayesian inference {y,Y,θ,D} from which an a posteriori distribution is deduced. The most reduced total evidence in the inference models considered above is connected with the likelihood inference and fiducial inference in which the total evidence is {y,θ}. The most concise total evidence of all is the one which includes the data {y} alone. That is why it is important to know what kind of concept of inference uncertainty can be deduced for it. We are dealing with the uncertainty of a single event $P_\Theta(y)$, which is later shown, can be interpreted as Rényi's (1970) concept of uncertainty of one event with probability $P_\Theta(y)$. This will be called Rényi's local uncertainty. It is significant in an inference situation in which the total evidence cannot be expanded by adding sets of events to the data. What follows is that we shall show the inference function in Edward's (1972) method of support can be interpreted as the difference between two Rényi's local uncertainties.

The synthesis of the inference models led to two additional requirements which are connected with the measurement of inference uncertainty. Measurements can be made either with concepts of probability or with information-theoretical concepts, the latter of which makes it possible to use the relatively reduced total evidence $\{y,\theta\}$. This will be shown in Chapter 3.

2-4 Choice of Models in an Inference Situation

In a statistical inference situation the choice of models concerns the models in the triplet $[M,P,I]$. The choice should be examined case by case, as is emphasized by Bartlett (1971). The statistician makes choices in which he, from among statistical and inference models, looks for those he thinks are best suited to the situation. According to Menges (1973), we are here dealing with a genuine decision situation in which the statistician has to put the different alternatives in order of preference on the basis of some metacriterion. By the metacriteria of a statistical inference situation we refer to those hidden preferences which the statistician must make use of before the inferences can be made.

In the previous chapters the examination of a statistical inference model and an inference situation led us to the three requirements connected with the metacriteria of the inference situation. They concern the decomposition of an inference situation into the model triplet [M,P,I], inference uncertainty and the notion of required total evidence.

The choice of the statistical model [M,P] can be made from two directions. It is either a substantive model of the phenomenon being examined as such, in which randomness derives from the phenomena themselves, or an instrumental statistical model is in question. In the latter case we are referring to such a statistical model $P_{\Theta}(y)$ which is a probability model. In that case randomness derives from the data producing system or the measurement in general. A good example of these cases is the property of large sample statistics to approach their limit distribution, which in many cases is a normal distribution. In choosing a statistical model this finding is worth exploiting.

The choice of the statistical inference model I{ } is linked with: a) the display of inference results, b) how inference uncertainty is measured and c) what is the observed total evidence in an inference situation. The examination of inference models indicated that the results can be left open,in which case the user of the results must make the decisions in his own inference situation. Both the Bayesian and the fiducial inference, for example, display an a posteriori distribution. On the other hand, in N-P test theory a genuine decision is made between the choice of alternatives.

Actually,only the definition of total evidence introduces a criterion which directs the choice of an inference model. This is so because of the fact that the required total evidence and inference are linked with each other. This can be seen in the tabulation below, which is a summary of the previous discussion. Anyone choosing an inference model should keep in mind that a certain inference model requires its own supplements in total evidence. The choice of an inference model thus presupposes the justification of the legitimacy of these additions.

Table 1.

The supplements of total evidence to be added to the data {y} and the connection of certain inference models with them in measuring inference uncertainty.

Inference model (Always observation {y})	Sample space Y	Repetition of sample space $Y^{(\infty)}$	Parameter space Θ	Decision space D
1. Bayesian inference	yes	no	yes	yes
2. Neyman-Pearson test theory	yes	yes	yes	no
3. Kullback's information	yes	yes	yes	no
4. Significance test	yes	yes	no	no
5. Method of support	no	no	yes	no
6. Fiducial inference	no	no	yes	no

The examination of statistical total evidence has shown that unobserved sets of events can be reduced from the very richest total evidence in the way that the inference uncertainty can be measured from those left over. The most reduced total evidence is naturally the single event {y}. In Chapter 3 it will be demonstrated that local uncertainty can be expressed with it supplemented by the parameter space Θ.

3 THE METHOD OF SUPPORT AS INFERENCE MODEL FOR INSTANT SAMPLE

In scientific investigations a widely accepted principle is that of repetition of observed experiments and samples. Then, conclusions drawn by scientists are based on this principle. In actual research activity, however, this requirement is rarely realized as is seen from many warnings exposed elsewhere-see Henkel and Morrison (1969), Sterling (1959) and Pearson (1962). A typical example is an empirical investigation, where the data is generated by an experiment or sample, which is only once performed. We define this kind of data as an instant sample.

Definition 2. An instant sample is a probability sample or an observed experiment which is not repeated or aimed to be repeated.

There are no hindrances to the use of an instant sample in statistical inference, because it represents the concerned existing or hypotetical population in a proper way. In the direction of measuring inference uncertainty there is a noticeable restriction since the instant sample provides total evidence whose extent is only the observed sample {y}. We have no further information carried by repeated data so that inference uncertainty cannot be measured either in a frequentist sense as, for example, linked with the N-P test theory. In this we have to keep to the local uncertainty supplied by an instant sample.

It is evident that we need an inference model which can be used with instant samples. If the earlier defined Requirements 1-3 are kept in mind the total evidence provided by the instant sample should match with the total evidence required by the model to be chosen . Our taxonomy of inference models on page 22 contains no inference model whose total evidence is only the data {y}. At least one supplement is to be done. If this is a parameter space Θ, possible models are likelihood or fiducial inference. Here we have chosen the method of support, whose properties are to be examined next.

The basis of the method of support is the observed likelihood. For the various definitions we refer to Fraser's (1979), Edwards' (1972) and Birnbaum's (1969) presentations, where the law of likelihood and the likelihood principle are thrown into a single axiom. In this study we have taken the definition directly from Edwards (1972), in which the frame of analysis is the statistical model [M,P]. Here the unknown parameter Θ of the model is varied by two hypotheses H_i: $\Theta = \Theta_i$ and H_j: $\Theta = \Theta_j$. The specifications of the hypotheses are here interpreted as two statistical models $P_{\Theta_i}(y)$ and $P_{\Theta_j}(y)$.

The law of likelihood: The data { y } supports the statistical model $P_{\Theta_i}(y)$ better than $P_{\Theta_j}(y)$, if the observed likelihood of the first statistical model $L_y(\Theta_i)$ is greater than that of the latter $L_y(\Theta_j)$.

This law gives us some indication of how to interpret likelihoods. This is to be done without any probability concepts if Hacking's (1965) ideas are to be followed. He has transferred the law of likelihood for discrete distribution in a proposition including a language of logic, which uses no nonlogical terms. Thus we can use it as a joint proposition and formulate a confirmation function c(|) accordig to Seidenfeld (1979), which brings together observed likelihood and the measurement of inference uncertainty through the following equivalence relation

$$c([M,P]_i|y) > c([M,P]_j|y) \iff L_y(\Theta_i) > L_y(\Theta_j).$$

This relation also demonstrates how we can measure inference uncertainty without any probability assertion. Confirmation function is defined as a rational person's assessment of the relative support for various hypotheses given any well-formed evidential basis. This interpretation for the law of likelihood sustains our aims to provide a measure for inference uncertainty without probability concepts. It should be noted that we do not need the above mentioned language of logic in searching our information theoretical interpretation for the law of likelihood.

The use of likelihood presupposes that it is statistically sufficient for inference. This is ascertained in the likelihood principle.

The likelihood principle: All the information in the data {y} concerning the mutual plausibility of two statistical models is contained in the observed likelihood ratio of these models $\lambda = L_y(\Theta_i)/L_y(\Theta_j)$.

The likelihood principle contains the sufficiency concept in the sense of measuring inference uncertainty. This concept is not the same as sufficient statistic in estimation, the existence of which is often verified by the factorization theorem. Fraser (1979), Birnbaum (1969) and Edwards (1972) have demonstrated that at least on the basis of a weak likelihood principle the data {y} can be replaced by the likelihood $L_y(\Theta)$, which at least contains the same inference uncertainty as the data (see Fraser p. 74). The likelihood ratio is,then,as the ratio of two sufficient statistics also a sufficient statistic.

Birnbaum demonstrates in another way as well that likelihood is a sufficient statistic in the sense of statistical evidence. As is known, likelihood differs only by an arbitrary constant from the joint probability of a sample. Consequently, if $P_\Theta(y)$ is a sufficient statistic in the sense of statistical evidence then $L_y(\Theta) = c\ P_\Theta(y)$ is it as well. Statistical evidence is defined by Birnbaum as the acceptance order of statistical models.

A part of the difficulty of the likelihood principle arises from the fact that the joint probability of a sample does not necessary depend on the population values. Birnbaum (1969) gives an example of this misconvenience concerning random sampling without replacement from a finite population. In the narrower area for the case of parametric hypotheses, we consider the likelihood function to be exhaustive with regard to the parameter given the class of possible likelihood functions. Thus observed likelihood and the class of likelihood functions, possible under the joint probability of a sample, contain all the inferential content of the data, given that the model under consideration is appropriate.

Finally the law of likelihood and the likelihood principle are thrown into a single axiom.

The likelihood axiom: The relative plausibility of two statistical models i and j is displayed sufficiently in the sense of measuring inference uncertainty from the data {y} as a likelihood ratio $\lambda = L\ (\Theta_i)/L_y(\Theta_j)$.

The likelihood axiom asserts that the ratio of likelihoods is useful only for the comparison of rival hypotheses. This property is included in the method of support, whose central inference concept is the natural logarithm of a likelihood ratio. Really, support is a one-to-one transformation of the likelihood ratio and thus preserves all the properties of the likelihood ratio itself. Three main concepts of the method of support are given in Definitions 3-5.

Definition 3. Support function is the natural logarithm of the likelihood function $S_y(\Theta) = \log L_y(\Theta) = \log P_\Theta(y) + C_y$.

The likelihood function has an arbitrary constant as coefficient, which adds the term C_y into $S_y(\Theta)$. In chapter 3-2,dealing with the norming of the support function,the choice of this constant will be examined. In the following we will drop the reference to y from $S_y(\Theta)$ and index the support function according to the alternative statistical models $[M_j, P_j]$, i.e. $S_j(\Theta)$.

According to the likelihood axiom,an observed value for support function is meaningless in an inferential sense. After we have given the information theoretical interpretation for the method of support it is seen that observed likelihood also has inferential content in a local sense. For a moment we shall proceed in the frame of that axiom and define the difference between two models as observed support.

Definition 4. The difference of two observed supports is the support $S\{i;j\} = S_i(\Theta) - S_j(\Theta)$, which indicates to what extent the support given by the data increases or decreases when the statistical model i is compared with the statistical model j.

By means of a suitable order of the models {i,j} or {j,i}, support can be restricted as non-positive, in which case its interpretation is: support measures to what extent a statistical model gets less support than the model better supported by the data. Defined in this way decrease in support is in question, when we move from one statistical model to another.

Definition 5. The use of support function and support as a statistical inference model is called the method of support.

There are many examples on the application of the method of support in statistical literature. However, Edwards (1972) must be seen as one of the few consistent users of the method. In particular, he emphasizes its use in testing scientific hypotheses, where even a small difference between two hypotheses, which can be revealed by support often is significant. Yet one crucial point has been revealed in the empirical researches performed by Edwards (1972), Cole (1975), Kalbfleisch and Support (1970), Fraser (1979) and so on, namely, the clear interpretation for the support difference and its ability to measure inference uncertainty is lacking. This problem is considered in the following chapter.

3-1 Information Theoretical Interpretation for the Method of Support

We distinguish three main interpretations for the likelihood ratio: Bayesian, information theoretical and a practical one. In the Bayesian version (see Birnbaum p.136) the posterior probabilities depend upon the data only through its likelihood function so that relative posterior probabilities differ from the corresponding relative prior probabilities in the factor which is just the likelihood ratio. The same arguments are considered by Lindley (1956), Good (1950) and Pearson (1962). In the information theoretical interpretation a Bayesian view has many times been borrowed so that an a prior calculated information is increased or decreased by the information carried by the data quoted as the logarithm of the likelihood ratio - see for example Theil(1972) and Lindley (1956). Edwards (1972) uses a practical interpretation for the log likelihood ratio and asserts that it measures increase or decrease of support provided by the data. This is one version of the interpretation of the confirmation function, which measures the relative support for one hypothesis relative to another. However, the discussion of Hacking (1965) and Seidenfeld (1979) should also be kept in mind. The foregoing has revealed those interpretational difficulties which efficiently prevent the common use of likelihood ratio or its logarithm for statistical inference.

Our interpretation is based on information theoretical concepts without any Bayesian reasoning. First we shall show that the Kullback's information and the Shannon's entropy are not good for an information theoretical interpretation for an observed likelihood ratio. The solution is to be found in Rényi's local uncertainty concept.

Kullback's information is the expectation

$$I\{i:j\} = \sum_{y \varepsilon Y} P_{\Theta_i}(y) \log \frac{P_{\Theta_i}(y)}{P_{\Theta_j}(y)} .$$

It denotes the average information carried by the data in the discrimination between the models $[M,P]_i$ ja $[M,P]_j$. According to Definition 1, $P_\Theta(y) = L_y(\Theta)$, which for Kullback's information implies the formula

$$\begin{aligned} I\{i:j\} &= E_{\Theta_i}[\log P_{\Theta_i}(y) - \log P_{\Theta_j}(y)] \\ &= E_{\Theta_i}[\log L_y(\Theta_i) - \log L_y(\Theta_j)] \\ &= E_{\Theta_i}[S\{i:j\}] \end{aligned}$$

The last expression indicates that Kullback's information and support are equal only in the sense of expectation. Thus Kullback's information is deleted as an information theoretical interpretation for an observed likelihood ratio in an instant sample.

Shannon's entropy is one of the central uncertainty concepts of the coding theory, for which several sets of axioms have been developed. From these Ash's (1965) approach is quoted here.

Consider a finite probability distribution $p_i \geq 0$ $(i = 1,\ldots,m)$ $\sum_i^m p_i = 1$. Then Shannon's entropy is

$$H_m = H_m(p_1,p_2,\ldots,p_m) = -\sum_{i=1}^{m} p_i \log_2 p_i,$$

which has e.g. the following properties:

1.) $H_m \geq 0$

2.) $H_m = 0$, if $p_{i_0} = 1$ and $p_i = 0$ $(1 \leq i \leq m\colon i \neq i_0)$,

3.) $H_m(p_1,\ldots,p_m) \leq H_m(\frac{1}{m},\ldots,\frac{1}{m})$.

In the case of a single event Shannon's entropy can only be interpreted as $E[-\log_2 p_i]$, for which a weak presentation corresponding to support can be developed,

$$\begin{aligned} H_m(P_{\Theta_i}) - H_m(P_{\Theta_j}) &= E_{\Theta_i}[-\log_2 P_{\Theta_i}(y)] - E_{\Theta_j}[-\log_2 P_{\Theta_j}(y)] \\ &= E_{\Theta_i}[-\log_2 L_y(\Theta_i)] - E_{\Theta_j}[-\log_2 L_y(\Theta_j)] \\ &= \{E_{\Theta_i}[-S(\Theta_i)] - E_{\Theta_j}[-S(\Theta_j)]\}\log 2. \end{aligned}$$

So by using the concepts of Shannon's entropy it is difficult to find an information theoretical interpretation for support. The difference of two Shannon's entropies is namely minus the difference of the expectations of the two supports $E_{\Theta_i}[S(\Theta_i)]$ and $E_{\Theta_j}[S(\Theta_j)]$. The coefficient -1 has no practical significance since instead of measuring increase in uncertainty we can think of measuring decrease in certainty. At the first glance the expectation property excludes Shannon's entropy in its original sense as an information theoretical interpretation for the method of support, since when expectation is formed the probabilities of the unobserved events are needed, as well.

Statistical inference connected with instant sample and, consequently, the main field of application of the method of support has one observation {y} and the probabilities $P_{\Theta_i}(y)$ linked with it as its special feature. The impossibility of repetition brings about that the statistician does not have the possibility to get further data.

A natural and useful model for this kind of situation is yielded by Rényi's (1970) incomplete probability distribution and Rényi's local uncertainty concept, defined on it. These will be discussed next.

We examine a finite set of positive numbers $\underline{\underline{p}} = \{p_1, \dots, p_m\}$, the sum of which $v(\underline{\underline{p}})$ has the property

$$0 < v(\underline{\underline{p}}) = \sum_{i=1}^{m} p_i \leq 1.$$

If $v(\underline{\underline{p}}) = 1$ an ordinary probability distribution is in question, but if $v(\underline{\underline{p}}) < 1$ we are dealing with an incomplete probability distribution.

Rényi (1970) (p. 579) assigns to every incomplete probability distribution an uncertainty concept by defining

$$H_R(\underline{\underline{p}}) = \frac{\sum_{k=1}^{m} p_k \log_2(1/p_k)}{\sum_{k=1}^{m} p_k} .$$

Especially, for the case m = 1, $\underline{\underline{p}} = \{p\}$ we get

$$H_R(\underline{\underline{p}}) = -\log_2(p),$$

which can be interpreted as uncertainty associated with occurence of an event with probability p (cf. Rényi (1970) p. 572). In what follows we use the notation

$$H_R(p) = H_R(\underline{\underline{p}}) \text{ , if } \underline{\underline{p}} = \{p\}.$$

Definition 6. The Rényi's local uncertainty of a single event {y} in relation to the probability distribution $P_\Theta(y)$ is $H_R(P_\Theta(y)) = -\log_2 P_\Theta(y)$.

The linking of statistical models with local uncertainties makes it possible to compare the models with one another by using differences $H_R\{i;j\} = -\log_2 P_{\Theta i}(y) - (-\log_2 P_{\Theta j}(y))$. According to Rényi, those differences can be interpreted as a gain or loss of uncertainty when the statistical model $[M,P]_i$ is changed into the model $[M,P]_j$.

In connection with this information concept Good's (1950) and later Pitman's (1979) works must be mentioned, according to which S{i;j} is the discrimination between statistical models based on one observed event $y \varepsilon Y$. Without information theoretical argumentation this refers to the ability of a single observation and not of its expectation to measure inference uncertainty in comparing two statistical models. Later Good together with Osteyee (1974) has presented an axiomatic system for the uncertainty of one event. They arrive at the measure -log P(A), but this does not fit as naturally an instant sample case as Rényi's local uncertainty concept.

The connection between support and Rényi's local uncertainty is obtained in the following way: from Definition 6 it follows that

$$H_R(P_{\Theta_i}(y)) = -\log_2 P_{\Theta_i}(y) \text{ and}$$

$$H_R(P_{\Theta_j}(y)) = -\log_2 P_{\Theta_j}(y),$$

which are put into the Definition 3 of support,

$$\begin{aligned} S\{i;j\} &= S(\Theta_i) - S(\Theta_j) \\ &= -\log P_{\Theta_i}(y) - (-\log P_{\Theta_j}(y)) \\ &= (-\log 2)\{\log_2 P_{\Theta_i}(y)\} - (-\log 2)\{\log_2 P_{\Theta_j}(y)\} \\ &= \{-\log 2\}\{-H_R(P_{\Theta_i}(y)) + H_R(P_{\Theta_j}(y)\}. \end{aligned}$$

We have thus proved the following Theorem 1.

Theorem 1. The support S{i:j} is minus the difference of two Rényi's local uncertainties multiplied by the module log 2.

Theorem 1 gives the basis for the statistical analysis of instant samples since it gives a clear interpretation to inference uncertainty measured from the total evidence including only two elements, the observed data {y} and a parameter space θ. This directly corresponds to the total evidence required by the method of support. Although Rényi uses a 2-based logarithm system, which derives from the definition of Shannon's entropy it has no significance for the interpretation of support. The difference of two local uncertainties in Rényi's terminology means either gain or loss of information.

Consequently, the interpretation of support is thus fully information theoretical. Although we are dealing with minus the difference of Rényi's local uncertainties,it does not make any difference to the statistician whether he communicates the results of the comparison of the models i and j as increase of uncertainty (Rényi's local uncertainty) or as decrease of certainty (method of support).

The comparison of the method of support and the information theoretical concepts yielded an unambiguous result. Support is minus the difference of two Rényi's local uncertainties multiplied by the module log 2. Interpretationally it corresponds to the confirmation function c(|) of inductive logic in the sense of plausibility order. In a statistical inference situation we should keep to Edwards' communication concept, according to which support difference measures the decrease or increase of support when we move from one statistical model to another. The use of Rényi's local uncertainty presupposes a considerably concise total evidence {y,θ}. The questions of support communication and norming still remain open. They will be discussed in chapters 3-2 and 3-4.

3-2 Communication of Inference Results

In the method of support the production of inference results begins with the determination of an observed support function and support. The observed support function is displayed either graphically or as observed values of communication parameters. The most important of them are connected with the maximum of the support function and with the behaviour of the function in the neighbourhood of this value. The communication of the support function is connected with interval and point estimation and in the shape of support with the testing of statistical hypotheses. The communication concepts are presented in the case where the support function S(Θ) is a regular continuous function of one parameter with unambiguous first and second derivates. The communication parameters are included in definitions 7 to 12. The words evaluator and evaluation are introduced as substitutes for estimator and estimation according to Edwards (1972).

Definition 7. The evaluate $\hat{\Theta}$ is the solution $\max_{\Theta} S(\Theta) = S(\hat{\Theta})$.

Evaluation is connnected with the maximum likelihood estimation. The value of the support function at $\Theta = \hat{\Theta}$ yields maximum support for the model and thus the least local uncertainty. The evaluate is a communication parameter linked with the position of the support function.

In comparing statistical models it is interesting to know how the support function behaves in the neighbourhood of its maximum. Derivates are suitable for the communication of this property. The second derivate, much discussed in literature, primarily comes into question.

Definition 8. Minus the second derivate of support function is the information, $I_\Theta = - \frac{dS^2}{d\Theta^2}$. The $I_{\hat{\Theta}}$ is called observed (Fisher) information.

The observed (Fisher) information is geometrically interpreted as the curvature of the support function in the neighbourhood of the evaluate. It denotes the speed by which the support reduces when we move from the evaluate to the parameter values in its neighbourhood. Pitman (1979) gives a well-argumented interpretation for Fisher information, which in his opinion measures the sensitivity of a statistical model in the neighbourhood of the evaluate. More meaningful communication parameters of the support function are two transformations of the observed (Fisher) information in which the order of magnitude of the observed support can also be expressed. These are radius of curvature of the support curve at its maximum and the square root of the radius.

Definition 9. The reciprocal of observed (Fisher) information $\hat{w}^2 = 1/I_{\hat{\Theta}}$ is called observed radius of curvature.

By the circle of curvature belonging to the point $S(\Theta)$ of the support function we mean a circle whose radius is $\hat{w}^2$ and whose centre is situated on the normal of the observed support function, on the concave·side of the curve. The radius is a nonlinear function of the evaluate and, for instance, it does not directly communicate to what extent support decreases if, for example, we move from the evaluate $\hat{\Theta}$ to the parameter value $\Theta \neq \hat{\Theta}$. In this sense the square root $\hat{w}$ of the radius, which is geometrically thought to be the chord of a circle with the radius $\hat{w}^2$, is a more perspicious communication parameter for the shape of support curve at its maximum. If it is put on the perpendicular of the observed support function we can, from the points common to its end points and the curve of the support function,formulate the space of two parameter values $\{\Theta_U,\Theta_L\}$, in the field of which the support function value $S(\Theta) \geq S(\Theta_U) = S(\Theta_L)$. The square root $\hat{w}$ of the radius thus represents a comparable measure from one inference situation to another, because the $\{\Theta_U,\Theta_L\}$ totally depends on the interval $\hat{w}$. In this sense it is defined as accuracy.

Definition 10. Accuracy is the square root $\hat{w}$ of the observed radius of the support function curvature at its maximum. Notationally $\hat{w} = \frac{1}{\sqrt{I_{\hat{\Theta}}}}$.

Accuracy in the method of support is analogous to the standard error of estimation theory. It can be used in the same sense , however, with the difference that inference uncertainty here equals support.

The most important communication concept of the method of support is support $S\{i:j\}$ itself, which is the result of the comparison of two statistical models. If support has been defined in the way that the larger support is substracted from the smaller we get a normed support as result. We label this communication form of the support as $\hat{S}\{i:j\} \in (-\infty,0]$. Norming is dicussed later in greater detail in Chapter 3-4. As such it is difficult to link them with some observable frame of comparison unless some transformations are made. Depending of the inference situation Edwards (1972), for example, recommends the transformation exp $\hat{S}\{i:j\}$, by means of which support is transformed to the odds, like 1:m. An interesting communication interpretation has been presented by Good (1950). He takes the concept of decibel in acoustics to be used as the communication form of support, in which case $\hat{S}\{i:j\}$ implies the noise corresponding to $10(\log 10)|\hat{S}(i:j)|$db. It must be noted that Good does not present support at the level of local support but as the difference of a priori and observed support. The alternative communication forms of support are presented in Definition 11.

Definition 11. The observed support $\hat{S}\{i:j\}$ is communicated
a) as a number $\hat{S} \varepsilon (-\infty,0]$
b) as the odds $n:m = \exp \hat{S}\{i:j\}:1$
c) in decibels $(10 \log 10)|\hat{S}\{i:j\}|$.

The communication parameters of the method of support in Definitions 7-11 are primaly applicable to such an inference situation in which the statistical model mainly remains unchanged and only its parameter Θ is varied. If different statistical models are compared the only thing to be communicated is the support $S\{i:j\}$ or its transformation, which in both cases is a point function. Its use has justifiably been criticized by Fraser (1979) and Menges (1973). The situation is different, however, if support is used at the level of ordinal scale like Lindsey (1974a) and Kalbfleisch and Sprott (1970) have done in their own analogous applications. In the comparison of the statistical models $[M,P]_1,\ldots,[M,P]_k$ the maximum value of each support function $S_j(\hat{\Theta})$ is determined, after which the order of these observed values yields plausibility order of the statistical models, which is also regarded as one form of communication of the method of support.

Definition 12. The order of observed maximum supports $S_j(\hat{\Theta}_j)$ is the plausibility order of the corresponding statistical models.

This form of communication is deduced from the law of likelihood because the logarithm of likelihood preserves the order relation. Communication at the level of ordinal scale is a natural alternative in a situation where supports are compared with one another without them being calculated from the same support function. Different models have, of course, different support functions. Comparison between models, which is more precise than that based on plausibility order, can be accomplished if the support functions for each model can be deduced from the same function type or if the statistical models can be transformed to the same statistical model by means of reparametrization, for example. These questions are linked with the norming of the support function, which will be discussed in Chapter 3 - 4.

The form of communication of inference results is a transformation of Rényi's local uncertainty. Consequently, it is at the same time inference uncertainty in the sense of the method of support. Within a given statistical model it is natural to compare different θ values with the evaluate $\hat{\theta}$ using support $\hat{S}\{\theta:\hat{\theta}\} = S(\theta) - S(\hat{\theta})$. A suitable specification is to state the accuracy $\hat{w}$. In the comparison between statistical models the method of support leads to (point) function, by means of which the plausibility order of the corresponding statistical models is displayed. The direction of the order is determined in relation to the statistical model which has obtained the maximum support. The model in question represents the least local uncertainty in the sense of Rényi's uncertainty concept.

3-3 Evaluators

In connection with the communication concepts the point evaluate $\hat{\Theta}$ describing the position of the maximum of the support function was defined. The evaluates are connected with the point estimation of statistical inference. According to Fraser (1979) and Menges (1973), point estimation is no statistical inference at all, but calculation of a statistic, which describes the data. In the method of support the point evaluate is connected with local uncertainty, which measures inference uncertainty. In that way point evaluates, in the sense of the method of support, are part of statistical inference where the communication of inference results and the measuring of inference uncertainty are combined.

As an estimation method the calculation of evaluates means minimizing local uncertainty. In this sense estimates produced by the method of support are estimates of the least local uncertainty (LLU estimates). They are tied to one observation like the instant sample and to the statistical model used. Below three kinds of LLU estimates will be defined: the m-unit support neighbourhood, the sum evaluate and the joint evaluate.

The concept in the method of support corresponding to interval estimation is the determination of the m-unit support neighbourhood. Definition 13 presupposes that support function is a regular function of the parameter.

Definition 13. The m-unit support neighbourhood is the set of all parameter values at which the support is not more than m-units below the maximum i.e. $S(\Theta) - S(\hat{\Theta}) \geq - m$.

For a one-dimensional parameter the m-unit support neighbourhood leads in a regular case to the m-unit support limits, which are displayed as $\{\Theta | S(\Theta) - S(\hat{\Theta}) \geq - m\}$. This resembles the construction of a confidence interval. Because it is to be operated with the concept of local uncertainty the m-unit support limits are not always placed symmetrically around the evaluate $\hat{\Theta}$. The LLU estimation differs from the usual interval estimation. For example, the set $\{\Theta | S(\Theta) - S(\hat{\Theta}) \geq - m\}$ may consist of several disjoint intervals. Also, we don't have any confidence coefficient in the construction of the m-unit support neighbourhoods. It should be observed that the m-neighbourhood is contained in the parameter space $\boldsymbol{\Theta}$, which is not always the case in the classical interval estimation.

In the following the point evaluate $\hat{\Theta}$ and the accuracy $\hat{w}$ will be extended to inference situations in which the data (treated as an instant sample) either consist of several subdata independent of one another or it can be classified into mutually exclusive groups. First the support functions $S_h(\Theta_h)$ of the subdata are determined. These are used in the calculation of joint evaluates by means of summing or weighing. It is assumed below that the subdata are independent of one another and that the corresponding support functions are of the form,

$$S_h(\Theta) = a_h + b_h\Theta + c_h\Theta^2$$

thus the quadratic approximation of $S(\Theta)$ is adequate. Then,

$$S_h(\Theta) = S_h(\hat{\Theta}_h) - \frac{1}{2\hat{w}_h^2}(\Theta - \hat{\Theta}_h)^2 \text{ ;}$$

and the sum support function of the subdata $y^{(1)},\ldots,y^{(h)},\ldots,y^{(k)}$ which are independent of one another, is

$$S(\Theta) = \sum_{h=1}^{k} S_h(\Theta) = \sum_{h=1}^{k} S_h(\hat{\Theta}_h) - (1/2) \sum_{h=1}^{k} \frac{(\Theta_h - \hat{\Theta}_h)^2}{\hat{w}_h^2} \text{ .}$$

This combination property of support function leads to the concept of joint evaluate, which is

$$\hat{\hat{\Theta}} = \sum_{h=1}^{k} \frac{\hat{\Theta}_h}{\hat{w}_h^2} \Big/ \sum_{h=1}^{k} \hat{w}_h^{-2},$$

and whose accuracy is

$$\hat{w}_{\hat{\hat{\Theta}}} = \surd \sum_{h=1}^{k} \hat{w}_h^{-2}.$$

The joint evaluate is applicable in an inference situation where we want to compare two statistical models, the evaluate of the one being the subdata specific evaluate $\hat{\Theta}^{(h)}$ and of the other the joint evaluate $\hat{\hat{\Theta}}$. The decrease in support is then caused by the division of the data by grouping or by combining different data, and it is

$$S_N\{\Theta^{(h)}:\hat{\hat{\Theta}}\} = \sum_{h=1}^{k} \frac{1}{2\hat{w}_h^2} (\hat{\Theta}_h - \hat{\hat{\Theta}})^2.$$

The point and interval evaluates and their various combinations at the same time correspond to both the communication concepts and the measurement of inference uncertainty. They are analogous with point and interval estimations in other inference models. Above they were discussed in a restricted case,in which the support function is a regular function of the parameter and the function, in addition, quadratic. These restrictions, in general, hold in the case of large instant samples.

3-4 Norming of Support Function

The norming of the support function aims at producing a form of communication which is comparable from one inference situation to another. The central concept of communication namely the support should be normed that it measures the same property in different situations. This can be achieved by the norming of the support function, in which case the same support function $\hat{S}(\Theta)$ is always used, or by the norming of the statistical model, in which case the support is always used in connection with the same statistical model $\hat{P}_{\Theta}(y)$.

In the norming of the support function its range is defined and such a transformation of the function is looked for, by means of which $S(\Theta)$ is always transformed into same function $\hat{S}(\Theta)$ so that

a) $\hat{S}(\Theta) \, \varepsilon \, [K_L \, , K_U]$, where K_L, K_U, K_C are constants and $\hat{S}(\Theta)$ is the normed support function.

b) $\max\limits_{\theta \varepsilon \boldsymbol{\Theta}} \hat{S}(\Theta) = K_C$ or

c) $q\{S(g(\Theta))\} = \hat{S}(\Theta)$, where g is a suitable chosen one-to -one function of θ and q indicates functional transformation to express $S(\Theta)$ in the normed form as $\hat{S}(\Theta)$.

The first alternative starting point is norming the statistical model itself with regard to its own evaluate vector. It follows from Definition 7 that if $S(\Theta)$ is regular and it has only one maximum, then $\max\limits_{\Theta} S(\Theta) = S(\hat{\Theta})$. The normed support function is then

$$\hat{S}(\Theta) = S(\Theta) - S(\hat{\Theta}). \tag{3.1}$$

Norming means the parallel displacement of the coordinate axis which includes the properties a) and b), for

$$\hat{S}(\Theta) \, \varepsilon \, (-\infty, 0]$$

$$\max \hat{S}(\Theta) = 0$$

This norming retains the original form of the support function.

The second alternative is to find such a transformation of the support function by means of which the result is always the same type of function. Edwards (1972) recommends the quadratic support function which corresponds to the statistical model of the normal distribution $N(\theta,\sigma^2)$ (σ^2known). Kalbfleisch and Sprott (1970) come to the same conclusion when examining the relative likelihood $R(\theta)$. An unknown support function can often be approximated by the first three terms of the Taylor polynomial expanded about the evaluate $\hat{\theta}$, which also yields the quaratic support function

$$\hat{S}_N(\theta) = S(\hat{\theta}) + (1/2)(\theta - \hat{\theta})^2 \frac{d^2S}{d\theta^2}(\hat{\theta}) \qquad (3.2)$$

$$= a + b\theta + c\theta^2,$$

where the constants a, b and c are functions of the evaluate and accuracy.

Norming in this case has been made in the support function and thus it does not necessarily follow from the statistical model of the normal distribution. The observed support function of the normal distribution model $P \leftrightarrow N(\theta,\sigma^2)$ is always of the form $\hat{S}_N(\theta)$ as has been pointed out by Edwards (1972). If the original support function is asymmetrical in the neighbourhood of its evaluate, the quadratic approximation is not always useful. Asymmetricality can often be corrected by a one-to-one transformation of the parameter. The quadratic support function is important in the case of large instant samples for which the distributions of the statistics approach their own, in many cases normal, limit distributions.

The third alternative in norming is to reparametrize the original statistical model $P_\theta(y)$ always into the same statistical model $\hat{P}_\theta(y)$. The natural starting point in this case is the statistical model which best fits the data; the multinomial distribution has often been recommended (see for example Lindsey (1974 a and b). The multinomial distribution is the simplest statistical model for which the observed sample is most probable. In philosophical jargon it corresponds to a model for an event which had to happen. When put in the order of magnitude, the observations form the discrete n-tuple $[y_{(1)},\dots,y_{(i)},\dots y_{(n)}]$. By combining possibly tied observations with each other, corresponding frequency distribution is obtained. This is,

$$y_1,\dots,y_i,\dots,y_k$$

$$n_1,\dots,n_i,\dots,n_k,$$

where y_i is the result of measuring a sampling unit i and n_i is the observed frequency.

The statistical model of the above frequency distribution is the multionomial distribution

$$\hat{P}_\Theta(y) \leftrightarrow P(N = n) = \frac{n!}{n_1! \dots n_k!} \prod_{i=1}^{k} \Theta_i^{n_i},$$

where $0 \leq \Theta_i \leq 1$ when $i = 1,2,\dots,k$,

$$\sum_{i=1}^{k}\Theta_i = 1 \text{ and } \sum_{i=1}^{k} n_i = n.$$

The support function of the parameter vector Θ is with a properly chosen constant

$$S_M(\Theta) = \sum_{i=1}^{k} n_i \log \Theta_i ,$$

which yields the solution of the evaluate vector

$$\hat{\Theta} = \{\hat{\Theta}_1 = n_1/n,\dots,\hat{\Theta}_i = n_i/n,\dots,\hat{\Theta}_k = n_k/n\}.$$

Thus, the least local uncertainty is

$$\max_{\Theta} S_M(\Theta) = \sum_{i=1}^{k} n_i \log \hat{\Theta}_i = S_M(\hat{\Theta})$$

The statistical model under study is transformed into a multinomial model by the following reparametrization, so that

$$\Theta_i = \begin{cases} P_\Theta(y = y_i), & \text{a discrete statistical model} \\ f_\Theta(y_i)\,\Delta y_i, & \text{a continuous statistical model where } \Delta y_i \text{ is a unit of measurement.} \end{cases}$$

Using the above described multinomial model as the statistical model $P_\Theta(y)$ we always get the same normed support function, namely

$$\hat{S}_M(\Theta) = \sum_{i=1}^{k} n_i \,(\log \Theta_i - \log \hat{\Theta}_i) \leq 0. \qquad (3.3)$$

The normed support function $\hat{S}_M(\Theta)$ is commensurate between different statistical models, and thus the plausibility order of the models can be achieved by means of it. A more accurate result is not obtained because Rényi's local uncertainty due to approximation by the multinomial model $\hat{P}_\Theta(y)$ is cancelled when the support $\hat{S}_M(\Theta_i) - \hat{S}_M(\Theta_j)$ in the comparison of two models is calculated. The result is the support difference on the condition that the statistical models have first been reparametrized into multinomial models.

Reparametrization often creates a grouping bias which is caused, for example, by the discretization of a continuous statistical model into a multinomial model. As a matter of fact, we should make a distinction between two procedures: the discretization of the continuous model and the classification of the data. These procedures are often mixed up with each other. Below, it is shown how classification, in the sense of support, always reduces the informativeness of the data. Let the original discrete frequency distribution be k-categorical, in which case the local uncertainty of the multinomial model is

$$S(\hat{\Theta}) = \sum_{i=1}^{k} n_i \log \hat{\Theta}_i .$$

If categories are combined, it follows that at least in one category $\Theta'_j = \Theta_i + \Theta_{i+1}$, and $n'_j = n_i + n_{i+1}$, and so

$$S_k(\hat{\Theta}) = \sum_{i=1}^{k} n_i \log \hat{\Theta}_i \leq \sum_{i=1}^{k'} n'_i \log \hat{\Theta}'_i = S_{k'}(\hat{\Theta}),$$

where k and k´are the numbers of the corresponding categories.

By combining of all categories we get $S_1(\Theta) = N \log 1 =$ ●. If we examine two models simultaneously, one of which is $\hat{P}_\Theta(y)$ and the other is, for example, a statistical model of a continuous random variable, like the normal distribution, $P \leftrightarrow P_\Theta(y) \sim N(\Theta,\sigma^2)$ the range of which is divided by classification, then alone by decreasing the number of categories a complete fit in the sense of the support S{i:j} is obtained. We are here dealing with a problem which in statistical literature is treated as optimal classification. In the empirical part of the present study, we use Mineo's (1979) method of natural classes which is interpreted by means of the concepts of support. It is included in the discussion of the regression residuals in Chapter 4-3-4.

The norming in the method of support aims at such a presentation of the support which makes it possible to measure the decrease of support when two statistical models are compared with each other. The observed value of the support is in that case always non-positive and it is also used like an ordinal scale as the plausibility order of the models. The alternative ways of norming, presented above, are shown in the table below. These notations will be used later.

Observed support	Explanation or relation
$S_j(\Theta)$	The support function of the statistical model $[M,P]_j$ with a given value of the parameter vector.
$\hat{S}_j(\Theta)$	$= S_j(\Theta) - S_j(\hat{\Theta})$
$\hat{S}\{i:j\}$	$= \hat{S}_i(\Theta) - \hat{S}_j(\Theta)$, on condition that $\hat{S}_i(\Theta) \leq \hat{S}_j(\Theta)$
$S_N(\Theta)$	$= a + b\Theta + c\Theta^2$, quadratic approximation
$\hat{S}_M(\Theta)$	$= \sum_{i=1}^{k} n_i(\log \Theta_i - \log \hat{\Theta}_i)$, multinomial support

Normed supports can be used for the determination of statistical tests and evaluates corresponding to estimation. The aim of norming is to find a measure of inference uncertainty commesurate from one inference situation to another. The decrease in support is here considered as such a measure. It corresponds to the inference risks used in the other statistical inference models analyzed earlier in this study. A special case in this connection is the support $\hat{S}_M(\Theta)$ in which the statistical model under comparison is a multinomial model best supported by data. $S_M(\hat{\Theta})$ is, however, cancelled if two statistical models, i and j, are compared with each other by determining the difference of the supports. In this way we get the support which provides the plausibility order between statistical models in the sense of support yielded by the data. This kind of an interpretation is not contradictory to the fact that there always is in question the difference between two Rényi's local uncertainties.

3-5 Support Tests

In a statistical test, the hypothesis H_o concerns either the probability distribution $P_\Theta(y)$ or its parameter Θ. Thus the hypothesis is expressed as $H_o : P_\Theta(y) = P'_\Theta(y)$ or $\Theta = \Theta_o$. In the method of support, Edwards (1972) considers a statistical test as an insertion into the support function from which the support between hypothetical and arbitary models or parameter values is calculated.. A natural way of making the insertion is to use a normed support function in which case we are comparing the hypothetical parameter value Θ_o with the evaluate $\hat{\Theta}$.

Definition 14. The support test of the statistical hypothesis H; $\Theta = \Theta_o$ is the insertion of $P_{\Theta_o}(y)$ into the normed support function $\hat{S}(\Theta)$, $\hat{S}_N(\Theta)$ or $\hat{S}_M(\theta)$ from which the support measuring inference uncertainty is obtained.

The support test is a hierarchial test the first stage of which is the comparison of statistical models and the second stage is the comparison of the parameter vectors of the chosen statistical model. Borth (1975), for example, deals with a test analogous to the support test in such way that the expected total entropy is divided into a model component and a parameter component. Lindsey (1974b) uses $\hat{S}_M(\Theta)$ for the comparison of statistical models alone. As a statistical test, the support test produces the plausibility order of the hypotheses, and some order of magnitude of the support, for example $|S\{i:j\}| \leq m$, can be used as the criterion of plausibility. If the support is smaller than m, all the hypotheses belong to the same equivalence class and so their mutual plausibility does not matter. However, Edwards (1972) warns of this kind of standard technique which in the frequentist inference has led to scientifically questionable inferences, as Sterling (1959) has demonstrated.

The inference uncertainty of the support test is expressed as the communication forms of the method of support: the support, the odds or desibels, depending on the inference situation. The support test will be discussed in more detail in connection with the regression analysis in Chapter 4-2.

3-6 The Method of Support and Inference Situation

The properties of the method of support and Requirements 1-3 of a statistical inference situation are considered. In Theorem 2 it is proven that the total evidence provided by an instant sample with a supplement by a parameter space coincides with the total evidence required by the method of support. Thus we shall conclude that the method of support is an inference model for instant samples.

The general requirements of a statistical inference situation concern the decomposition of an inference situation into the model triplet [M,P,I], the existence of the total evidence and the measuring of inference uncertainty. Requirement 1 refers to the organization of a statistical inference situation into the model triplet [M,P,I]. Its components M and P do not as such place special restrictions on the use of the method of support. The component I denotes the observed support function $S(\theta)$. Thus, there exists the relation $I \leftrightarrow S(\theta)$ between the inference model and the method of support.The only restriction is the condition that it must be possible to write down the statistical model into the form of a likelihood function the transformation of which is $S(\theta)$. It is convenient with regard to the inference results if $S(\theta)$ is a regular function with only one maximum because it is natural to represent the support as the difference between the least local uncertainties. It is possible to influence the regularity of the support function by the choice of the statistical model. Thus the method of support is a statistical inference model in the sense of Requirement 1.

Requirements 2 and 3 concern the total evidence at the statistician's disposal and the measuring of inference uncertainty. For the choice of an inference model Requirement 3 presupposes that the extent of the required total evidence coincides with the total evidence provided by the inference situation.

As was mentioned earlier, the method of support needs a total evidence which includes only two components: data y and parameter space Θ, abbreviated as $\{y,\Theta\}$. According to Theorem 1, we can measure with this total evidence inference uncertainty expressed as the difference of two Rényi's local uncertainties. Thus the method of support fulfils Requirement 2. On the other hand, recalling Definition 2 we see that the total evidence provided by an instant sample is only data y. If we add a parameter space Θ as a supplement we get the total evidence at the statistician's disposal whose extent is $\{y,\Theta\}$. To add a parameter space is very natural, because it represents those unknowns which are of interest at all. The above coincidence of two total evidences leads to the following Theorem 2.

Theorem 2. The method of support is a statistical inference model for instant samples and non-repeated designed experiments.

The method of support has been used as a statistical inference model in the sense of Theorem 2, although the support has not been interpreted as inference uncertainty. Edwards (1972), for example, has often used the method of support as an inference model for non-repeated experiments in genetics. Cole (1975) has applied the method of support to medical instant samples gathered from various geographical areas. His samples have been probability samples including 100-2000 persons and they have not been repeated. Other practical studies are to be found which have made use of the method of support and so proved the usefulness of this inference model. How frequent is an instant sample in social sciences is revealed by Hagood (1969). For example an interview of people at a certain moment t_o represents the population at that moment. Continuous changes in social and cultural phenomena exclude the possibility of collecting a sample or experimental data in successively similar occasions. In particular, it is possible to speak of uniqueness in connection with the empirical data of these phenomena, uniqueness which refers to the population at the moment of measurement or to the circumstances in which the designed experiment is realized. The need for inference models without a frequency principle is thus urgent in this field.

The concept of total evidence gives a sound ground to prove our Theorem 2. which supplies an inference model for those research situations where the frequentist principle is not in force. This concept is of great use also as a method to clarify and analyze inference models. For example, the often argued misuse of frequentist measurment of inference uncertainty, would have been better indicated if total evidence or other corresponding concepts had been used. See for earlier considerations Hacking (1965), Seidenfeld (1979) for the philosophical part and Hagood (1969), Henkel and Morrison (1969) and Sterling (1959) for the behaviouristic part, and Savage (1954) and Pearson (1962) for the statistician's part.

4 AN APPLICATION OF THE METHOD OF SUPPORT TO THE LINEAR REGRESSION ANALYSIS

The linear regression analysis is one of the most commonly used statistical methods. In the linear regression analysis, statistical inference means an estimation of regression coefficients and testing hypotheses. For performing and the stages of the regression analysis, see Seber (1977). Below, we concentrate only on those stages in which the method of support is needed. Our starting point is the decomposition of the statistical inference situation into the model triplet [M,P,I].

$M \leftrightarrow E(y) = X\Theta$, where y is the vector of observations of the dependent variable ($n \times 1$),
X is the known matrix of the independent variables ($n \times p$).

$P \leftrightarrow y \sim N(X\Theta, I\sigma^2)$, where N refers to the multivariate normal distribution and Θ is the parameter vector ($p \times 1$).

Alternatively, the statistical model above can be represented by means of the ($n \times 1$) vector of errors ε in the following way:

$M \leftrightarrow Y = X\Theta + \varepsilon$, where $E(\varepsilon) = 0$ and $V(\varepsilon) = I\sigma^2$.

$P \leftrightarrow \varepsilon \sim N(0, I\sigma^2)$,

or equivalently, i.e., the density of the distribution of the ε vector is

$$f(\varepsilon_1,\varepsilon_2,\ldots,\varepsilon_n) = \frac{1}{\sigma^n(2\pi)^{n/2}}\, e^{-\varepsilon'\varepsilon/2\sigma^2} .$$

In statistical studies and literature, the presentation of inference models usually ends here. Good planning, however, presupposes the display of the statistical inference model. Thus the model triplet [M,P,I] of the linear regression analysis is the following (the method of support as the statistical inference model):

$M \leftrightarrow Y = X\Theta + \varepsilon$

$P \leftrightarrow \varepsilon \sim N(0, I\sigma^2)$

$I \leftrightarrow S(\Theta)$

The models are defined in more detail.

The Mathematical Model M{ }

The mathematical models of the regression analysis are clearly hierarchic. A typical variance analytic model can be regarded as the basic model and the final model is the one that fits data as closely as possible.

Mathematical model	Explanation of the model
$M^{(1)} \leftrightarrow E(y) = \Theta_o$	The H_o hypothesis of the variance analysis. (No effect on the levels of the explanatory variable).
$M^{(2)} \leftrightarrow E(y) = \Theta$	The H_1 hypothesis of the variance analysis.
$M^{(3)} \leftrightarrow E(y) = X\Theta$	The general linear regression.
$M^{(4)} \leftrightarrow E(g(y)) = q(X)\Theta$	Linear regression between the transformations of the variables Y and X.

The objects of interest in the linear regression analysis are the model $M^{(3)}$ and its generalization $M^{(4)}$, the appropriateness of which as an explanatory or predictive model of the phenomenon studied is examined. The models $M^{(1)}$ and $M^{(2)}$ are used in comparison by means of which it is decided a) whether the linear regression analysis is worth undertaking at all (model $M^{(1)}$) or b) how well the chosen linear model explains the variations in the phenomenon (model $M^{(2)}$).

The Statistical Model $P_\Theta(y)$

The statistical model of the linear regression analysis is connected with the distribution of the dependent variable Y, if independent variables are regarded as constants. It is generally assumed that the observations are independent of each other and so it is realistic to assume in the case of a large sample that the statistics in use follows the multivariate normal distribution. The properties of the multivariate normal distribution are not dealt with here. The linear regression analysis is linked with the determination of the mean vector of the model in question. Thus the mathematical models $M^{(1)}$-$M^{(4)}$ each yield a different statistical model because their mean vectors are different.

The Statistical Inference Model $S(\Theta)$

In the linear regression analysis, the method of support is defined by means of the likelihood of the parameter vector. When the statistical model is a multivariate normal distribution, the likelihood for the parameter vector $\{\Theta,\sigma^2\}$ is

$$L(\Theta,\sigma^2) \propto (\sigma^2)^{-n/2} \exp\{-(1/2\sigma^2)\ (y - X\Theta)'(y - X\Theta)\},$$

and the support function is

$$S(\Theta,\sigma^2) = -\frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\{(y - X\Theta)'(y - X\Theta)\},$$

from which by solving $\partial \log L/\partial\Theta = 0$ and $\partial \log L/\partial\sigma^2 = 0$ we get the evaluates of the parameters, if $X'X$ is nonsingular. These are

$$\hat{\Theta} = (X'X)^{-1}X'y \quad\text{and}\quad \hat{\sigma}^2 = \frac{RSS(\hat{\Theta})}{n}.$$

where the residual sum of squares is $RSS(\hat{\Theta}) = (y - X\hat{\Theta})'(y - X\hat{\Theta})$. In the following discussion the nuisance parameter σ^2 is eliminated by the maximum relative likelihood method (see Edwards (1972), pp.111-15).

It is also possible to write down the support function as the function of the residual sum of squares including the parameter vector under study. Using the notations above we get

$$S(\Theta^{(j)}) = \sup_{\sigma^2} S(\Theta^{(j)},\sigma^2) = -\frac{n}{2}\log\frac{RSS(\Theta^{(j)})}{n} - \frac{n}{2},$$

where the residual sum of squares depends on the chosen statistical model $M^{(j)}$. Because the support function is quadratic for a normal statistical model, it follows that

$$\hat{S}_N(\Theta^{(j)}) = S(\hat{\Theta}^{(j)}) - \frac{1}{2}(\Theta^{(j)} - \hat{\Theta}^{(j)})'B^{(j)}(\Theta^{(j)} - \hat{\Theta}^{(j)})$$

where $B^{(j)}$ is the observed Fisher information matrix, the elements of which on the main diagonal are the terms $\hat{w}_i^2$ indicating the accuracy of the evaluates and the other elements are second order derivates $\partial^2 S/\partial\Theta_i\partial\Theta_k|_{\Theta = \hat{\Theta}}$. The plausibility order of the hypothetical values of the parameter vector can be solved by determining the local uncertainties with the normed support function $\hat{S}_N(\Theta)$. If the m-unit support region is chosen in advance,its boundaries are to be found in the quadaratic support function as the hyperellipsoid defined by

$$2m = (\Theta^{(j)} - \hat{\Theta}^{(j)})'B^{(j)}(\Theta^{(j)} - \hat{\Theta}^{(j)})$$

which defines the m-unit support region.

Figures are a typical form of communication in the method of support. They are succesful mainly in connection with such parameter vectors which have one, two or three components at most. It is because of this that a multiparameter model should either be reparametrisized into a model with only one parameter or one parameter at a time should be considered. In the case of a model with one parameter, the support can be determined by means of the sum of squared residuals

$$\log \frac{RSS(\Theta^{(j)})}{n} = -\frac{n}{2}(a + b\Theta + c\Theta^2)$$

where the coefficients of the parabola are solved by assigning different values to the parameter. The parabola yields the value $\{-b/2c, (nc)^{-1/2}\}$ for the communication parameter $\{\hat{\theta}, \hat{w}\}$. An application of this will be presented in the empirical part of this study, see Chapter 4-3-3.

In the linear regression analysis, the multinomial support $\hat{S}_M(\Theta^{(j)})$ measures two things simultaneously: the fitting of parameter vector in the statistical model and the fitting of the probability model. With this support, it is possible to express the local uncertainty of the statistical model $[M,P]_j$ with regard to the multinomial model best supported by data. When determining the multinomial support, it is worthwhile to divide the data into independent sets. As it is known, in the multinormal regression the conditional distribution $f(y_i|x_{(i)})$ of the dependent variable is a normal distribution with parameters $(x_{(i)}\Theta^{(j)}, \sigma^2)$ where $x_{(i)}$ is the i'th row vector of $\mathbf{X}$. By fitting the multinomial distribution according to the levels 1,...,i,...,k of the independent variables, we get at each level the multinomial supports $\hat{S}_M(\Theta)$ which by direct summing yield the support of the statistical model $[M,P]_j$. Different ways of using the normed support function will be considered in connection with tests in the regression analysis.

The method of support resembles the descriptive methods of the regression analysis, for it is based on the OLS (Ordinary Least Squares) estimation in connection with the multinomial statistical model. In addition to the observed regression, it yields the local uncertainty of the inference results as the multinomial support $\hat{S}_M(\Theta)$ and the plausibility of the hypothetical models in the regression analysis. Chapters 4-1 and 4-2 deal with the estimation of the regression coefficients and their support tests.

4-1 Evaluates of Regression Coefficients

The determination of the numerical values of the unknown parameter vector $\{\Theta^{(j)}\}$ in the regression analysis is evaluation which also includes the determination of the accuracy of the evaluates. Because the evaluates in the multinormal statistical model are the same as the OLS or ML estimates, they are produced by corresponding estimation methods. As the normed support $\hat{S}_N(\Theta)$,$\hat{S}(\Theta)$, etc. has been defined generally for some parameter vector it is possible to regard the hypothetical parameter vector as an evaluate which is compared to the evaluate vector best supported by the data. The latter evaluate is the same as the ML estimate

$$\hat{\Theta}= (X'X)^{-1}X'y.$$

It is easy, for example, to examine the vector estimators yielded by different estimation methods by means of the method of support with the m-unit hyperellipsoid $(\Theta - \hat{\Theta})'B(\Theta - \hat{\Theta})$. The accuracies of the evaluates are obtained by taking square roots of the inverse numbers of the main diagonal elements in the observed Fisher information matrix B (on the condition that they are non-negative), which yields the vector $\hat{\omega}$. Because of the quadratic nature of the support function, one parameter at a time can be separated from the parameter vector to formulate a separate support function

$$S_N(\Theta_j) = S(\hat{\Theta}_j) - (\Theta_j - \hat{\Theta}_j)^2/2\hat{w}_j^2 \text{ ,}$$

and by means of it, the plausibility order of the different values of the point evaluates can be found.

A neighbourhood evaluate for a parameter vector is a hyperellipsoid, but for a single parameter it is the solution of the m-unit support limits $\{\Theta_{j_L}, \Theta_{j_U}\}$ determined from the support function $\hat{S}_N(\Theta_j)$.

In some connections, it is useful to apply the joint evaluate of Definition 14 when determining the evaluates of the regression analysis. The joint evaluate was detemined on the basis of the combinational properties of the data. The regressions of two different data can be combined or the data can be divided into independent sets of data and the regressions of the subsets can be estimated from them. The determination of the support function is in these cases accomplished by arranging the data (combined) into categories 1,...,h,...,k, in each of which there are n_h observations. The support function is applied to each category and so the normed support of the parameter Θ_j in the model {j}is in a one dimensional case

$$\hat{S}_h(\Theta_j) = S_h(\Theta_j) - S_h(\hat{\Theta}_j)$$

The combining of the supports of the subsets of the data yields for the model {j} the total support as follows:

$$\sum_{h=1}^{k} S_h(\Theta^{(j)}) = -(1/2)\sum_{h=1}^{k} n_h\{(a_h + b_h\Theta \quad + \quad c_h\Theta^2 \quad - \quad RSS(\Theta^{(j)})\}$$

The joint evaluate of the total data is the solution of the derivative equation

$$\hat{\Theta} = - \frac{\sum_{h=1}^{k} n_h b_h}{2\sum_{h=1}^{k} n_h c_h} \quad .$$

By using the second derivative we get the accuracy of the joint evaluate

$$\hat{w} = (\sum_{h=1}^{k} n_h c_h)^{-\frac{1}{2}} \quad .$$

The use of the joint evaluate in the regression analysis makes sense, for example, if it is possible to divide the data into subdata on the basis of some dummy variables. It can then be determined directly from the support function how much the support decreases when the evaluates for each data are substituted with the joint evaluate. Its magnitude is

$$S(\hat{\theta}) - S(\hat{\theta}^{(j)}) = \sum_{h=1}^{k} \frac{1}{2\hat{w}_h^2} (\hat{\theta} - \hat{\theta}^{(j)})^2.$$

The fundamental difference between the concepts evaluate and estimate lies in the fact that estimation aims at prediction and for example in the frequentist inference it is clearly interpreted in that way. An evaluate refers to the correction of a more or less vague a priori belief, the amount of which is measured as the support, based on data alone.Because evaluation closely resembles the ML estimation it is often confused with it, as Edwards (1972) points out. Compared with estimation, evaluation is a form of statistical inference because inference uncertainty is included in the notion of the least local uncertainty estimators. In this study, however, we have not rejected Edwards's concepts because the evaluate clearly distinguishes them from the ML and OLS estimates.

4-2 Support Tests of the Regression Analysis

In the linear regression analysis, statistical hypotheses are classified into two groups: on the one hand, the hypotheses of the mathematical model and on the other, those of the probability model. The hypotheses of the mathematical model concern the parameter vector. The tests of the probability model are connected with the homogeneity of variance, the normal distribution of residuals and their mutual independence. The hypotheses are, however, nested and thus a hierarchy of support tests will here be developed for them. In addition, it will be shown that the transformations of the statistics t, F and χ^2 are support tests as well.

Seber (1977) has considered the hierarchic test of the regression analysis as an application of the Neyman-Pearson test theory. His treatment represents the most frequently used testing procedure in the linear regression analysis. Fraser (1979) divides a hierarchic test of the regression analysis into two stages: the first step is to get some assessment how valid the distribution assumption is and the second is to perform the tests of the parameter vector as significance tests. He names this as an adaptive procedure. An example of the information theoretical approaches is Borth's (1975) method of dividing the expected joint entropy (Shannon's entropy)determined from the statistical model into two components: entropy of the probability model and that of the parameter vector. By means of his own information concept, Kullback (1959) deals with the hierarchic test of the parameter vector Θ in which the hypotheses are nested and thus information increases on each step of the hierarchy. In this study, it will be shown how by combining the first step of Fraser's adaptive procedure and Kullback's hierarchic test of the parameter vector we obtain a hierarchic support test of the hypotheses which are nested. As we move from the multinomial model representing the least local uncertainty to the model specified by the regression analysis, inference uncertainty increases on each stage.

We are dealing with a hierarchic support test when the dependent variable y can be defined according to the levels 1,...,l,...,k of the independent variables. The data are then of the form:

$y_{11}, y_{12}, \ldots, y_{1n_1}$ are n_1 repeat observations at the level 1 of the independent variables $(x_{(1)1}, \ldots, x_{(1)p})$

$y_{21}, y_{22}, \ldots, y_{2n_2}$ are n_2 repeat observations at the level 2 of the independent variables

...

$y_{k1}, y_{k2}, \ldots, y_{kn_k}$ are n_k repeat observations at the level k of the independent variables.

According to the multinormal regression analysis we assume that on each level l of the independent variable Y_l follows the normal distribution with the parameter $(\mu_l, \hat{\sigma}_l^2) = (X_{(l)}\Theta, \hat{\sigma}_l^2)$. In addition, it is assumed that the observations of the dependent variable are independent of each other both inside the levels and between them.

The basis of the hierarchic support test is the multinomial statistical model which is fitted on each level of the dependent variable. This yields the local uncertainty $S_M(\hat{\mu}_l)$ which is the least local uncertainty on each level. Because the observations on each level are independent of each other (assumption), we get the total local uncertainty

$$S_M(\Theta) = \sum_{l=1}^{k} S_M(\hat{\mu}_l)$$

which is the sum of the least local uncertainties. Hierarchy in connection with this test implies that at the same time as we start from the least local uncertainty, the total uncertainty measured as support values increases with the specification of the hypotheses H_o, $H_1, \ldots, H_h$. The steps of the hierarchic test of the regression analysis are described below.

Step 1. The least local uncertainty $S_M(\hat{\mu}_l)$ is determined on each level of the dependent variable and summed up as the local uncertainty $S_M(\Theta)$.

Step 2. The normal distribution is discretisized on each level of the dependent variable in such a way that the evaluate vector $(\hat{\mu}_l, \hat{\sigma}_l^2)$ is the parameter vector of the fit and the method of natural class interval proposed by Mineo (1980) is used for the discretization. The supports for each level are then calculated from the standardized support function $\hat{S}_M(\mu_l)$ and, by summing up, these supports yield the total support $\hat{S}_M(\Theta)$. At the first stage of the fitting, the expected cell frequencies of the theoretical normal distribution are applied. If the local uncertainty $S_M(\Theta)$ is subtracted from this the results is the local uncertainty $S_{N_I}(\Theta)$ of the discretisized normal distribution.

Step 3. A fitted distribution is made in which the discretisized normal distribution of Step 2 is transformed in such a way that the cell frequencies for the classes are the observed cell frequencies. This yields the local uncertainty $S_{N_{II}}(\Theta) \leq S_{N_I}(\Theta)$. The total local uncertainty $S_{N_{II}}(\Theta)$ should be used as the basis of the comparison because there the evaluate vector has been formed of the estimates of the least local uncertainty and the hypotheses of the regression analysis have not yet been used.

Step 4. At this step, all the usual hypotheses of the parameter vector Θ of the regression analysis can be tested. These hypotheses concern the lack of fit, linearity and the homogeneity of variance.

The hierarchic support test formally listed

Hypothesis	Explanation	Local uncertainty
H_o: $y_1 \sim$ Multinom (Θ, n_i)	Least local uncertainty	$S_M(\hat{\Theta})$
$H_1^{I}: y_1 \sim N_I(\mu_1, \sigma_1^2)$	Normal fit I	$S_{N_I}(\hat{\Theta}, \hat{w}_1^2)$
$H_1^{II}: y_1 \sim N_{II}(\mu_1, \sigma_1^2)$	Normal fit II	$S_{N_{II}}(\hat{\Theta}, \hat{w}_1^2)$
$H_2: y_1 \sim N_{II}(X_{(1)}\Theta, \sigma_1^2)$	General linear hypothesis (presupposes no homogen.of variance)	$S_{N_{II}}(\hat{\Theta}_{reg}, \hat{w}_1^2)$
$H_3: y_1 \sim N_{II}(\Theta_{(1)}, \sigma^2)$	General linear hypothesis (presupposes homogen. of variance)	$S_{N_{II}}(\hat{\Theta}_{reg}, \hat{w}^2)$
$H_4: y_1 \sim N_{II}(\Theta_o, \sigma_1^2)$	Lack of fit (presupposes no homogen. of variance)	$S_{N_{II}}(\hat{\Theta}_o, \hat{w}_1^2)$
$H_5: y_1 \sim N_{II}(\Theta_o, \sigma^2)$	Lack of fit (presupposes homogeneity of variance)	$S_{N_{II}}(\hat{\Theta}_o, \hat{w}^2)$

The hierarchy of the hypotheses is

$$\begin{tikzcd}
 & H_4 \arrow[dr, phantom, "\subset"] & & & & \\
H_5 \arrow[ur, phantom, "\subset"] \arrow[dr, phantom, "\subset"] & & H_2 \arrow[r, phantom, "\subset"] & H_1^{I} \arrow[r, phantom, "\subset"] & H_1^{II} \arrow[r, phantom, "\subset"] & H_0 \\
 & H_3 \arrow[ur, phantom, "\subset"] & & & &
\end{tikzcd}$$

in which the local inference uncertainty between the hypotheses forms a corresponding order. With respect to some supports it must be mentioned that the difference between the hypotheses H_1^{I} and H_1^{II} measures the loss of information due to classification, as it was already pointed out in Chapter 3-4. An example of this is the consideration of the evaluates of the normal distribution in Chapter 4-3-4. The supports $S\{H_5;H_3\}$ and $S\{H_4:H_2\}$ measure the increase in uncertainty which is due to lack of fit of the linear regression in explaining the phenomenon under study. In all, the hierarchic support test includes at the same time the tests of the normal distribution and those of the homogeneity of variance in the regression analysis. In his interpretation of the support tests, Edwards (1972) does not recommend the use of a special fixed support level. Similarly, Kullback (1959) in his study does not include fixed discrimation rates in his information concept but regards them as statistics in the frequentist inference. In the present study, the support measures directly inference uncertainty as well. The comparison with the hypothesis H_o makes it justifiable to define the test procedure above as a test of the least local uncertainty.

In the linear regression analysis, the statistics χ^2, t and F are usually used as test statistics and they are displayed by the computer programmes in use. Following Edwards (1972) we show next how we can use them as arguments for determing local uncertainties of the appropriate hypothesis. Furthermore the following formulas lead to the normed support function $\hat{S}(\Theta)$. Thus the tests are performed by inserting $\Theta = \Theta_o$ into it, according to Definition 14.

Student's test as a support test

The statistic t is in a normal statistical model connected with the hypothetical value of the mean θ, when the variance σ^2 is unknown. The normed support for θ is then

$$\hat{S}(\theta) = \frac{\nu+1}{2} \log \left(1 + \frac{t_\nu^2}{\nu}\right)$$

where $t_\nu = g(\theta) = \frac{\bar{y}-\theta}{s} \sqrt{\nu}$, ν is the degree of freedom,

$$\bar{y} = \Sigma\, y_i/n \text{ and } s^2 = \frac{1}{n-1} \Sigma\, (y_i - \bar{y})^2$$

Karl Pearson's χ^2 test as a support test

The statistic χ^2 concerns the hypothesis either of location (a) or that of scale (b) in a normal statistical model. It is a common practice to test the goodness of fit by means of the χ^2 statistics. In the latter case the hypotheses of location and scale are often confused with each other, as pointed out by Edwards (1972). Testing the location hypothesis H_o: $\theta = \theta_o$, when the variance σ^2 is known under a normal probability model, leads to the normed support function for Θ, which is

$$\hat{S}(\theta) = -\frac{1}{2} \chi_1^2$$

where $\chi_1^2 = g(\theta) = \frac{n(\bar{y}-\theta)^2}{\sigma^2}$.

Testing the scale hypothesis H_o: $\sigma^2 = \sigma_o^2$, when the mean Θ is known, yields under a normal probability model the normed support function for σ^2,

$$\hat{S}(\sigma^2) = \frac{\nu}{2}[\log \chi_\nu^2 - \log \nu] - \frac{1}{2}[\chi_\nu^2 - \nu] ,$$

where ν is the number of degrees of freedom and $\chi^2_\nu = g(\sigma^2) = \frac{ns^2}{\sigma^2}$, $s^2 = \frac{1}{n-1}\Sigma(y_i - \bar{y})^2$.

An important observation in the hypothesis of scale with regard to interpretation is the fact that in the goodness of fit test the value $\chi^2_\nu = \nu$ of the statistic yields the value zero to the support. Thus the best fit with regard to the local uncertainty is obtained with the value $\chi^2_\nu = E(\chi^2_\nu) = \nu$ and not with $\chi^2_\nu = 0$ which is characteristic of the frequentist inference.

The F test as a support test

The ratio of two sampling variances $F = s^2_1/s^2_2$ is connected with the testing of the poorness of fit, the general linear hypothesis and other corresponding hypotheses. From this test statistics it is possible to deduce two normed support functions. Consider an inference situation where it is to be tested a hypothesis that two populations have the variances σ^2_1 and σ^2_2, whose the ratio is $\xi = \sigma^2_1/\sigma^2_2$. The normed support function for parameter (σ^2_1,σ^2_2) is

$$\hat{S}(\sigma^2_1,\sigma^2_2) = S(\sigma^2_1,\sigma^2_2) - S(\hat{\sigma}^2_1,\hat{\sigma}^2_2)$$

$$= \hat{S}_1(\sigma^2_1) + \hat{S}_2(\sigma^2_2),$$

when the two samples are mutually independent. If the generalized hypothesis concerning parameter ξ, whose evaluate is $\hat{\xi} = \hat{\sigma}^2_1/\hat{\sigma}^2_2$ leads to the normed support function

$$\hat{S}(\xi) = S(\xi) - S(\hat{\xi})$$

$$= S(\xi\ \hat{\sigma}^2_2,\hat{\sigma}^2_2) - S(\hat{\sigma}^2_1,\hat{\sigma}^2_2)$$

$$= \hat{S}_1(\xi\ \hat{\sigma}^2_2) + \hat{S}_2(\hat{\sigma}^2_2)$$

$$= \hat{S}_1(\xi\ \hat{\sigma}^2_2).$$

The support for ξ is easily expressed in terms of $F = \frac{n_1 s_1^2}{n_2 s_2^2} \frac{n_2 - 1}{n_1 - 1}$, $\nu_1 = n_1 - 1$ and $\nu_2 = n_2 - 1$ as

$$\hat{S}(\xi) = -\frac{1}{2}[\nu_1 \log(1 + \frac{\nu_2}{\nu_1} \frac{\xi}{F}) + \nu_2 \log (1 + \frac{\nu_1}{\nu_2} \frac{F}{\xi}) + \nu_1 \log \nu_1 + \nu_2 \log \nu_2 - (\nu_1 + \nu_2) \log(\nu_1 + \nu_2)]$$

It is assumed above that $F \sim F_{(n_1-1),(n_2-1)}$ when $\xi = 1$.

It is characteristic of the support tests to measure inference uncertainty directly. So there is no need to assign any support limits as "critical values". We can directly calculate the support increase available, without any conventional statistical tables. Thus they are appropriate for the statistical tests of instant samples.

4-3 Empirical Application to the IEA Data

The international evaluation of educational achievement is an extensive research project aiming at the comparison of learning achievements between different countries. Its largest data consist of interviews of schoolchildren in twelve different countries, gathered in 1970. The data include observations about approximately 250 000 pupils, 50 000 teachers and 9 700 schools. The data are located at the Institute of International Education at the University of Stockholm and it is available for researchers in the educational field. Copies of the international data have been given to the institutes participating in the IEA project, one of which is the Institute of Educational Research in Jyväskylä. On a national level, the data consist of two-stage cluster samplings of pupil, school and teacher populations.

The choice of the IEA data for the empirical application in this study is due to the fact that the data correspond to the previously defined large instant samples which are regarded as the central area of the application of the method of support. These data cannot be repeated in a sampling situation or in an designed experiment. For example, in Finland at the time of the sampling in 1970 basic education was given within the elementary school system which, by 1980, was in the most part reorganized into a new compulsory school system. Thus, the data form a large instant sample in terms of the number of observations. The choice of the IEA data has also been influenced by the fact that they are easily accessible in the Institute of Educational Research at the University of Jyväskylä. The data have not been utilized as a whole but only a part of them have been used, as specified in the description of the Finnish IEA data bank (see Chapter 4-3-2).

Internationally, the IEA data have been studied quite extensively. This becomes evident, for example, from the comprehensive bibliography of Munck's study (1979). The most important findings are presented in Chapter 4-3-1, where the statistical and inference models used are listed. The substantive model of the regression analysis applied to the data in this study is presented there.

4-3-1 Inference Models used by IEA Researchers

The IEA data have been an object of a massive research activity for almost ten years in different countries. Some of these studies will be discussed here mainly from the point of view of the inference model used and the way in which the inference has, in general, been realized. The studies which will be discussed have intentionally been chosen from those in which the substantive model has been displayed as a linear regression analysis.

In addition, we refer to the work of the hypothesis committee of the IEA research project whereby researchers of the educational field would have had a chance to define an a priori distribution for an eventual Bayesian inference.

The most extensive application of regression and variance analyses so far has been carried out by Peaker (1975), the statistical advisor of the IEA project. Comber and Keeves (1973) have also used regression analysis complemented by path modelling. As for statistical inference,frequentist inference has mainly been used in applying the partial F-test to insert new variables until the regression equation is satisfactory. They have used as a critical value for acceptance $F \geq 2$. The significance of regression coefficients has not been tested and no diagnostic examinations have been carried out either. The significance levels of 5%, 1% and 0.1% have been used. In both studies it has obviously been assumed that the asympotic properties of the used ML estimates are in force because of the large sample size and so a multinormal distribution is admissible as a statistical model for the distribution of the statistics.

A model choice different from the classical regression analysis is represented by Noonan's and Wold's (1977) application of the latent variable technique to IEA data. Their starting point is to describe large statistical data by means of a structural model which is as simple as possible and has as few parameters as possible. Consequently, this application does not contain inductive statistical inference. The computational stage has been realized by the use of NIPALS procedures (Nonlinear Iterative Partial Least Squares). The three variables of the regression analysis are latent variables which are combinations of observed variables. The coefficient of determination $100R^2\%$ has been regarded as one criterion of finding the structure of the following model:

$$Y = \hat{\theta} H + (1-\theta)S$$

$$= (.61)H + (.26)S \; , \quad 100R^2\% = 58.2\%.$$

Interpretationally this regression model is interesting because there it can be seen what fraction $\theta \varepsilon [0,1]$ of that part of learning achievements, which is divided between the home and school variables, can be explained by means of these variables. They obtain as one estimated value for $\theta:(1-\theta) = 2:1$. Since the analysis includes no distribution assumption there is no knowledge of the inference uncertainty linked with this fraction. The application of the method of support carried out here aims at estimating the same parameter θ and, in addition, at determining the accuracy of the evaluate to be yielded as the LLU estimate.

Munck (1979) has applied Jöreskog's LISREL statistical model and LISREL IV computer program in his model building. The choice of a school achievement model is started with a simple and accurately specified model which by means of the LISREL approach is developed to a more general model. She has aimed at a structural model based on the linear regression analysis whose dependent variable and independent variables are latent variables combined from observed variables. No statistical inference model has been recognized although the compatibility of the models is tested by means of the chi square test according to the LISREL IV program. Munck uses the significance levels of 5%, 1% and 0.1% frequentistically. In this case, for example, it would have been useful to aim at a statistical model whose measure of goodness of fit, denoted by χ^2_e would have been in the neighbourhood of its own expectation. (See Chapter 4-2 for Pearson's χ^2 test as a support test).

An application of decision making theory has been performed by Bulcock et al. (1977) who have evolved a statistical decision function for school achievement. In addition to the data and preconceptions a cost function has been needed. He has primarily aimed at producing research findings for the use of school planning authorities. After the computation of results Bulcock draws relatively scanty and questionable conclusions, as for as their utility is concerned.

We have no knowledge of any empirical Bayesian inference, although the IEA project would have offered an interesting opportunity for such a study. Indeed, at the beginning of the IEA project in 1966 and 1967 multidisciplinary conferences were arranged where besides educational researchers, sociologists, political economists and other behavioural and social scientists were present. The aim of the conferences was to deduce structural hypotheses before the collection of the data. A large scientific community was assembled to discuss an a priori research situation in which it would have been possible to look for a priori distributions applicable in Bayesian analysis. The result of the conferences was a number of hypotheses but no a priori distributions.

In a real research situation Bayesian inference is,according to Hogarth (1975), prevented by the fact that the members of the scientific community should know the probability concept before it can be used as a measure of an a priori knowledge. On the other hand, Hogarth's second requirement is suitable for the IEA scientific community. He claims that an a priori distribution should be combined as some kind of a priori probability distribution of a group of researchers as the mean of the a priori distributions of all the researchers. As is known, the IEA community has a large group of researchers and no researchers are working on their own. It seems that scientific communities do not have enough knowledge to utilize Bayesian inference.

The main emphasis of the empirical IEA research carried out so far has been on descriptive methods. No attention has been paid to sampling errors in the data. It has obviously been thought that the numerical magnitude of the data provides the results with some certainty. However, in contradiction to the descriptive strategy, frequentist statistical inference has been used as a criterion in choosing variables (partial F tests) and in the fitting of models (chi square tests), for example.

4-3-2 Description of the Finnish IEA Data Bank

The Institute for Educational Research at the University of Jyväskylä has planned and realized an IEA sampling in Finland. We are dealing with the data bank of what can be called the survey sample of school achievement in six subjects whose collection and filing has been discussed by Saari (1977). The target population of the IEA study was the population of all pupils at the age of compulsory education, which was grouped according to age. From these the subpopulation I has been chosen for this research. At the moment of the IEA measurement in 1970 it consisted of ten-year-old Finnish speaking pupils in normal classes. At the time the total population consisted of 73 369 pupils.

The sampling method used was a two-stage cluster sample. The primary sampling unit was school, which in the whole country numbered 4741 in 1970. A sample of 97 schools was chosen by means of stratified sampling in such a way that the stratification was made on the basis of the location of the schools (administrative district), the degree of urbanization of the municipality in which the schools were situated, the category and size of the schools determined by the number of pupils. At the second stage the pupils in the age group in question were chosen from the clusters or, in this case, from schools in the following way: in small schools up to 40 pupils all of the pupils were chosen and in the other schools with regard to the total number of the pupils proportionally the same number or approximately 13 pupils from each school were chosen. The sampling method at the last stage was systematic according to the month of birth. The sample included the total of 1331 pupils, which form about 1.8% of the target population of 73 369 pupils (sampling ratio 1:55). Here we had to drop out 44 pupils whose sampling records were incomplete because of outlying or missing observations in some variables. Then, the total number of the sampling units was limited to the 1287 pupils. The teacher sample consisted of teachers who taught the pupils in question in the sampled schools at the time of sampling. The 350 teachers included in the sample represent about 7% of the total of 4592 primary and secondary school teachers in the country (sampling ratio for teachers 1:13).

From the description of the data it is clear why it has been chosen for an illustration of the method of support. Firstly, it is a large sample ($n \geq 1000$), in which case the asymptotic properties of statistics can be utilized in inference. Secondly, we are dealing with a two-stage cluster sample with the properties of a probability sample obtained from a finite and real population. Thirdly, it must be noted that the sampling has not been repeated and it cannot be repeated as such because the school system has after 1970 (year of measurement) been changed into the new comprehensive school (1980). Consequently, the degree of uncertainty in inference must be expressed by means of the sample data available and not on the basis of an imaginary repetition principle. In short, the IEA project has a large sample obtained by instant sampling and thus it is most suitable for the use of the method of support.

4-3-3 Empirical Results and their Interpretation

The above discussion on the IEA data and some earlier studies specify the inference situation, which will here be dealt with by referring to Theorem 1 and to Requirements 1 to 3 of the inference situation. The substantive problem is the same as in Noonan's and Wold's (1977) application of the NIPALS modelling. The score Y by any one pupil in a science test can be explained by means of the school type S and the home type H variables. These together and separately explain a certain part 100 R^2% of the variation of the score Y. In addition , we want to know how the explanation is divided between the variables H and S. This is measured by the fraction parameter $\theta \in [0,1]$ which consists of our parameter space Θ. Our aim is to evaluate the least local uncertainty estimate for θ, to determine its accuracy and to perform certain support tests for some hypothetical fraction values. The data $\{y\}$ are an instant sample of IEA and so the observed total evidence is $\{y,\Theta\}$.

A statistical inference situation must first be formulated into the model triplet $[M,P,I]$. As it was already pointed out in Chapter 4-3-2, the Finnish IEA data is a large ($n > 1000$) instant sample. Consequently, the statistical model must be one which uses the pair $\{y,\Theta\}$ as the total evidence. The method of support, fiducial inference and relative likelihood inference are such inference models. Of these the method of support is chosen, but for comparison some results of the other two inference models are also reported. On the other hand, the large size of the instant sample guarantees that for the statistics to be computed we can use a normal distribution as their statistical model. The model triplet of this inference situation is;

$$\boxed{\begin{array}{l} M \leftrightarrow y = \theta H + (1-\theta)\, S + \varepsilon \\ P \leftrightarrow \varepsilon \sim \mathrm{NID}(0, I\sigma^2) \\ I \leftrightarrow \hat{S}_N(\theta) = a + b\theta + c\theta^2 \end{array}} \tag{4.3.3.1}$$

The mathematical model is that of Noonan and Wold (cf. p. 60) and the probability model is derived from the discussions on pages 46-48.

All the variables, indexed below as $x_{u\nu}$, are taken from the Finnish IEA Data Bank, which is located at the Institute for Educational Research at the University of Jyväskylä. We use three blocks of IEA variables; Home, School and Achievement.

I. Learning Achiement (y)

$y = x_{12}$ Science Tests A & B Corrected Score -10,...,60

II. School Block (S)

$s_1 = x_{55}$	School Environment Score	-11,...,11,
$s_2 = x_{57}$	Learning in the School Score	-6,...,6,
$s_3 = x_{59}$	Grade in School	3 or 4,
$s_4 = x_{62}$	Number of Students in Class	1,2,...
$s_5 = x_{78}$	Regular Science Lessons Available	1,2,...
$s_6 = x_{82}$	Observations made in Science	1,2,...
$s_7 = x_{87}$	Number of Grades in the same Classroom	1,...,5
$s_8 = x_{92}$	Need for further Education	1 or 2

III. Home Block (H)

$h_1 = x_{60}$	Father's Occupation	1,...,10
$h_2 = x_{63}$	Homework Hours/Week in all Subjects	hrs/day
$h_3 = x_{77}$	Position in Birth Order	1,2,... .

The relation of the latent variables S and H to the observed variables is assumed to be linear. The first step is to fit the following linear regressions

$$y = \sum_{i=1}^{8} \omega_i s_i + \delta \quad \text{and} \quad y = \sum_{j=1}^{3} \lambda_j h_j + \zeta$$

from which we get the OLS coefficients $\hat{\omega}_i$ and $\hat{\lambda}_j$.

Consequently, the estimated latent variables S and H are;

$$S = \sum_{i=1}^{8} \hat{\omega}_i s_i \quad \text{and} \quad H = \sum_{j=1}^{3} \hat{\lambda}_j h_j \; .$$

The non-linear property linked with the evaluation of the fraction θ can be seen if we examine in more detail the structure of the mathematical model defined by means of latent variables. This procedure is justified because in the present case the observed variables are not multicollinear.

In fact we can write

$$Y = \theta \Sigma \hat{\lambda}_i h_i + (1-\theta) \Sigma \hat{\omega}_j s_j +$$

$$= \theta \hat{\lambda}_1 h_1 + \theta \hat{\lambda}_2 h_2 + \theta \hat{\lambda}_3 h_3 + (1-\theta) \hat{\omega}_1 s_1 + \ldots + (1-\theta) \hat{\omega}_8 s_8 + \varepsilon$$

By using reparametrisation $\theta \hat{\lambda}_i = \hat{\beta}_i$ and $(1-\theta)\hat{\omega}_j = \hat{\alpha}_j$, we obtain the linear regression,

$$Y = \hat{\beta}_1 h_1 + \hat{\beta} h_2 + \hat{\beta}_3 h_3 + \hat{\alpha}_1 s_1 + \ldots + \hat{\alpha}_8 s_8 + \varepsilon.$$

Thus learning achievement has been written as the weighted sum of the home and school variables, in which the weights are non-linear functions of the original parameters. In this kind of latent structure the regression coefficients can be estimated iteratively by means of the NIPALS method by Noonan and Wold (1977).

The method of support leads to another kind of estimation procedure which is based on the quadratic support function of a normal statistical model. In the regression analysis the support function can be determined by means of the residual sum of squares RSS(θ) obtained through the OLS principle. The quadratic support function has thus the following form for the fraction θ,

$$\hat{S}_N(\theta) = (-\frac{n}{2})[a' + b'\theta + c'\theta^2 - \log RSS(\hat{\theta})].$$

The parameter triplet {a',b',c'} of the observed support function is thus determined by varying the parameter θ and calculating the corresponding residual sum of squares RSS (θ). Some of these values are in the table below:

Values for θ	.0	.1	.2	.3	.4	.5	.6	.7	.8	.9	1.0
log RSS(θ)	7.09	7.02	6.97	6.93	6.91	6.91	6.94	6.98	7.03	7.11	7.20

The observed support function for θ is a parabola (see Fig.1),

$$\hat{S}_N(\theta) = (-\frac{1287}{2})[7.09 + (-.81)\theta + (.92)\theta^2 - 6.91].$$

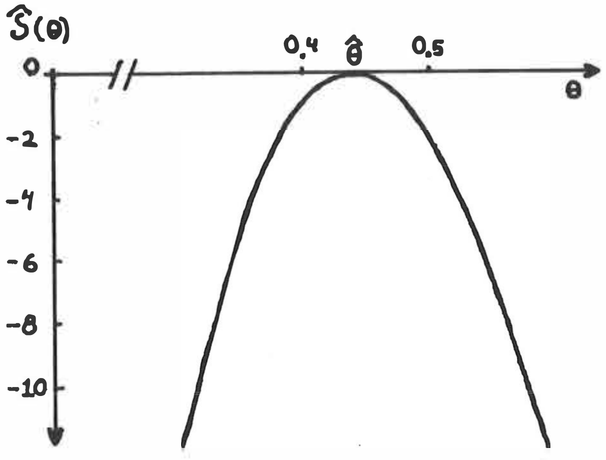

Figure 1. Support curve for Home/School fraction θ ε [0,1]. Finnish IEA six subjects data.

The least local uncertainty (LLU) estimate $\hat{\theta}$ and its accuracy $\hat{w}$ from the observed support function are

$$\hat{\theta} = -\frac{1}{2}(\frac{b'}{c'}) = -\frac{1}{2} \cdot \frac{(-.81)}{(+.92)} = .44 ,$$

$$\hat{w} = (n\ c')^{-\frac{1}{2}} = 0.03 .$$

In order to communicate the computation results m-unit support limits can be determined for the parameter θ as the solution of the equation;

$$|\hat{S}_N(\theta)| = \frac{1}{2\hat{w}^2}(\theta - \hat{\theta})^2 \leq 2,$$

from which it follows that the 2-unit support limits for the fraction θ **are**:

$$.44 \pm .05 \quad \text{or} \quad \hat{\theta} = .44\ (.39,\ .49).$$

By transforming the domain of the fraction θ to be percentual it is possible to state as an evaluate that in the IEA data about 44 ± 5% of the pupils learning achievement is explained by the home latent variable and the rest by the school latent variable from that part which is explained by these variables together.

For the sake of comparison the computional results will here also be given the relative likelihood $\hat{R}_N(\theta) = e^{\hat{S}_N(\theta)}$ and the fiducial distribution $f_f(\theta)$ which in this case is derived from $\theta \sim N(\hat{\theta},\hat{w}^2)$. In the technical appendix (table 1), the values of the support function, relative likelihood and fiducial density functions have been tabulated on some values of the fraction θ. The fiducial distribution for the fraction θ should be defined as a double-sided truncated normal distribution, because the parameter belongs to the interval $\theta \varepsilon [0,1]$ instead of $\theta \varepsilon [-\infty,+\infty]$. Indeed, the truncation is of little value as is seen from the following tail areas,

$$F_f(0) = \int_{-\infty}^{0} f_f(\theta)\, d\theta = F_f(-15.15) \simeq 0 \quad \text{and}$$

$$1 - F_f(1) = \int_{1}^{+\infty} f_f(\theta)\, d\theta = 1 - F_f(+19.34) \simeq 0.$$

and, consequently, no correction is worth doing.

Fiducial distribution can be used in an inference situation in the same way as Bayesian a posteriori distribution. For example, the fiducial probability $P_f\{\theta_L \leq \theta < \theta_U\} \geq .95$ indicates the upper and lower bound of the location of a fiducial mass of a given size. The 95% fiducial condifidence interval for the fraction θ computed of the IEA data is

$$P_f\{.382326 \leq \theta < .495990\} \geq .95 \quad \text{or}$$

$$P_f\{\ \theta \varepsilon [.44 \pm .06]\} \geq .95.$$

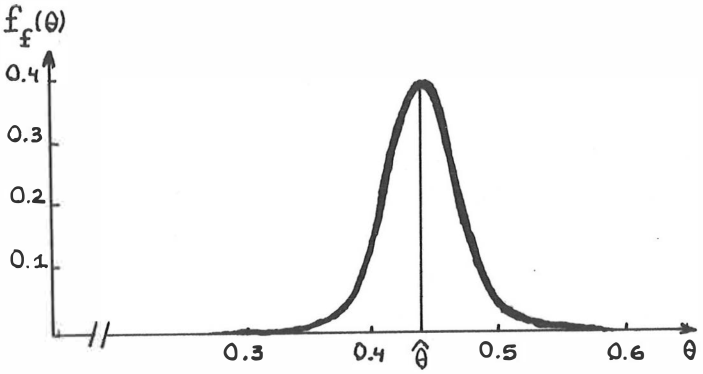

Figure 2. Fiducial density function for Home/School fraction θ ε [0,1]. Finnish IEA six subjects data.

Both in the method of support formula and in the fiducial distribution the value for the evaluate and for the mean is the same θ = 0.44. The accuracy of the support function and the standard deviation w of the fiducial distribution are also the same. Consequently, only the norming of the inference model used in inference is different. The support function denotes to what extent the local uncertainty increases (or as in Edward's terminology support decreases) when the parameter value moves away from the evaluate $\hat{\theta}$. In a fiducial formula the parameter values are not compared with each other, but an area of the parameter space $\theta \varepsilon [0,1]$, whose P_f probability exceeds a given bound by getting up to 95%, for example, is indicated. In the choice of the area we can, however, centralize it about the evaluate $\hat{\theta}$.

The relative likelihood $\hat{R}_N(\theta) = e^{\hat{S}_N(\theta)}$ is a transformation of the formula for the support function. A noticeable advantage is the use of a likelihood unit which can, for example, be denoted percentually 100 $\hat{R}_N(\theta)$%. The relative likelihood $\hat{R}_N(\theta)$ must not be mistaken for probability. Consequently, by means of this inference model the domain of the parameter θ which has a relative likelihood higher than some a priori chosen relative likelihood $\alpha \varepsilon [0,1]$ can be stated. For instance, a 2.5% likelihood interval is $\theta \varepsilon [.36, .51]$ if the function 100 $\hat{R}_N(\theta)\% > 2.5\%$.

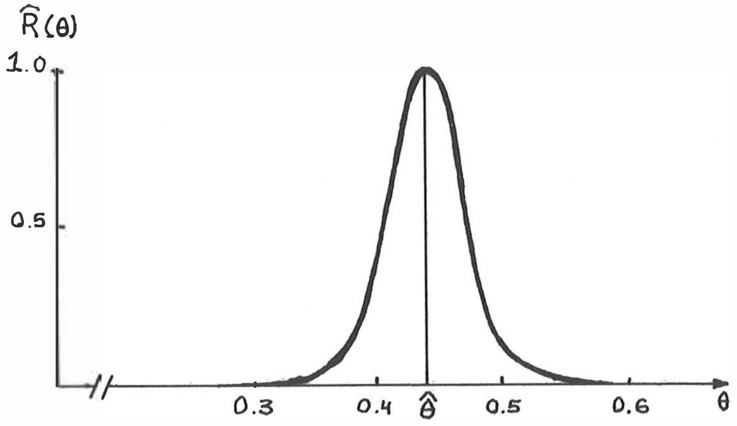

Figure 3. Relative likelihood function of Home/School fraction $\theta \varepsilon [0,1]$. Finnish IEA six subjects data.

The one to one transformation between relative likelihood andthe support function indicates that we are dealing with the use of a different measuring scale.

Although the interpretation of the percentual relative likelihood 100 $R_N(\hat{\theta})\%$ is easy for the user of the inference results an incorrect inference may occur if it is confused with probability concepts. Only in fiducial inference there is a probability concept in the background. It is assumed that there is at the statistician's disposal a certain confidence mass, which can be represented as a fiducial distribution, in the parameter space. Since there is no frequency interpretation for this probability concept the fiducial inference results are closely tied to the inference situation which clearly refers to the concept of an instant sample.

In order to compare the inference models $\hat{S}_N(\theta)$, $\hat{R}_N(\theta)$ and $F_f(\theta)$ some intervals of the fraction θ have been listed in the table below.

Inference model and arbitrary chosen value for inference uncertainty	Range for θ (Finnish IEA Data)
1. Support function $\hat{S}_N(\theta)$ The 2-unit support limits	$.379 \leq \theta < .497$
2. Relative likelihood 100 $\hat{R}_N(\theta)$ The 2.5% likelihood interval	$.360 \leq \theta < .509$
3. Fiducial probability $F_f(\theta)$ 95% fiducial confidence interval	$.380 \leq \theta < .498$

Because the measure for inference uncertainty varies according to the inference model, they are not directly comparable to each other, if the width of range is considered. The width of a confidence interval depends on the value of the chosen criterion alone. On the other hand, with the value $\hat{\theta} = .44$ all inference functions obtain their maximum values. For further ranges see Table 2 in Appendices.

In order to supplement the statistical inference some hypotheses linked with the home/school fraction θ can be examined. In advance the following ones would seem to be of interest:

Null and alternative hypothesis		Substantive interpretation
$H_o^{(1)}$; $\theta = 0$	$H_1^{(1)}$; $\theta > 0$	Home has no effect
$H_o^{(2)}$; $\theta = 1$	$H_1^{(2)}$; $\theta < 1$	School has no effect
$H_o^{(3)}$; $\theta = .5$	$H_1^{(3)}$; $\theta =.5$	Effect is divided evenly between home and school
$H_o^{(4)}$; $\theta = .3$	$H_1^{(4)}$; $\theta =.3$	Effect is divided between home and school in odds 1:2.

The test statistic of a support test is, according to Definition 15, an insertion into a standardized support function, which in this case is $\hat{S}_N(\theta)$. The test result directly yields a support decrease with regard to the hypothesis best supported by the data, which here has the value $\theta = \hat{\theta} = .44$.

The support test yields the following result for each hypothesis:

Null hypothesis and the value of the test statistic	
$H_o^{(1)}$; $\hat{S}_N(0) \approx -\infty$	No support obtained from the data
$H_o^{(2)}$; $\hat{S}_N(1) \approx -\infty$	No support obtained from the data

For both hypotheses it can be inferred that neither the school nor the home variable alone explains learning achievement if both of these variables are considered together.

$H_o^{(3)}$; $\hat{S}_N(\frac{1}{2}) = -2.20$	The data support this kind of hypothesis to some extent. The hypothetical $\theta = .5$ is at the distance of nearly twice the accuracy $\hat{w}$ from the hypothesis best supported by the data.
$H_o^{(4)}$; $\hat{S}_N(\frac{1}{3}) = -11.52$	Little support obtained from the data

The result of the support test indicates how quickly local uncertainty increases when we move away from the neighbourhood of the evaluate. On the other hand, a support test is not very illustrative because the result is one number which denotes the change of local uncertainty in Rényi's terminology. Instead of support tests, in the case of one parameter, the observed support function and its graph should be displayed. This method has earlier been used by the already mentioned Edwards , Cole and Dickey; Dickey, however, has not used it in the sense of the method of support.

Although the aim of presenting the numerical results has been to show the ability of the method of support to function as an inference model of an instant sample, it is worth emphasizing the substantive results connected with the estimation of the fraction θ. Firstly, the regression analysis which was carried out explained 62.3% of the total variation of the score Y. This figure can be regarded as rather high. Secondly the evaluate of the home/school fraction was $\hat{\theta} = .44$, which yields the odds 1:1.3 when we are considering how the foregoing degree of explanation is divided between these two variables. The odds is in favour of the school variable and, consequently, is in accordance with expectations as Noonan and Wold (1977) have pointed out in connection with their NIPALS modelling. Their odds is 2:1 according to the empirical computations, but in their opinion the method overestimates the influence of home. Peaker (1975), who has the odds 6:1, has the greatest deviation in favour of home. This result is, however, due to the fact that in the regression analysis in question the home block variables were entered first as independent variables and thus explain a larger portion due to the interaction with the school block variables.

As for the measuring techniques the IEA data are soft and of behavioural nature to which, with regard to the distribution assumption,hard statistical methods have here been applied. For the reliability of the results it should thus be confirmed to what extent such conditions hold. Since there is an instant sample as data the use of the method of support as a statistical inference model cannot be questioned. It still remains open in what way the statistical distribution assumption of a normal distribution and the other assumptions of the regression analysis hold. They will be examined in the next Chapter 4-3-4.

4-3-4 Adequacy of the Underlying Assumptions

The assumptions in the regression analysis usually concern the residual vector and the examination of its properties of this serves as a diagnostic check of the inference situation in question. It forms an inference situation of its own in which, in the case of an instant sample, the method of support should be applied. Inference is here an application of the support tests.

The regression residuals are assumed to have the following properties: they a) are mutually uncorrelated, b) have homogeneous variance, c) have a normal distribution and d) are unbiased. From the properties a) and b) it follows that $D^2(\varepsilon) = \sigma^2 I$ and the property d) means that $E\{\varepsilon\} = 0$. In addition, it follows from the properties a) and c) that all the residuals are mutually independent. Each of these properties can be studied as a statistical hypothesis and appropriate statistical tests can be performed. We consider here only the normality assumption which is examined by means of the test for goodness of fit and of the support determined from that test.

In this study, the value of the home/school fraction θ must always be fixed before the regression residuals can be calculated. A natural value for θ is the evaluate $\hat{\theta} = .44$. The corresponding 1287 regression residuals are presented in Table 2 of Appendices. The evaluate for the parameter $\{\mu,\sigma^2\}$ of the residuals is $\{0, 3.58^2\}$ which will be applied to make the normal distribution fit. Because the observed frequency distribution is always discrete, we must, for the fitted distribution, discretisize the continuous random variable which in this case is the normal distribution. It was pointed out earlier that discretization decreases local uncertainty which, measured as support, may create discrepancy between the observed distribution and its fit. Thus, it is important to classify the variable in such a way that it follows the form of the theoretical distribution as closely as possible.

In his treatment of the classification of data, Lindsey (1974 a) has emphasized the number of classes. However, Mineo (1979) has demonstrated that, in the discretization of a continuous random variable, the number of classes is not as essential as their width. Classification in general has been treated in statistical literature and, for example, one method presupposes such widths of classes in which there is in each of the classes the same percentage of observations. One of these methods, put forward by Mineo (1979), brings about the most accurate goodness of fit in connection with the χ^2 test. This kind of classification is called the method of natural class interval. On the other hand, Lindsey has used the support $\hat{S}_M(\Theta)$ in the manner of the multinomial model but has got nowhere because when the number of classes decreases, the goodness of fit increases and naturally becomes perfect when there is exactly one class.

In our study, we have combined Lindsey's support and Mineo's natural classification consisting of classes of different lengths. Because classification is of great significance here, it is considered in more detail below.

The basis of the natural classification is the ordered n-tuple of observations $y_1,\ldots,y_i,\ldots,y_n$, where ties are also possible. We are aiming at a classification in which variance inside the classes is minimized compared to variance between the classes. The criterion for the minimization is either the squared deviation (D) calculated from neighbouring observations or the variance (V) obtained from the formulae:

$$D_j = \frac{n_j n_{j+1}}{n_j + n_{j+1}} (y_{j+1} - y_j)^2 \text{ or } V_j = \frac{n_j n_{j+1}}{(n_j + n_{j+1})^2} (y_{j+1} - y_j)^2,$$

where j = 1,2,...,m-1 and thus forms in all m-1 statistics calculated of neighbouring observations. The smallest of these is chosen and the corresponding classes are combined by choosing as the new statistic of the class the mean

$$\bar{y}_s = \frac{y_j n_j + y_{j+1} n_{j+1}}{n_j + n_{j+1}},$$ the corresponding frequency of which is $n_s = n_j + n_{j+1}$.

After this we obtain a new empirical distribution with m-1 classes. The combining procedure is repeated so many times that a distribution with the required number of classes is reached. It is reasonable to fix the final number of classes a priori. In our study, for example, we have chosen k = 20 as the number of classes. The frequency distribution of the regression residuals is presented in Figure 4. It is based on a statistical model (see formula 4.3.3.1) the evaluate of which gets the value $\hat{\theta} = .44$. The class widths have been determined by Mineo's method.

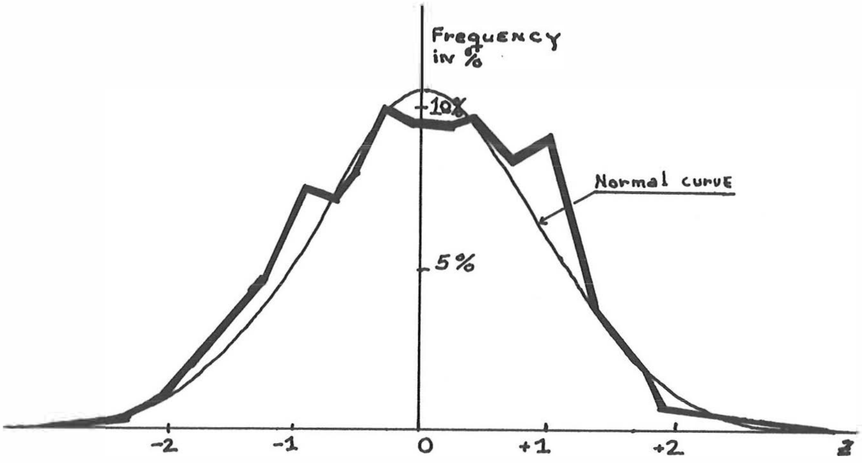

Figure 4. Histogram of the distribution of 1287 regression residuals and a normal distribution curve (class lengths are unequal) Finnish IEA six subjects data.

In examining the distribution of the regression residuals, three fits are made which form a plausibility order with regard to the goodness of fit. They are the multinomial fit, the normal fit I, in which the cell frequencies of the classes are identical with the observed cell frequencies, and the normal fit II, in which the frequencies of the classes are the fitted frequencies. We use the two fitted normal distributions I and II to cancel out the additional uncertainty caused by the discretization of the continuous normal distribution. Here we refer to our discussions concerning the loss of uncertainty in Chapter 3-4. The multinomial model represents the least local uncertainty, as demonstrated in Chapter 3-4, and it is obtained from the following formula:

$$S_M(\hat{\theta}) = \sum_{i=1}^{20} n_i \log \hat{\theta}_i = 3 \log .0023 + 6 \log .0047 + \ldots + 6 \log .0047$$

$$= -2415.09 \ , \quad \text{where } \hat{\theta}_i = n_i/n.$$

Two normal fits are deliberately made, the first of which, the fit I, represents a model in which classification does not increase local uncertainty. The local uncertainty of the normal fit I can be calculated by using the mean of the residual distribution $\bar{\varepsilon} = 0$ as the mean of the fit and $D(\varepsilon) = 3.58344$ as standard deviation. The bounds of the classes are identical with those of the residual distribution and their frequencies are identical with the frequencies n_i of the residual distribution. The local uncertainty for the normal fit I is thus

$$S_{N_I}(\theta) = \sum_{i=1}^{20} n_i \log \theta_i = 3 \log .0027 + 6 \log .0051 + \ldots + 6 \log .0059$$

$$= -3440.32 \ , \quad \text{where } \theta_i = \Phi(y_{U_i}) - \Phi(y_{L_i}),$$ where U and L denote the lower and upper bounds of the natural classes determined from the normal distribution.

The normal fit II is an accurate fit when a continuous random variable is approximated by a discrete one. There the probabilities for each class are the same as in the fit I and the frequencies for each class are exactly as expected by the fit. Thus they differ from those of observed frequencies n_i. Local uncertainty based on these facts represents local uncertainty determined without observations, for this fit is assigned to the data on the basis of the bounds of classes, standard deviation and the total number of observations. (For details see Appendices, Table 2). Thus, the local uncertainty of the normal fit II is

$$S_{N_{II}}(\tilde{\theta}) = \sum_{i=1}^{20} \tilde{n}_i \log \tilde{\theta}_i = 3 \log .0027 + 7 \log .0051 + \ldots + 6 \log .0059$$

$$= -3439.69 \text{ when } \tilde{\theta}_i = \theta_i \text{ and } \tilde{n}_i = n\,\tilde{\theta}_i .$$

The comparison of the three distribution fits yields a plausibility order, measured with supports, in which the support of the two normal fits is relatively small

$$\hat{S}\{N_I;\ N_{II}\} = -3440.32 + 3439.69 = -0.63.$$

According to this, the distribution of observations follows the same distribution as a discrete normal distribution with observations of the same order of magnitude. This leads us to the conclusion that the residuals follow a normal distribution. If we calculate the corresponding statistic of the χ^2 test, it gets the value

$$\chi^2_{20-3} = 7.61 \in [7.56,\ 30.19] = [\chi^2_{.025},\ \chi^2_{.975}].$$

With this measure, a relatively high goodness of fit is achieved. As it was pointed out in Chapter 4-2, too high a goodness of fit does not, however, guarantee that there could not exist some hypothesis, or in this case a distribution, whose local uncertainty were even smaller. As a matter of fact, such a distribution is the multinomial fit, the local uncertainty of which is $S_M\{\hat{\theta}\} = -2415.09$. Measured with support, both the normal fit I and II are very far from this multinomial distribution which represent the least local uncertainty model. This casts some doubts upon our mathematical and statistical models. There are to be found other models with a better fit.

A high goodness of fit between the discrete normal distribution and the observations was achieved with the choice of a natural class interval. At the same time, we aimed at decreasing one variance. Thus it is possible to calculate directly from the statistic χ^2 the support

$$S_{\chi^2}\{N_I; N_{II}\} = -\frac{17}{2}(\log 17 - \log 7.61) - \frac{1}{2}(17 - 7.61)$$
$$= -2.13 .$$

This implies that we can search for a statistical model with a smaller variance as the one used and get an increased support.

The diagnostic examination of the regression residuals as a support test results in the conclusion that the residuals may follow a normal distribution. On the other hand, there exists an other statistical model which is considerably better supported by the IEA data and that is the multinomial model. Thus, the discrete normal distribution used as a reference distribution is locally not the best distribution fit of the phenomenon to be analyzed.

5 CONCLUSIONS

Following remarks concern some restrictions on the use of the method of support and the diversification on inference uncertainty and uncertainty of the phenomen on to be studied. In addition, we shall give indications for further research activity in the LLU -estimation and -test theory.

First we shall point out that a well-performed inference situation must be decomposed as the model triplet [M,P,I]. In addition the content of total evidence at the statistician's disposal, which conveys the choice of the inference model, must be checked. If the method of support is used, the first thing to be checked is that the joint probability of sample exists and that the likelihood function is dependent on population parameters of interest. For this reason we notice that when the distributional properties of a parent population are unknown we can use the large sample properties of ML estimators which are applied in our inference situation. As a result we are generally dealing with large instant samples, as mentioned earlier. This restricts the use of the method of support to large samples. Naturally, there are really no obstacles for an analysis of small samples, too, if the joint probability of the sample can be properly evaluated.

It cannot be denied that communication results achieved by the method of support are relatively modest compared to those of richer inference models. But it must also be noticed that, in the case of an instant sample, the observed total evidence is so scarce that inference uncertainty can only be measured as differences in local uncertainties, unless imaginary components are deliberately added to the required total evidence. Thus, we can recommend the method of support primarily for an inference model of statistical instant samples.

Secondly, we shall direct attention to the two distinct uncertainty concepts which appeared in our empirical application. With regard to the diagnostic considerations we must pay attention to the observation concerning the commonly used fit of regression residuals based on a normal distribution. In a case like that, it is possible to determine by means of the method of support the least local uncertainty of a reference distribution, here the normal distribution, by classifying observations to ensure the best possible fit. In this way, the local uncertainty of the reference distribution can be determined and, by inserting the classified observations obtained from the data, the local uncertainty of this, in a traditional sense, distribution fit can be calculated. The support between these two distributions measures, in an ordinary sense, the fitting of observations with the normal distribution. Here, for example, the goodness of fit was high enough when a suitable classification was applied. But the most important observation is that on uncertainty of the phenomen on under study and uncertainty caused by sampling fluctuation or inference uncertainty are separated from each other. The used normal distribution as a reference distribution represents substantive fluctuation and, as we saw, it gets very little support in comparison with the multinomial distribution best supported by data. So we recommended the search of some other substantive probability model with to smaller variance. Accordingly, total uncertainty prevailing in an inference situation is decomposed in to two parts: one of substantial type and the other of inferential type. Analysing instant samples thus presupposes two uncertainty concepts, both of which are Rényi's local uncertainties, one for inference uncertainty and one for the entropy of the phenomen on itself.

In the present study, the method of support was applied to an instant sample gathered from Finnish schools in 1970. It is not possible to repeat this kind of sampling, even in a historical sense. A similar kind of situation is presumably often encountered in other behavioural scientific research, too.

We evaluated some LLU-estimates concerning Home/School fraction θ and performed LLU-tests for a few hypothetical values for θ. Compared to the results of some early studies in the same field, the LLU-results were in good agreement with them. For further research we recommend the evaluation of the general statistical properties of the LLU-tests and - estimates. It can easily be seen that some locality properties are linked with them and an application to the superpopulation sampling theory is possible.

FINNISH SUMMARY

Tieteellisen tutkimustyön tilastollisessa osassa kohdataan monesti tilanne, jossa havaintoaineisto on todennäköisyysotos tai havaittu koeasetelma, jota ei ole toistettu eikä edes aiottu toistaa. Tällaista havaintoaineistoa nimitetään tässä kertaotokseksi, jolle etsitään tilastollinen päättelymalli, koska päättelyepävarmuuden mittaamisen osalta toistoperiaatteeseen nojautuvat päättelymallit eivät siihen sovellu.

Lauseessa 2 on osoitettu miten Edwardsin (1972) hioma tukifunktiotekniikka sopii kertaotoksen tilastolliseksi päättelymalliksi. Todistus nojautuu siihen miten päättelymallin edellyttämän ja päättelijän käytössä olevan kokonaistiedoston tulee vastata toinen toisiaan, jotta päättelytuloksiin otosaineiston mukanaan tuoma päättelyepävarmuus on mitattavissa. Kokonaistiedosto koostuu esimerkiksi havaintoaineistosta (ja sen toistosta), otos- ja parametriavaruudesta. Kertaotoksen tapauksessa kokonaistiedostona on havaintoaineisto ja parametriavaruus.

Tukifunktiotekniikan keskeinen päättelykäsite on logaritminen uskottavuusosamäärä, jolle tämän tutkimuksen lauseessa 1 osoitetaan informaatioteoreettinen tulkinta. Todistus perustuu Rényin (1970) epätäydellisen todennäköisyysjakauman ja sille määritellyn epävarmuuskäsitteen varaan. Siten logaritminen uskottavuusosamäärä on kahden paikallisen Rényi-epävarmuuden erotus. Kertaotoksen tapauksessa on kysymyksessä yksi havaittu tapahtuma, joka tässä mielessä riittää päättelyepävarmuuden mittaamiseen.

Tukifunktiotekniikan sovellus tilastolliseen estimointi- ja testiteoriaan johtaa pienimmän paikallisen epävarmuuden (engl. lyh. LLU) estimaattoreihin ja testeihin. Näiden toimivuus teoreettisella tasolla on osoitettu viivallisen regressioanalyysin estimoinnissa ja hypoteesien testauksessa. Empiirinen laskenta liittyy kansainvälisen koulusaavutustutkimuksen (IEA) Suomen havaintoaineistoon, joka on laaja kertaotos, ja josta on määrätty eräitä LLU-estimaatteja ja -testejä.

APPENDICES

Table 1. Three inference uncertainty measures for estimation of Home/School fraction θ ε [0,1]. Finnish IEA Six-Subject data*.

Selected values for fraction θ	Measure for inference uncertainty		
	Support function $\hat{S}_N(\theta)$	Relative likelihood function $\hat{R}_N(\theta)$	Fiducial cumulative distribution function $F_f(\theta)$
.000	- ∞	.000	.000
...	...	...	...
.330	-6.907	.001	.000
.340	-5.846	.003	.000
.350	-4.727	.010	.001
.360	-3.726	.024	.005
.380	-2.081	.125	.025
.381	-2.000	.135	.228
...	...	...	...
.439	± .000	1.000	.500
...	...	...	...
.497	-2.000	.135	.772
.498	-2.080	.125	.975
.510	-3.887	.020	.995
.520	-4.605	.010	.998
.530	-4.908	.007	.999
.540	-6.907	.001	1.000
...	...	...	...
1.000	- ∞	.000	1.000

$$\hat{S}_N(\theta) = \frac{n}{2}(a + b\theta + c\theta^2 - \log RSS(\hat{\theta})$$

$$= -\frac{1287}{2}(7.09 - .81\theta + .92\theta^2 - 6.91)$$

$$\hat{R}_N(\theta) = \exp \hat{S}_N(\theta)$$

$$F_f(\theta) = \Phi\left(\frac{\theta - \hat{\theta}}{\hat{w}}\right) = \Phi\left(\frac{\theta - .44}{.03}\right)$$

* Finnish IEA Six-Subject Survey data is in posession of the Institute for Educational Research, University of Jyväskylä. Saari (1977) has reported how the data have been documented for research use.

Table 2. Observed frequency distribution of regression residuals and its normal fit (e). Selected value for Home/School fraction is the evaluate $\hat{\theta} = 0.44$. Finnish IEA data (see footnote table I). Class interval lengths used are unequal.

Number of class	Class upper point	Frequencies			
		Observed		Expected (e)	
		Pupils	%	Pupils	%
1	<-2.65	3	0.23	3	0.27
2	-2.42	6	0.47	7	0.51
3	-2.09	6	0.47	14	1.05
4	-1.75	26	2.02	28	2.18
5	-1.42	48	3.73	49	3.77
6	-1.17	62	4.82	56	4.32
7	-0.86	97	7.54	95	7.39
8	-0.64	91	7.07	85	6.62
9	-0.41	103	8.00	103	7.98
10	-0.16	128	9.95	123	9.55
11	+0.09	123	9.56	128	9.95
12	+0.33	121	9.40	120	9.34
13	+0.62	125	9.71	133	10.31
14	+0.90	110	8.55	108	8.35
15	+1.34	118	9.17	121	9.40
16	+1.68	57	4.43	56	4.36
17	+1.96	34	2.64	30	2.32
18	+2.24	15	1.17	14	1.08
19	+2.52	8	0.62	8	0.66
20	∞	6	0.47	6	0.59
Totals		1287	100.00	1287	100.00

REFERENCES

Ash, R.B.(1965). *Information Theory*. New York: John Wiley and Sons.

Bartlett, M.S.(1971). When is inference statistical inference? *Foundations of Statistical Inference*, 20-31, eds. Godambe, V.P. and Sprott, D.A. Toronto: Holt, Rinehart and Winston.

Birnbaum, A.(1969). Concepts of statistical evidence. In *Philosophy, Science and Method*, 112-43, eds. Morgenbesser, S., Suppes, P. and White, M. New York: St Martin's Press.

Borth, D.M.(1975). A total entropy criterion for the dual problem of model distribution and parameter estimation. *J.R. Statist. Soc.*, *B*, *37*, 77-87.

Bulcock, J.W., Fägerlind, I. and Emanuelsson, I.(1974). Education and the socioeconomic career II: A model of the resource conversion properties of family, school, and occupational environments. Research Report 10. University of Stockholm: Institute for International Education.

Cole, T.J.(1975). Linear and proportional regression models in the prediction of ventilatory function. *J.R. Statist. Soc.*, *A*, *138*, 297-325.

Comber, L.C. and Keeves, J.P.(1973). *Science Education in Nineteen Countries*. Uppsala: Almqvist and Wiksell.

Dickey, J.(1973). Scientific reporting and personal probabilities: Students hypotheses. *Ann. Math. Statist.*, *44*. 285-305.

Dickey, J. and Freeman, P.(1975). Population-distributed personal probabilities. *J. Amer. Statist. Ass.*, *70*, 362-364.

Edwards, A.W.F.(1972). *Likelihood*. Cambridge:University Press.

Fraser, D.A.S.(1979). *Inference and Linear Models*.London: McGraw-Hill Inc.

Good, I.J. and Osteyee, D.P.(1974). *Information, Weight of Evidence, the Singularity between Probability Measures and Signal Detection.* Lecture Notes in Mathematics 376. Berlin: Springer-Verlag.

Good, I.J.(1950). *Probability and the Weighing of Evidence.* London: Charles Griffin.

Hacking, I.(1965). *Logic of Statistical Inference.* Cambridge: University Press.

Hagood, M.J.(1969). *Statistics for Sociologistis.* New York: Henry Holt Inc.

Henkel, E.R. and Morrison, D.E.(1969). Significance tests reconsidered. *The Amer. Sociologist, 4.* 131-140.

Hogarth, M.R.(1975). Cognitive processes and the assessment of subjective probability distributions. *J. Amer. Statist. Ass., 70,* 271-289.

Husén, T., ed.(1967).*International Study of Achievement in Mathematics: A. Comparison of Twelve Countries.* Stockholm: Almqvist and Wiksell.

Kalbfleisch, J.D. and Sprott, D.A.(1970). Applications of likelihood methods to models involving large numbers of parameters. *J.R. Statist. Soc., B, 32,* 175-208.

Kullback, S.(1959). *Information Theory and Statistics.* New York: Wiley and Sons.

Lindley, D.V.(1956). On a measure of the information provided by an experiment. *Ann. Math. Statist., 27,* 986-1005.

Lindsey, J.K.(1974a). Comparison of probability distributions. *J.R. Statist. Soc., B,* 36, 38-47.

Lindsey, J.K.(1974b). Construction and comparison of statistical models. *J.R. Statist. Soc., B, 36,* 418-424.

Menges,G.(1973). Inference and decision. In *Selecta Statistica Canadiana, vol. I,* 1-16. Delhi: Hindustan Publishing Corporation.

Mineo, A.(1979). A new grouping method for the right evaluation of the chi-square test of goodness-of-fit. *Scand. J. Statist. 6*, 145-153.

Munck. I.M.E.(1979). *Model Building in Comparative Education*. Stockholm: Almqvist and Wiksell International.

Noonan, R. and Wold, H.(1977). NIPALS path modelling with latent variables. *Scand. J. Educ. Research, 21*, 33-61.

Peaker, G.F.(1975). *International Studies in Evaluation VIII: An Empirical Study of Education in Twenty-Two Countries: A Technical Report*. Stockholm: Almqvist and Wiksell.

Pearson, E.S.(1962). Some thoughts on statistical inference. *Ann. Math. Statist. 33*, 394-403.

Pitman, E.J.G.(1979). *Some Basic Theory for Statistical Inference*. Cambridge: University Printing House.

Preuss, L.G.(1980). A class of statistics based on the information concept. *Commun. Statist. - Theor. Meth., A, 9*, 1563-1585.

Rényi, A.(1970). *Probability Theory*. Amsterdam: North-Holland.

Saari, H.(1977). Finnish data files of the IEA six subject study. Description of the documentation of data. Bulletin 81. University of Jyväskylä: Institute for Educational Research.

Savage, L.J.(1954). *The Foundations of Statistics*. New York: John Wiley and Sons.

Seber, G.A.F.(1977). *Linear Regression Analysis*. New York: John Wiley and Sons.

Seidenfeld, T.(1979). *Philosophical Problems of Statistical Inference*. Dordrecht: D. Reidel Publishing Company.

Sprott, D.A. and Kalbfleisch, J.D.(1969). Examples of likelihoods and comparisons with point estimates and large sample approximations. *J. Amer. Statist. Ass., 64*, 468-484.

Sterling, T.D.(1959). Publication decisions and their possible effects on inferences drawn from tests of significance - or vice versa. *J. Amer. Statist. Ass.*, *54*, 30-34.

Suppes, P.(1966). Probabilistic inference and the concept of total evidence. In *Aspects of Inductive Logic*, *28-42*, eds. Hintikka, J. and Suppes, P. Amsterdam: North-Holland.

Theil, H.(1967). *Economics and Information Theory.* New York: American Elsevier Publishing Company.

Wilkinson, G.N.(1977). On resolving the controversy in statistical inference. *J.R. Statist. Soc.*, *B*, *39*, 119-172.