Cowles Commission Discussion Papers: Statistics, November 25, 1947

THE IDENTIFICATION OF STRUCTURAL CHARACTERISTICS

by T. C. Koopmans

General Formulation of the Identification Problem

R. A. Fisher [F] distinguished the following problems in statistical inference (re-numbered by me):

1.) Specification of the mathematical form of the population from which the data are regarded as a sample.

3.) Estimation, i.e. choice of statistics to serve as estimates of parameters of the population.

5.) Distribution of estimates and other statistics.

Of course, 3) and 5) are closely connected problems, which can only be solved in combination. Subsequently, Neyman, Pearson and Wald added to the list:

4.) Testing hypotheses or deciding between courses of action,

a group of problems treated in combination with the corresponding problems under 3).

It is the purpose of this discussion to suggest a re-formulation of the specification problem, appropriate to many applications of statistical methods, and to point out the consequent emergence of an additional group of problems.

In many fields the objective of the investigator's inquisitiveness is not just the distribution of observable variables, but a physical structure projected behind this distribution, by which the latter is thought to be generated. The word "physical" is used merely to convey that the structure concept is based on the investigator's ideas as to the "explanation" or "formation" of the phenomena studied, briefly, on his theory of these

phenomena, whether they are classified as physical in the literal sense, biological, psychological, sociological, economic or otherwise.

Specific examples of such postulated structures are contained further on in this note (example from econometrics), and in the contribution of Rasch (biometrics) and Reiersøl (psychometrics) to this discussion. Some generalizations from these examples are attempted here.

In each of the cases considered, the distributional specification applies directly to certain non-observable variables, variously referred to as disturbances, specific factors, etc. We shall refer to these as latent variables, denoted by a vector u. In addition, certain structural relationships are specified which connect the latent variables with observable variables, here to be called apparent variables and denoted by a vector y. The specification is therefore concerned with the mathematical forms of both the distribution of the latent variables and the relationships connecting apparent and latent variables.

The term "mathematical form" carries a suggestion of parametric specification which obviously is not the only possible type. We shall therefore employ terms and concepts introduced by Hurwicz [H] which cover both parametric and non-parametric specification. By a structure $S = (F, \Phi)$ we understand a specific probability distribution function $F(u)$ of the latent variables -- thought of as given numerically, either by a cumulative distribution surface or curve or table, or parametrically with numerical values of the parameters -- combined with a specific structural relationship (or set of simultaneously valid relationships)

$$\Phi(y, u) = 0$$

between apparent and latent variables -- again given numerically by curves, surfaces or parameters -- which permits unique determination of the apparent

variables from any given values of the latent variables. The corresponding probability distribution

$$G\,(y|S)$$

of the apparent variables is said to be generated by the structure S.

Using the term model for a set of structures, the specification problem can now be re-formulated as follows:

1*.) Specification of a model which by hypothesis contains the "true" structure generating the distribution of the apparent variables.

In particular, parametric specification of the structural relationship(s) consists in first instance in an indication of the mathematical form (e.g. linearity) of this (these) relation(s). It may be, and in all our examples is, supplemented by prescription of the numerical values of some of the parameters, or of given functions thereof. These prescriptions have sometimes been called a priori restrictions on the structure, specified by the model. It will be more appropriate here to call them parameter specifications, leaving open the question to which extent they are indeed determined a priori.

If a model (a set of structures) is regarded as the object of specification, it will be clear that a new problem of inference arises, which logically precedes all problems of estimation or the testing of hypotheses. It is implied in the definition of structure that a given structure S generates one and only one probability distribution G(y|S) of the apparent variables. However, statistical inference from observations can relate only to characteristics of the distribution of the apparent variables. The limit of statistical inference is an exact knowledge of this distribution function, a limit not attainable but approachable if very large samples

can be taken. Anything not implied in this distribution is therefore not a possible object of statistical inference.

Thus it is a question of great practical importance whether a statement converse to the one just made is valid: can the distribution of apparent variables, generated by a given structure S contained in a model $\mathcal{J}$, be generated by only one structure in that model? This is by no means implied in the definitions given, and it is not generally true. Whether or not it is true in a particular instance depends -- as illustrated in our examples -- always on the model, and often on the given structure besides. If it is true, we shall say that the model $\mathcal{J}$ (uniquely) <u>identifies</u> the given structure S.

If a structure S is not identifiable by a model $\mathcal{J}$, some of its characteristics may still be uniquely determinable. By a <u>structural parameter</u>

$$\theta(S)$$

we understand a functional of the structure S (This definition applies, of course, equally to the case of non-parametric specification of the functions F, $\mathcal{E}$ defining the structure). We further define that two structures S and S' are <u>equivalent</u> if they generate the same distribution,

$$G(y|S) = G(y|S') \quad \text{for all } y,$$

of apparent variables. Then we say that a model (uniquely) identifies a parameter $\theta(S)$ in a given structure S, if that parameter has the same value in all structures S', contained in the model, and equivalent to S. This definition can obviously be extended to characteristics $\chi(S)$ of a structure S other than parameters, such as the functional form of a relationship represented by a component of the vector Φ, etc.

In terms of these concepts, we can now add the following group of analytical problems to our list:

2.) Identification problems, i.e. study of the parameters or other characteristics of a given structure which are identified by the model.

While the problems involved are undoubtedly problems of inference, they proceed from the model rather than from the observations. It is perhaps debatable whether these problems can be called problems of statistical inference, except in the obvious sense that the object of their study is to explore the limits of statistical inference.

However this may be, interpenetration of the pre-statistical analysis of identifiability with problems of statistical inference arises from the fact, amply illustrated by our examples, that the identifiability of a structural characteristic $\chi(S)$ often depends not only on the model, but also on the given structure S. Thus, each structural characteristic χ divides the model $\mathcal{M}$ exhaustively into two submodels

$$\mathcal{M} = \mathcal{M}_\chi + \mathcal{M}_{\bar{\chi}}$$

(of which one may be empty), such that $\chi(S)$ is uniquely identified by the model if S belongs to $\mathcal{M}_\chi$, and not uniquely identified if S belongs to $\mathcal{M}_{\bar{\chi}}$. We shall call $\chi(S)$ uniformly identified if $\mathcal{M}_\chi$ coincides with $\mathcal{M}$.

The subdivision of $\mathcal{M}$ into $\mathcal{M}_\chi$ and $\mathcal{M}_{\bar{\chi}}$ has an important property: If S belongs to $\mathcal{M}_\chi$, then all structures equivalent to S also belong to $\mathcal{M}_\chi$, and a similar statement holds for $\mathcal{M}_{\bar{\chi}}$. This property follows directly from the definition of identifiability of $\chi(S)$ above. Its meaning is that the identifiability of $\chi(S)$ depends only on the distribution $G(y) = G(y/S)$ of apparent variables generated by S. To the subdivision of the model corresponds a subdivision

$$\mathfrak{M} = \mathfrak{M}_\chi + \mathfrak{M}_{\bar{\chi}}$$

of the set

$$\mathfrak{M} \equiv \mathfrak{M}(\mathcal{M})$$

of all distribution functions $G(y|S)$ generated by the structures S of $\mathcal{M}$, into the subset $\mathcal{M}_\chi$ containing those distribution functions $G(y|S)$ generated by structures S for which $\chi(S)$ is uniquely identifiable, and the subset $\mathcal{M}_{\bar{\chi}}$ containing $G(y|S)$ generated by structures for which the opposite is true.

Hence, whenever the identifiability of $\chi(S)$ cannot be decided in the same sense (affirmatively or negatively) for all structures S of $\mathcal{M}$ as a result of either $\mathcal{M}_\chi$ or $\mathcal{M}_{\bar{\chi}}$ being empty, then the identifiability of the characteristic $\chi(S)$ of the "true" structure S generating the observations is a property of the distribution $G(y|S)$ of the observations. This identifiability is equivalent to the statistical hypothesis

$$G(y|S) \text{ belongs to } \mathcal{M}_\chi,$$

and as such it is subject to statistical testing.

Often the model consists of one general specification supplemented with a number of particular specifications which are "detachable pieces" in the sense that they can be removed, added or replaced by alternatives to construct alternative models. We may define the general specification as a set $\mathcal{M}_o$ of structures which is postulated to contain the model $\mathcal{M}$ as a subset. Particular specifications can then be defined as subsets $\mathcal{M}_1, \mathcal{M}_2, \ldots$ of $\mathcal{M}_o$ of which the model is the intersection (set product)

$$\mathcal{M} = \mathcal{M}_1 \times \mathcal{M}_2 \times \ldots$$

An example is that of parametric specification of the "form function" $\mathcal{F}(y,u)$ of the structural relationships and of the distribution function $F(u)$ of latent variables as the general specification, and specifications of the values of certain parameters of $\mathcal{F}$ and F as particular specifications.

In such situations, it is an important question whether a given particular specification is in principle subject to statistical test,

and if so, which minimum set of other particular specifications must (together with the general specification) be entered into the "maintained hypothesis" in order that that given particular specification be subject to statistical test. A formal answer to this question, facilitating specific answers in each concrete case, can be given as follows.

Let a model $\mathfrak{M}$ be narrowed down to an alternative model

$$\mathfrak{M}_1' = \mathfrak{M} \times \mathfrak{M}_1$$

by a particular specification $\mathfrak{M}_1$. This particular specification will be called <u>overidentifying</u> if the set $\mathfrak{H}(\mathfrak{M}')$ of all distribution functions $G(y/S')$ of apparent variables generated by a structure S' of $\mathfrak{M}'$ is a proper subset of the set $\mathfrak{H}(\mathfrak{M})$ of all distribution functions $G(y|S)$ generated by a structure S of $\mathfrak{M}$. A statistical test of an overidentifying particular specification can then be constructed by choosing as the maintained hypothesis

$$G(y) \text{ belongs to } \mathfrak{H}(\mathfrak{M})$$

and as the hypothesis subject to test

$$G(y) \text{ belongs to } \mathfrak{H}(\mathfrak{M}')$$

The particular specification $\mathfrak{M}_1$ remains subject to test if the model $\mathfrak{M}$ is stripped of such other particular specifications which are not necessary for the overidentifying character of $\mathfrak{M}_1$, although of course the outcome of the test may become either less or more certain as a result.

A frequent case of an overidentifying specification is that where a parameter $\theta(S)$ already identified for almost all structures S of $\mathfrak{M}$, is

restricted by ϕ_1 to a prescribed value, or to a prescribed point set not containing all points of its domain for all S of $\mathcal{M}$.

The reason for the choice of the term "overidentifying" lies in a conjecture that, in an important class of cases, all characteristics of $\gamma(S)$ identified for almost all S of $\mathcal{M}_1'$, would also be identified by a model $\mathcal{M}_2'$ for almost all S of $\mathcal{M}_2'$, where $\mathcal{M}_2'$ is either $\mathcal{M}$ or a model

$$\mathcal{M}_2' = \mathcal{M} \times \mathcal{M}_2$$

obtained by adding to the general specification incorporated in $\mathcal{M}$ a particular specification $\mathcal{M}_2$ which is less restrictive than $\mathcal{M}_1$ (i.e. such that $\mathcal{M}_1$ is a proper subset of $\mathcal{M}_2$). We shall not attempt to give a proof of this conjecture, or to determine in which class of cases it is valid.

(to be continued)