

Some Aspects of Linear Estimation Problems
and Problems of Statistical Estimation

Statistics: 338

By

Erling Sverdrup

1. The problem that is going to be treated is of a rather general nature, but could be well illustrated by the following familiar example.

The model is defined by means of a system of linear structural equations where the disturbances are serially independent but not necessarily normal. All predetermined variables are exogeneous. It is well known that the least square method applied to the reduced form equations gives you estimates of the parameters in the reduced form which are:

1. consistent
2. unbiased
3. of least variance among the linear unbiased estimates.

If the model is "just identified" we can find estimates of the parameters in the structural equations by applying to the estimates of the reduced form parameters the same transformation which transforms the true reduced form parameters into the true structural parameters. By this transformation our property 1 above will still hold (Slutsky [1] p. 75); but properties 2 and 3 will, in general, not hold.

This raises the problem of which parameters are the ones for which we want certain optimum properties to hold. Which parameters do we want to estimate and which optimum properties do we want them to have?

Since the ultimate aim of any statistical inference is to predict something, and since in particular the notion of a "structural" equation is incomprehensible without reference to the prediction purpose, it is obvious that we cannot answer

this question without stating our prediction problem. Furthermore, if the sole purpose of our model is to solve a well defined prediction problem then everything we do, also the estimation, must be subordinate to this purpose. The arbitrariness in choice of the estimation principle, which is present in most statistical treatments, is no longer satisfactory.

The common way of solving the prediction problem is this: first, estimates of the unknown parameters are found, which fulfill some arbitrary chosen optimum population properties. Then an optimum way of predicting on the basis of known parameters is worked out. By this two-step procedure we have, of course, no guarantee that we have optimum prediction on the basis of our a priori knowledge and the observations. The inefficiency of this two-step procedure has been pointed out by Haavelmo [2] p. 109.

It is my purpose here to analyze the mechanism by which the prediction problem uniquely defines the intermediary step, namely, the statistical inference problem.

The prediction problem is taken in the general sense where some "action parameters" are involved, the variation of which may have repercussions on our stochastic process.

The statistical inference problem is taken in the Wald sense as defined by a weight function and a class of subsets of the space of admissible hypothesis. (Wald [3])

Of course, it must be realized that very often when setting up the model and estimating the parameters, the prediction purpose is not clear. The model is intended to serve a multitude of prediction purposes which may arise in the future. But even if this is the case, it might be of interest to see how a well defined prediction problem defines an estimation problem.

Thus, the aim of this discussion paper is to try to obtain conceptual clarity with regard to the link between estimation and prediction and not so

much to obtain workable rules for estimating and predicting. However, what is developed below may be helpful in deciding which weight functions, etc. to use in practical cases.

The general theory does not necessarily presume that our model is defined by means of stochastic equations.

2. If we shall be able to predict something it must be because there is a certain persistence in the mechanism which produces the data. This mechanism must be defined in such a way that it gives rules for "structural changes".

More rigorously, this mechanism can be defined by means of a probability measure.

$P_\eta(\beta, c)$ for the set β . η is a random variable of the form

$$\eta = \{\eta(\tau)\}_{\tau} = -\infty_1 + \infty$$

where $\eta(\tau)$ for fixed τ is a random variable with ℓ components. c is an action parameter belonging to a space Γ' . β is any set belonging to the least completely additive class of sets containing all cylindric sets with Borel basis. (Kolmogoroff [4] p.).

Our a priori knowledge consists in stating that $P_\eta(\beta, c)$ belongs to some space Ω' of functions of β and c with the properties mentioned above. From the stochastic process $P_\eta(\beta, c)$ we derive a new stochastic process $P_\xi(\beta, c)$ in which the points of time, when we make observations and predict, are involved. Let $t_1 < t_2 < \dots < t_n$ be the points of time for observation and let c_0 be the value of c before t_1 , c_1 the value of c between t_1 and t_2 , etc. c_n the value after t_n . Let $t \geq t_n$ be the time when we want to predict something. For any choice of $\bar{t}_1 < \dots < \bar{t}_p$ and p we define a random variable $\{\xi(\bar{t}_1), \dots, \xi(\bar{t}_p)\}$ where each $\xi(\bar{t}_j)$ has ℓ components. The c.d.f. for this random variable is defined as

$$\begin{aligned}
& P_{\xi}(\bar{\tau}_1) \dots \xi(\bar{\tau}_p)^{(b_1 \dots b_p; c)} \stackrel{=}{=} P_{\eta}(\bar{\tau}_1) \dots \eta(\bar{\tau}_{i_1})^{(b_1 \dots b_{i_1}; c_0)} \cdot \\
& \cdot P_{\eta}(\bar{\tau}_{i_1+1}) \dots \eta(\bar{\tau}_{i_1+i_2})^{(b_{i_1+1} \dots b_{i_1+i_2}; c_1)} \Big| \eta(\bar{\tau}_1) \leq b_1 \dots \\
& \eta(\bar{\tau}_{i_1}) \leq b_{i_1} \Big) \dots P_{\eta}(\bar{\tau}_{q+1}) \dots \eta(\bar{\tau}_p)^{(b_{q+1} \dots b_p; c)} \Big| \dots \\
& \eta(\bar{\tau}_q) \leq b_q \tag{2.1}
\end{aligned}$$

where

$$\begin{aligned}
& \bar{\tau}_1 < \bar{\tau}_2 < \dots < \bar{\tau}_{i_1} \leq t_1 < \bar{\tau}_{i_1+1} < \dots < \bar{\tau}_{i_1+i_2} \leq t_2 < \dots \\
& \dots < \bar{\tau}_q \leq t < \bar{\tau}_{q+1} < \dots < \bar{\tau}_p
\end{aligned}$$

By $\eta(\bar{\tau}_j) \leq b_j$ we mean $\eta^i(\bar{\tau}_j) \leq b_j^i$, $i = 1, 2, \dots, l$. If there is more than one t between any two $\bar{\tau}$ we insert $\bar{\tau}$'s between these/and get a series of $\bar{\tau}$'s and t 's with the property given above. We now use the consistency property and symmetry property of Kolmogoroff [4], and define the c.d.f. for $\xi(\bar{\tau}_1) \dots \xi(\bar{\tau}_p)$ for any $\bar{\tau}_1 \dots \bar{\tau}_p$. The probability measure $P_{\xi}(\beta, c)$ is then uniquely defined for the same class of β 's as for the η -process. c is confined to a subset Γ of Γ' . Note that the c.d.f. for $\xi(\bar{\tau}_1) \dots \xi(\bar{\tau}_p)$ if all $\bar{\tau}$'s are less than t is independent of c .

We introduce the notation

$$X = \{ \xi(t_1), \dots, \xi(t_n) \} \tag{2.2}$$

for the observed random variable and

$$Y = \{ \xi(\bar{\tau}) \} \bar{\tau} > t \tag{2.3}$$

for the future random variable. x is a sample point. X and Y is the space of all X and Y .

We consider

$$P_Y(S) = \Pr(Y \in S \mid X = x) \tag{2.4}$$

For any $P_{\eta} \in \Omega'$ and $c \in \Gamma'$, P_Y is uniquely determined. P_X is also uniquely determined for each P_{η} . Let the space of all P_X be Ω . Ω is, of course, generated by varying P_{η} within Ω' .

We now make the following fundamental assumption which is, in a way, a

generalization of the assumption that a model should be identified.

Assumption 1. To any two P_{η} in Ω' the corresponding two F_X in Ω will not be identical, i.e., F_X determines P_{η} uniquely.

If assumption 1 is fulfilled the $P_Y(S)$ is uniquely determined for each F_X and we write $P_Y(S | c, x, F_X)$.

The above set-up can easily be extended to the case where the components of X do not comprise all components of $\xi(t_j)$, $j = 1, \dots, n$, i.e., where not all components are observed.

In designing the statistical investigation, i.e., in choosing X , we will always attempt to make such a design that assumption 1 is fulfilled, i.e., that our model is identified. In trying to choose an identified model we must make use of our a priori knowledge, which is given by the space Ω' of all P_{η} .

If our a priori knowledge were only given by the space Ω of all F_X then this would be of no help in solving the design of statistical investigation problem (the identification problem), since F_X is only given if we know what we want to observe. In that case we can only state if our model is identified or not, we can not make it identified.

3. We now define the utility (or "gain") $V(y,c)$ of obtaining a value y of Y by taking action c at time t . $V(Y,c)$ is for each $c \in \Gamma$ a measurable function of y with respect to the class of all B .

The expected future utility is now

$$E V(Y,c) = \int_{\mathcal{Y}} V(y,c) d P_Y(y | c, x, F_X) \tag{3.1}$$

and this expectation is a function (al) of F_X , c and x . With known F_X and x we want to take the action c which maximizes the expected utility. The statistical material gives us x directly. It is a statistical inference problem to determine F_X . Note that this set-up is rather general. In some cases $V(y,c)$ can be given in money units, in other cases it might be possible to construct preference charts. Again, if $V(y,c) = 1$ if $y \in S$ and 0 otherwise, then we

simply want the probability of $y \in S$ as large as possible.

Let the set of all c for which $E V(Y,c)$ is maximized for given F_X and x be $\bar{c}(F_X, x)$.

That is, if $c \in \bar{c}(F_X, x)$, then

$$\int V(y, c) d P_Y(y | c, x, F_X) = \sup_{\delta \in \mathcal{F}} \int V(y, \delta) d P_Y(y | F_X, \delta, x) \quad (3.2)$$

Let \mathcal{F} be the class of all set $\omega \subset \Omega$ such that for any $\omega \in \mathcal{F}$

$$(i) \bigcap_{F_X \in \omega} \bar{c}(F_X, x) \text{ is nonempty for all } x. \quad (3.3)$$

$$(ii) \text{ for any } F_X^{(1)} \in \Omega, F_X^{(1)} \notin \omega,$$

$$\bar{c}(F_X^{(1)}, x) \bigcap_{F_X \in \omega} \bar{c}(F_X, x) = \emptyset, \text{ for some } x.$$

In other words, ω is the set of all F_X which would lead you to take the same action. Of course, the different ω may be overlapping.

It is easily seen that no set in \mathcal{F} is a ^{proper} subset of another, and consequently, if \mathcal{F} contains Ω it contains only Ω . In that case we can decide which action to take on the basis of our a priori information and we have no statistical problem.

We now make the assumptions

Assumption 2. $\sum_{\omega \in \mathcal{F}} \omega = \Omega$, i.e., the sets in \mathcal{F} covers Ω . This implies, of course, that \mathcal{F} is nonempty.

Assumption 3. For all $\omega \in \mathcal{F}$

$$\bigcap_{F_X \in \omega} \bar{c}(F_X, x)$$

contains just one element, which we denote $c(x, \omega)$.

On intuition, it is reasonable to presume that assumption 3 is fulfilled in "many cases" since the definition of \mathcal{F} implies that (3.3) "is on the verge of becoming empty."

It would be desirable to have necessary and sufficient conditions for assumptions 1 - 3 to be fulfilled in terms of $V(y, c)$ and $P_{\eta}(\beta, c)$. However, I have not been able to obtain such conditions.

We can now determine the loss in utility which we suffer if we state that the distribution function of X belongs to ω , whereas F_X is the true distribution.

$$W(F_X, \omega, x) = \sup_{\gamma} \int_{\mathcal{Y}} V(y, \gamma, x) dP_Y(y | F_X, \gamma, x) - \int_{\mathcal{Y}} V[y, c(\omega, x)] dP_Y[y | F_X, c(\omega, x), x] \quad (3.4)$$

This gives us the weight function to be used in the statistical inference problem, i.e., in choosing between the different members of \mathcal{F} .

In the serially independent case W will not depend on x . But it is interesting to see that in the case of serial dependence x will in general occur in W . This is due to the double purpose of x in a stochastic process. x shall tell us something about F_X (statistical inference) and at the same time enter directly as a kind of initial (boundary) condition for prediction.

After having defined \mathcal{F} and W we can proceed in the manner described by Wald [3]. We define a statistical procedure \mathcal{A} which determines for each x in the sample space, the element ω_x of \mathcal{F} we want to select. By means of the decision function ω_x we define the risk function

$$E W(F, \omega_x, X) = r(F, \mathcal{A}) \quad (3.5)$$

We can now apply Wald's minimax principle (or some other principle) to this risk function.

We have reached the following conclusion. If assumption 1 - 3 are fulfilled, then the prediction problem $\{R_{\eta}(\beta; c), V(y, c), X\}$ can be reduced to a statistical inference problem $\{\mathcal{F}, W(F_X, \omega, x)\}$.

After having solved the statistical inference problem, the action c is determined by

$$c = c(x, \omega) \quad (3.6)$$

4. In order to illustrate the general principle we shall first study a simple example outside the field of economics.

Example 1. The components of $\eta(\bar{t})$ are observations of the amount of crop on different plots. The components are independent and the process is

serially independent. The action parameters c express the application of two different fertilizers I and II. There is just one point of time for observation ($n = 1$) and this is also the point of time for prediction $t_1 = t_n = t$.

We observe p plots where fertilizer I has been applied and q plots where fertilizer II has been applied.

$$X = \left\{ X_1 \dots X_p \quad X_{p+1} \dots X_{p+q} \right\}$$

$$F_X(x) = \prod_{i=1}^p G(x_i - \mu_1) \prod_{i=p+1}^{p+q} G(x_i - \mu_2)$$

where G is a distribution function such that $\int z dG(z) = 0$.

We want to predict the future product y on a plot if fertilizer I ($c=1$) or fertilizer II ($c=0$) is used.

$$F_Y(y) = c H(y - \mu_1) + (1 - c) H(y - \mu_2)$$

where H is a distribution function. $\Gamma = \{0, 1\}$. Both fertilizers are equally expensive and we want as great a crop as possible. We can therefore set

$$V(y, c) = y$$

By easy calculation, we get (by 3.1)

$$E V(y, c) = c \mu_1 + (1 - c) \mu_2 + \int z dH(z)$$

$\bar{c}(F_X, x)$ is independent of x , and we get (by 3.2)

$$\begin{aligned} \bar{c}(F_X) &= \{1\} && \text{if } \mu_1 > \mu_2 \\ \bar{c}(F_X) &= \{1, 0\} = \Gamma && \text{if } \mu_1 = \mu_2 \\ \bar{c}(F_X) &= \{0\} && \text{if } \mu_1 < \mu_2 \end{aligned}$$

$\omega(c, x)$ is of course, also independent of x and we get (by 3.3),

$$\begin{aligned} \omega(1) &= \{F_X | \mu_1 \geq \mu_2\} && \omega(0) = \{F_X | \mu_1 \leq \mu_2\} \\ W[F, \omega(1)] &= 0 && \mu_1 \geq \mu_2 \\ W[F, \omega(0)] &= \mu_1 - \mu_2 && \mu_1 \geq \mu_2 \\ W[F, \omega(1)] &= \mu_2 - \mu_1 && \mu_1 \leq \mu_2 \\ W[F, \omega(0)] &= 0 && \mu_1 \leq \mu_2 \end{aligned}$$

Let A be a set in the sample space \mathcal{X} and $\bar{A} = \mathcal{X} - A$ its complement. The statistical procedure consists of accepting $\omega(1)$ if $x \in A$, otherwise $\omega(0)$.

By (3.4) we get

$$W(F, \omega_x) = \begin{cases} 0 & x \in A \\ \mu_1 - \mu_2 & x \in \bar{A} \\ \mu_2 - \mu_1 & x \in A \\ 0 & x \in \bar{A} \end{cases} \quad \begin{matrix} \mu_1 \geq \mu_2 \\ \mu_1 \geq \mu_2 \\ \mu_1 \leq \mu_2 \\ \mu_1 \leq \mu_2 \end{matrix}$$

Let us introduce the notation

$$p(A) = \Pr (x \in A)$$

$p(A)$ depends, of course, on μ_1 and μ_2 .

We now get for the risk function (by 3.5)

$$EW = r(F, A) = \begin{cases} (1 - p(A)) (\mu_1 - \mu_2) & \text{if } \mu_1 \geq \mu_2 \\ p(A) (\mu_2 - \mu_1) & \text{if } \mu_1 \leq \mu_2 \end{cases}$$

The statistical problem is now determined. Roughly speaking, it is seen from the last equation that we want $p(A)$ large if $\mu_1 > \mu_2$ and $p(A)$ small if $\mu_1 < \mu_2$ which corresponds "almost" to Neyman-Pearson's principle. $p(A)$ is, of course, nothing but the power function.

Note that a whole family of prediction problems [corresponding to different $H(s)$] leads you to the same problem of statistical inference.

Example 2. Let P_t be the price and Q_t the quantity sold of a commodity at time t . $X_t = \log Q_t$, $Y_t = \log P_t$.

$$\eta(t) = \{X_t, Y_t\}$$

$P_\eta(\beta, c)$ is defined by the stochastic equations,

$$\epsilon_{1t} = X_t - a Y_t - b \quad (\text{demand relation})$$

$$[(\epsilon_{2t} - X_t + \alpha(Y_t - c_1) + \beta)]c_2 + (1 - c_2)(c_1 - Y_t) = 0$$

where $c = \{c_1, c_2\}$ is the action parameter. c_2 is either 1 (free competition) or 0 (monopoly). Let T be a sales tax. $c_1 = -\log(1 - T)$ if $c_2 = 1$ and c_1 is equal to the price fixed by the monopolist if $c_2 = 0$. This gives a definition of Γ' . The past is characterized by free competition with varying tax. c_1, X_t, Y_t are observed at time t_1, t_2, \dots, t_n .

$$c_j = \left\{ c_{1t_j}, 1 \right\} \quad j = 1, 2, \dots, n.$$

$$X = \left\{ X_{t1} \dots X_{tn}, Y_{t1} \dots Y_{tn} \right\}.$$

At time t a monopoly is formed which wants to maximize its profit at time $t + 1$. The cost of producing a quantity Q is $AQ + B$.

We then get

$$V(Y, c) = e^{X_{t+1}} (e^{c_1} - A) - B$$

$(\xi_1 \tau, \xi_2 \tau)$ is a serially independent process and their distribution function is independent of a, b, α, β and τ . $E \xi_i \tau = 0, i = 1, 2$.

We presume a priori $a < -1$.

We get

$$E V(Y, c) = e^a c_1 + b (e^{c_1} - A) E (e^{\xi_1 \tau}) - B$$

We get

$$\bar{c} (F_X, x) = \left\{ \log \frac{A}{a+1} \right\}$$

\bar{c} has always just one element.

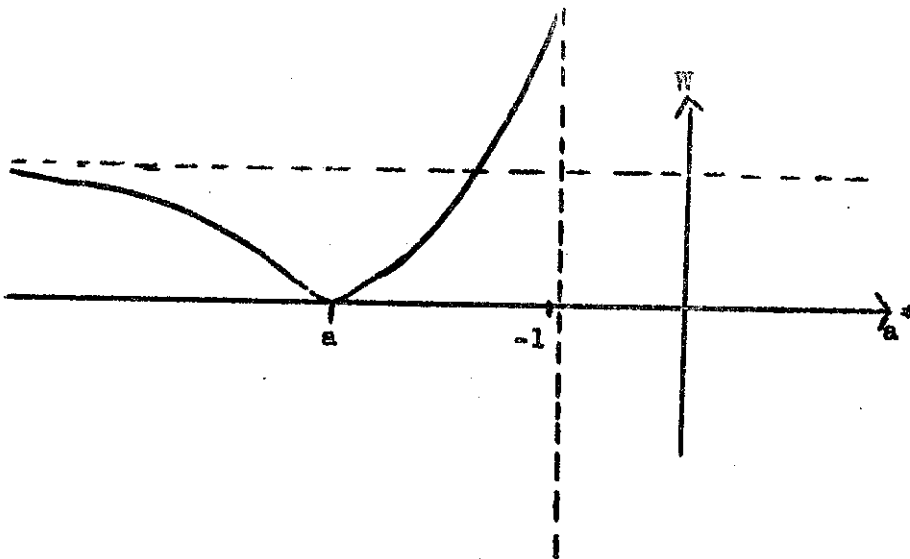
The set of All F_X which lead you to fix the same monopoly price is simply the set of all F_X with the same a . Let a^* be an estimate of a . We get the following weight function to be used when estimating the elasticity of demand.

$$A^{a+1} \left[\left(\frac{a}{a+1} \right)^a \left(-\frac{1}{a+1} \right) \left(\frac{a^*}{a^*+1} \right)^a \left(-\frac{1}{a^*+1} \right) \right] e^b E(e^{\xi_1 \tau})$$

This is loss in profit which the monopolist suffers due to the fact that he estimates the elasticity to be a^* whereas the true elasticity is a .

The last two factors can be left out without changing the minimax solution, and we get the following weight function.

$$W(a, a^*) = A^{a+1} \left[-\frac{a^a}{(a+1)^{a+1}} + \frac{a^{*a}}{(a^*+1)^{a+1}} \right]$$



After having found the elasticity of demand, the monopolist knows how to fix the price. He sets the price equal to

$$\log A + \log\left(\frac{a^*}{a^*+1}\right)$$

We have given an example of how the weight function in some cases can be determined. But what is much more interesting is the following: The fact that the monopolist wants to fix the price such that the profit is maximized gives you a unique definition of what is the "best" estimate of the elasticity of demand. This gives a clear meaning to the phrase, "the choice of estimation principle depends on the practical purpose of the statistical investigation."

Usually we want the model to serve many different purposes. In that case, there exists, in general, no unique "best" estimate and we are, of course, justified in making a more or less arbitrary choice of estimation principle.

Example 3. By "passive" prediction is meant prediction where the action parameter does not enter in the η -process. Of course, the action parameter must then be present in the utility function (if the statistical problem were not related to any kind of "action" at all, then the statistical investigation would have no practical aim). The passive prediction problem is the type of prediction mostly dealt with by statisticians. The following example is meant to illustrate how these problems can be considered as special cases of the type

problems outlined in 2 and 3.

Let $P_\eta = P_\xi$ be defined by

$$\xi(\tau) + a_1 \xi(\tau - 1) + \dots + a_p \xi(\tau - p) = \xi\tau$$

where $\xi(\tau)$ is a scalar, $E(\xi\tau) = 0$, $E(\xi\tau^2) = \sigma^2$, the distribution function of $\xi\tau$ is independent of $\{a_1 \dots a_p\} = a$ and τ and $\xi\tau$ is a serially independent process. Prediction takes place at time t . $\xi(\tau)$ has been observed at time $t, t - 1, \dots, t - n + 1$. ($n > p$).

$$X = \{\xi(t), \xi(t - 1), \dots, \xi(t - n + 1)\}.$$

We want a "point" prediction of $Y = \xi(t + 1)$, i.e., we want to determine a functional form f such that $f(X)$ can be used as a predictor for $\xi(t + 1)$.

We choose to determine f such that

$$\sup_a \int [\xi(t + 1) - f(X)]^2 dP_\xi$$

is minimized.

In order to see how this type of prediction comes under what has been described above it is only necessary to reason as follows. Since we want a "point" prediction of $\xi(t + 1)$ it must be because any two different values of $\xi(t + 1)$ would lead to different actions. There is then a one-to-one correspondence between action c and the predicted value of $\xi(t + 1) = Y$. Consequently action c could be identified with the predicted value of $\xi(t + 1)$. We then choose the utility function.

$$V(Y, c) = - [\xi(t + 1) - c]^2$$

We get

$$EV = - \sigma^2 - [a_1 \xi(t) + \dots + a_p \xi(t - p + 1) - c]^2$$

$$\bar{v}(a, X) = - [a_1 \xi(t) + \dots + a_p \xi(t - p + 1)]$$

It is easily seen that the class \tilde{f} consists of sets each of which correspond to one value of $a = \{a_1 \dots a_p\}$. We are led to point estimation of $a_1 \dots a_p$.

The weight function is

$$W(a, a^*, x) = (a_1 - a_1^*) \xi(t) + \dots + (a_p - a_p^*) \xi(t - p + 1)^2$$

where a^* is an estimate of a .

The predictor is then

$$f(X) = \hat{c}(a^*, X) = - [a_1^* \xi(t) + \dots + a_p^* \xi(t + p + 1)]$$

The introduction of the expression on the right hand side instead of $f(X)$ does not, of course, limit the possible forms of $f(X)$. We have further,

$$\begin{aligned} \text{risk function} = EW &= \int [E_{t_{t1}} \xi(t+1) - f(X)]^2 dP \\ &= \int [\xi(t+1) - f(X)]^2 - \sigma^2 \end{aligned}$$

This leads us to the following result.

The problem of finding minimax point estimator $f(X)$ for $\xi(t+1)$ can be solved as follows:

Estimate $a_1 \dots a_p$ by finding minimax of the risk function corresponding to the weight function $W(a_1 a_1^* x)$.

Then use the estimator

$$f(X) = - [a_1^* \xi(t) + \dots + a_p^* \xi(t - p + 1)]$$

5. Assumption 1 is really too restrictive. It could be replaced by

Assumption 1'. Consider

$$E V(Y, c) = \int_{\mathcal{Y}} V(Y, c) d P_Y(S | x)$$

as a function of c and x . For each admissible functional form for this function, consider the set of all P_η which leads you to this functional form. We require these sets of that to any of P_η there corresponds just one F_X for the whole set.

Even this assumption may be too restrictive in some cases, i.e., there exist important examples where the theorem in 3 is fulfilled without assumption 1' being fulfilled.

REFERENCES

- [1] Eugen Slutsky: "Ueber stochastische Asymptoten und Grenzwerte."
Metron, vol. 5, [1925].
- [2] Trygve Haavelmo: "The Probability Approach in Econometrics."
Econometrica, vol. 12, [1944], Supplement.
- [3] Abraham Wald: On the Principles of Statistical Inference.
Notre Dame Mathematical Lecture, No. 1, [1941].
- [4] A. Kolmogoroff: "Grundbegriffe der Wahrscheinlichkeitsrechnung."
Berlin, [1933].