COWLES COMMISSION DISCUSSION PAPER: STATISTICS NO. 344

Prediction Problems and the Theory of Statistical Decision Functions

by

Erling Sverdrup

CONTENTS

§1. Introduction

§ 1.1. This section will outline in a not too abstract mathematical language, some problems which are defined rigorously in § 3. An attempt will be made to emphasize the generality and importance of certain prediction situations in statistical problems connected with stochastic processes.

§ 1.2. Very often prediction problems can be formulated in the following manner: For the observable and future random variables you are given a certain class of joint probability distributions. On the basis of 1) the information that your true distribution is a member of the given class and 2) a set of observations of the observable random variables, you want to make a certain statement involving your future random variable.

To make the idea more concrete, consider as an example, the following situation. Let the observable random variable be a finite dimensional vector X and let the future random variable be a one dimensional vector Y. Let the class of probability distributions for (X, Y) be Ω and let P be an arbitrary member of Ω. We want to determine a rule $Y = g(X)$ such that for any observation X^0 of X the future random variable Y can be predicted as $Y^0 = g(X^0)$. It has been preposed by Haavelmo [1, page 109], among others, to apply the minimax principle to the following expression

$$\int [Y - g(X)]^2 \, dP \qquad (1)$$

where the integral is taken over the space of all (X, Y). This expression could be considered as a measure of the average goodness of the prediction function g if P is the true probability measure. According to the minimax principle we want to find a function $g^*(X)$ (if existing) such that

$$\sup_{P \in \Omega} \int [Y - g^*(X)]^2 \, dP = \inf_{g} \sup_{P \in \Omega} \int [Y - g(X)]^2 \, dP \, . \qquad (2)$$

Of course, functions of Y and G other than $[Y - g]^2$ could be applied in (1) and (2).

§1.3. A way of solving prediction problems which is very often recommended (see

e.g. Mann and Wald [2, page]) and very commonly applied in the following: First some "statistical inference" is made on the basis of an observation X^0 of the observable random variables X. By statistical inference, we mean either estimation of certain parameters in P which generate Ω or testing certain hypothesis concerning P. In general, any statement on the basis of X^0 which amounts to saying that P belongs to a proper subset ω_{X^0} of Ω is a statistical inference. The statistical inference method is chosen such that certain conditions and optimum properties are fulfilled, e.g. that of having uniformly smallest variances among unbiased estimates, or uniformly largest power among all tests at certain levels of significance. Usually these conditions and optimum properties are chosen without regard to the prediction problem which ultimately has to be solved. After having solved the statistical inference problem, the prediction about the future random variable is made on the basis of the information that P belongs to ω_{X^0}. The "best" way of predicting is usually defined relatively to the ω we have decided upon.

To illustrate the idea, let us consider the following situation. (U_i, V_i) $i = 1, \ldots N+1$ are $N+1$ independently identically distributed variables; each has binormal distribution with unknown mean and moment matrix. We want to predict $Y = V_{n+1}$ on the basis of an observation of

$$X = (U_1, V_1, U_2, V_2 \ \ldots\ U_N, V_N, U_{N+1}) .$$

The common way of doing this is to find estimates of α and β in $E(V|\ U) = \alpha + \beta U$ by the method of least squares. Let us call them α^* and β^*. (It is well known that these estimates fulfill some rather nice conditions and optimum population properties.) α^* and β^* are functions of $(U_1\ V_1 \ \ldots\ U_N\ V_N)$. As a predictor for V_{N+1}, then $V^*_{N+1} = \alpha^* + \beta^* U_{N+1}$ is used.

The inefficiency of this two-step procedure has been pointed out by Haavelmo [1, page 109]. By this method we have no guarantee that we have made the best possible prediction on the basis of the observation. Very often the "best" way of predicting

on the basis of x^0 and Ω is not even defined and we are not able to tell precisely in which way our predictor is good. We therefore conclude: If we have a prediction problem in mind, then the arbitrariness about the choice of the statistical inference principle is no longer satisfactory.

§ 1.4. In a great number of cases of practical importance, the problem is not that of a passive prediction such as described in the example in § 1.2, it is a problem of policy. In agricultural experimentations the problem is not always: How large will the crop be next year, more often it is: What method of cultivation should be used. (E.g. which fertilizer is the best).

As another example, let us consider the problem in "Collective Risk Theory" developed chiefly by Lundberg [3]. Here the problem is to predict whether an insurance company, with the funds they have at their disposal and the premiums they charge, will suffer a deficit in the future. It is obvious that treating this problem as a passive prediction problem is not adequate. If a deficit is predicted the insurance company will probably start to do something about it and possibly invalidate the prediction. The company will ask the statistician: What would your prediction be if we did so and so; or, What action should be taken to prevent a deficit (and at the same time obtain some other aims).

The same would probably be the case in predictions of the stock market. The fact that a prediction is made and relied upon will possibly invalidate the prediction. The prediction will have repercussions on the stochastic model we are considering and may change it completely.

The general prediction problem is then the following:

We are given a stochastic model (i.e. a class of probability measures) for our stochastic variables which, in general, depend on some action (or "policy") parameter and have an observation of the random variable in the past. We are also given a utility

function which depends on the outcome of our random variable in the future and on our action parameter. We are interested in making the expected value of the utility large. If the stochastic model is the same for all values of the action parameter, then we have a passive prediction problem.

The example of point prediction, mentioned in § 1.2 is covered by this set-up. The problem in this example can be interpreted as follows: Since we are interested in a point prediction, g of Y, it must be because different predictions of Y will lead to different actions. There must be a one to one correspondence between the prediction g and the action. But in that case g could be identified with "action." The utility function used in the example is then $-(Y - g)^2$. The model is in this case independent of the action.

The action parameter should always enter either in the utility or in the model (or in both); if it is absent both places such that our problem has no connection with human actions, then it is hard to see its practical importance.

Problems in planned economy are, of course, typical of the general set-up outlined above.

§ 1.5. "Predicting" something about a random variable is, of course, neither more nor less than saying something about its distribution function. A. Wald's [4] general theory of statistical decision functions is therefore nothing but a theory of prediction and his formulation covers almost every problem earlier treated within the field of what is usually meant by mathematical statistics. It seems, therefore, that there is nothing more to be said about the subject so far as general formulation is concerned.

However, it has become common in statistics to treat any statistical problem as if the ultimate aim always was to obtain a statistical procedure with some standard optimum properties regardless of the specific prediction purpose. This is probably the reason why Wiener cirtically remarks - "I may remark parenthetically that the modern

apparatus of the theory of small samples, once it goes beyond the determination of its own specially defined parameters and becomes a method of positive statistical inference in new cases, does not inspire me with any confidence, unless it is applied by a statistician by whom the main elements of the dynamics of the situation are either explicitly known or implicitly felt." Wiener [5, pages 34-35.]

It is the purpose of this discussion paper to show that if "the dynamics of the situation are ... explicitly known" and if the aim is "positive statistical inference in new cases", then the statistical decision does not go "beyond the determination of its own specially defined parameters."

A part of the requirement that "the dynamics of the situation are explicitly known" is the requirement that some specified parameters must be "identified" [in the sense in which this word is used by econometricians. (A definition will be given later.)]. Thus the problem of identifiability becomes, in a natural manner, a part of "positive statistical inference in new cases."

§ 2. Two Fundamental Concepts

§ 2.1. In this section we shall define two concepts which we shall apply later. The first is that of statistical inference (decision) in the sense of A. Wald, the second is that of "identified parameters in a model".

§ 2.2. Statistical inference defined.

Let X be a random (vector) variable in the space $\mathcal{X}$ and Ω a set of probability measures φ for X (i.e. for a class of Borel sets in $\mathcal{X}$). Let, further, ϕ be a class of subsets ω of Ω, the union of which covers Ω. The weight function $W(\varphi, \omega)$ is a function from $\Omega \times \phi$ to the real line.

Let x be an "observed value" of X. It is a "drawing" from one φ in Ω. The statistical inference consists in "deciding" to which ω in ϕ this φ belongs. More precisely, we want to find a function $\omega = d(x)$ from $\mathcal{X}$ to ϕ fulfilling certain optimum properties. This function is the decision function. The optimum properties of the

decision function d(x) is expressed in terms of the risk function

$$r[\varphi, d] = \int_{\mathcal{X}} W(\varphi, d(x)) d\varphi$$

(where it is assumed that d(x) is such that $W(\varphi, d(x))$ is measurable). Let D be a functional space of the d's. According to the minimax principle we want to find a function d^0 (if existing) such that

$$\sup_{\varphi \in \Omega} r[\varphi, d^0] = \inf_{d \in D} \sup_{\omega \in \Omega} r[\varphi, d].$$

(This is a special case of the von Neumann-Morgenstern zero-sum two person game.) On the basis of the statistical material x, we then make the statistical inference that $\varphi \in \omega = d^0(x)$.

The set-up can be generalized to cover the sequential (and design of experiment) situation. See Wald [4].

§ 2.3. Identification defined. Let M be a set in the q-dimensional space. To each member $\theta = (\theta_1 \ldots \theta_q)$ if M corresponds to a probability measure $\varphi(B; \theta)$ for B, where B is a Borel-set in the space $\mathcal{X}$. M is called a model, each member θ of this model, a structure. Two structures θ^1 and θ^2 are called equivalent if the corresponding probability measures are the same. Let us consider, say, the j^{th} component of θ. This j^{th} component θ_j is said to be identified if for any two equivalent structures θ^1 and θ^2 of M the j^{th} component is necessarily the same $\theta^1_j = \theta^2_j$. We have complete identification if this is true for j = 1, 2 ... q. This definition is the same as that given by Koopmans in Monograph 10 [6] and Haavelmo [1, page 92]. The definition is easily generalized to a nonparametric situation.

§3. The General Set-up

§ 3.1. We shall now try to give a more precise formulation of the prediction situations outlined in § 1.

§ 3.2. In order to predict something there must be a certain persistence in the mechanism which produces the data. This mechanism must also involve rules as to how your action influences your model (if it does). Let us express this in purely mathematical language.

Let Z be a functional space, each "point" ζ being a function of τ, where τ varies in a certain subset T of the real line.

$$\zeta = \zeta(\tau) \mid \tau \in T$$

where

$$\zeta(\tau) = (\zeta_1(\tau), \ldots \zeta_\ell(\tau)).$$

Let $\mathcal{Z}$ be a completely additive class of subsets S of Z. Let A be another functional space, where each point α is a certain number of functions of the real variable $\tau \in T$.

For each $\alpha = \alpha(\tau) \mid \tau \in T \in A$ we are given a class Ω_α of probability measures $\mu(S; \alpha)$ for the sets S of $\mathcal{Z}$. α is called the action parameter. (Some of the component functions of $\alpha(t)$ may be completely determined in A and may correspond to what an economist means by an "exogenous variable.")

§ 3.3. Let us introduce the point of time for prediction t.

Consider the set

$$T_t = [\tau \mid \tau \in T \text{ and } \tau \leq t].$$

We make

Assumption 1. Ω_α is such that the ("marginal") probability measure for $\zeta(\tau) \mid \tau \in T_t$ is independent of $\alpha(\tau) \mid \tau \in T - T_t$.

In other words, let $\mathcal{Z}_t$ be the set of all $\zeta(\tau) \mid \tau \in T - T_t$ generated by letting $\zeta(\tau) \mid \tau \in T$ run through all elements of $\mathcal{Z}$. Consider the class of all measurable sets in $\mathcal{Z}$ which contain $\mathcal{Z}_t$ as a subset. This class forms a Borel-field and a measure over the field is defined by means of μ. It may be called the marginal probability measure of $\zeta(\tau) \mid \tau \in T_t$. It is required in Assumption 1 that the class of these measures generated by varying μ in Ω_α is the same for any α for which $\alpha(\tau) \mid \tau \in T_t$ is given. (In other words, Assumption 1, expresses an axiomatic property about "action" α. You cannot, by an action, change what has already happened.)

We also make the

Assumption 2. For any $\tau \in T_t$, $\alpha(\tau)$ is completely determined by the fact that $\alpha \in A$. (In other words, our past action is known to us.)

Let $t_1 < t_2 < \ldots < t_n$ belong to T_t. They are the points of time when we observe $\xi(\tau)$. We introduce the notation

$$X = (\xi(t_1), \ldots \xi(t_n))$$

for the "observed" random variable. The probability measure for X is called φ_X. It can be derived from μ and due to the assumptions stated above, it is independent of α.

An example of a rather common type of A in prediction problems is the case where $\alpha(\tau)$ is defined by

$$\alpha(\tau) = c \qquad \tau \leq t$$
$$\alpha(\tau) = \gamma \qquad \tau > t$$

where c and γ are constants. c is a known number; A is generated by varying γ.

Returning to the general set-up, let Ω be the class of φ_X obtained by varying μ in Ω_α (for any arbitrary α since φ is independent of α). $\mathcal{X}$ is the space of all X.

Let

$$Y = \xi(\tau) \Big| \tau \in T - T_t$$

be the "future" random variable, $\mathcal{Y}$ the space of all Y. For any measurable $S \subset \mathcal{Y}$ let us consider the conditional distribution of S given that $X = x$.

$$\Pr\{Y \in S \mid X = x\}$$

This probability measure exists almost everywhere in $\mathcal{X}$ (Kolmogoroff [7, page 45-49]). We now make

Assumption 3. To any two μ in Ω_α the corresponding two φ_X in Ω are not identical, i.e. φ_X determines μ and therefore also $\Pr(Y \in S \mid X = x)$ uniquely.

This condition corresponds loosely, as we shall see later, to the condition about identifiability.

We shall see later that Assumption 3 could be substituted by a less restrictive assumption.

We introduce the notation

$$\varphi_Y(S \mid x, \varphi_X, \alpha) = \Pr(Y \in S \mid X = x).$$

The above set-up can, of course, be extended to the case where not all components of $\zeta(t_j)$, $j = 1, 2, \ldots n$ are components of X. It may, in some cases, be a statistical design of investigation to choose the points of time $t_1, \ldots, t_n$ and the components of $\zeta(t_j)$, $j = 1, \ldots, n$ such that Assumption 3 is satisfied, i.e. such that structures in the "the/model are identified." In trying to choose X we must of course make use of our a priori knowledge, which is given by the sets Ω_α of μ- s. If our a priori knowledge were only given by the set Ω of all φ_X, then this would be of no help in solving the design of statistical investigation problems (the identification problem), since φ_X is only given if we know what we want to observe. In that case, we can only state whether identification is present or not; we cannot insure that it is present. Of course, any statistical investigator who chooses his model in a sensible manner must at least have an implicit feeling of what his Ω_α's are.

Of course, design of statistical investigation, involves something more; we want to minimize the cost of the investigation. We shall not go into this here. However, it should be mentioned that the present set-up can be extended to include situations involving cost of investigation, and it can also be extended to include mixed sequential and design of statistical investigation situations.

§ 3.4. In §3.2-3.3 we have expressed explicitly "the dynamics of the situation." We shall now define the practical purpose of the statistical investigation. What do we want to predict?

We define a utility $V(Y, \alpha)$ which is a function from $\mathcal{Y} \times A$ to the real line, and for each $\alpha \in A$ is measurable with respect to Y. The expected future utility is now

$$EV = \int_{\mathcal{Y}} V(Y,q)\, d\, \varphi_Y(S \mid x, \varphi_X, \alpha)$$

and this is a function(al) of φ_X, α and x. With known φ_X and X we want to take the action α which maximizes the expected utility. The statistical material gives us x directly. It is a statistical inference problem to determine φ_X.

Note that this set-up is rather general. If we want to make the probability of the event $Y \in Q$, where Q is a measurable set in $\mathcal{Y}$, as large as possible, then we define

$$V(Y, \alpha) = 1 \qquad \text{if } Y \in Q$$

$$V(Y, \alpha) = 0 \qquad \text{otherwise.}$$

As another example, it may be mentioned that the case where $\zeta(\tau)$ is a scalar, $V(Y, \alpha) = (\zeta(t+1) - \gamma)^2$ (where γ is defined on page 8) and α does not actually occur in φ_Y. In that case we have a "passive point prediction problem." In some cases, it may be possible to give V in money units and in other cases a utility function, as it is defined in econometrics, may be constructed. However, questions about the definition and measurement of utility in this special sense, is not essential to our theory.

§ 3.5. By means of an observation x, we want to narrow the set of all probability measures φ_X in such a manner that the prediction problem can be solved. What exactly do we want to know about φ_X, how do we want to narrow the set of φ_X? More precisely, which class ϕ of subsets ω of Ω is relevant to the prediction problem.

Let the set of all $\alpha \in A$ for which EV is maximized for given φ_X and x be $\bar{A}(\varphi_X, x)$. That is $\alpha \in \bar{A}(\varphi_X, x)$, when and only when

$$\int_{\mathcal{Y}} V(Y, \alpha)\, d\, \varphi_Y(S \mid x, \varphi_X, \alpha) = \sup_{\alpha \in A} \int_{\mathcal{Y}} V(Y, \alpha)\, d\, \varphi_Y(S \mid x, \varphi_X, \alpha)$$

We now define the class ϕ in the following manner. It is the class of all sets $\omega \in \Omega$ such that

(i) $\bigsqcap_{\varphi_X \in \omega} \bar{A}(\varphi_X, x)$ is nonempty for all x

(ii) for any $\varphi_X^{(1)}$ belonging to Ω but not to ω

$$\bar{A}(\varphi_X^{(1)}, x) \bigsqcap_{\varphi_X \in \omega} \bar{A}(F_X, x) = 0 \text{ for some } x.$$

In other words, any ω is the set of all φ_x which, regardless of x, lead you to take the same action. There is no subset of Ω properly covering ω with the same property.

It is easily seen that no set in ϕ is a proper subset of another, and consequently, if ϕ contains Ω if 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 proper statistical inference problem.

We now make the assumption:

Assumption 4. $\sum_{\omega \in \phi} \omega = \Omega$, i.e. the sets in ϕ covers Ω. This implies, of course, that ϕ is nonempty.

This assumption is "very often" fulfilled.

§ 3.6. The final step is now to determine the weight function. The weight function is the loss in expected utility caused by assuming that $\varphi_x \in \omega$, whereas this is not true. If we make an error of this kind we take the wrong action.

The set

$$\varphi \Big|_{x \in \omega} \quad \bar{A}(\varphi_x, x) \text{ where } \omega \in \phi$$

contains for all x at least one element. In determining the utility we could have obtained, if we had taken the right action, we insert any of the elements of this set into the expected utility. In determining the utility if we make a wrong decision we have no guarantee that this utility is the same for any element of the set. In order to overcome this difficulty in determining the weight function we make the following assumption:

Assumption 5. For all $\omega \in \phi$ the set

$$\varphi \Big|_{x \in \omega} \bar{A}(\varphi_x, x)$$

contains just one element, which we denote $\alpha(x, \omega)$. This condition also seems to be fulfilled in "many cases." On intuition, it is a reasonable property of ϕ since by the definition of ϕ each ω is "on the verge of becoming empty."

Of course, all assumptions made are essentially assumptions on μ, V, and the random variable, X.

The loss in utility, if we say that the probability measure for x belongs to ω whereas φ_X is the true measure, is then,

$$W(\varphi_X, \omega, x) = \sup_{\alpha} \int_{\mathcal{Y}} V(Y, \alpha) \; d\varphi_Y(S \mid x, \varphi_X, \alpha)$$
$$- \int_{\mathcal{Y}} V(Y, \alpha[x, \omega]) \; d\varphi_Y(S \mid x, \varphi_X, \alpha[x, \omega])$$

In the case where μ is a serially independent process or - more generally - where X and Y are independent, W will not depend on x. But it is interesting to see that in the general case x will usually enter. This is due to the double purpose of the sample in a stochastic process. It tells us something about φ_X (statistical inference) and at the same time enters directly as a kind of initial condition for prediction. If we have a Markoff process, then W will only depend on $\zeta(t_n)$.

It is seen that if $\varphi_X \in \omega$, then W = 0. We now define the risk connected with any statistical decision function $\omega = d(x)$, as

$$r(\varphi_X, d) = EW(\varphi_X, d(X), X)$$

where d represents the functional form of d(x).

After having found a minimax solution $d^0(x)$ with respect to this risk function, the best action to take is

$$\alpha = \alpha^0[x] = \alpha(x, d^0(x)) .$$

§ 3.7. The method developed above corresponds to predicting by minimaxing the expected loss in expected future utility due to making wrong statistical decisions.

We could use a different principle, namely maximising with respect to the decision function the minimum with respect to φ_X of the expected future utility, expectation being taken over the space $\mathcal{Z}$ of all ζ. In the latter case we want to consider

$$\int_{\mathcal{Z}} \bar{V}(\zeta) \, d\mu$$

where

$$\bar{V}(\zeta) = V(Y, \alpha[X, d(X)])$$

for all ζ such that $\zeta(\tau)\big|_{\tau \in T - T_t} = Y$

and $(\xi(t_1) \ldots \xi(t_n)) = X$.

This integral is equal to

$$\int_{\mathcal{X}} \int_{\mathcal{Y}} V(Y, \alpha(X, d(X))) \, d\varphi_Y(S \mid X, \varphi_X, \alpha) \, d\varphi_X(T)$$

By Assumption 3 minimizing the expected future utility with respect to μ is the same as minimizing it with respect to φ_X. It is therefore seen that the prediction problem is reduced to the same statistical inference problem as before except that the weight function is now

$$W(\varphi_X, \omega, x) = - \int_{\mathcal{Y}} V(Y, \alpha(x, \omega)) \, d\varphi_Y(S \mid x, \varphi_X, \alpha).$$

Note that in this case W may be different from zero when $\varphi_X \in \omega$.

It is interesting to notice that if you want to find the prediction function $\alpha = \alpha^0[X]$ which maximins the future utility, then this is, except for the form of the weight function, the same as we have done above if Assumptions 4 and 5 are fulfilled. If Assumptions 4 and 5 are not fulfilled, then the prediction problem still makes sense as long as Assumption 3, about identifiability, is fulfilled. However, in that case, we have no guarantee that the prediction problem reduces to a problem of statistical inference in Wald's sense. It is surprising that Assumptions 4 and 5 are fulfilled in most practical cases and this makes it urgent to analyze these Assumptions further and express them in terms of μ, V and X.

In the end, it should be remarked that the minimax principle is not essential to the theory developed above. Other principles could be applied.

§ 4. Examples.

§ 4.1. In this section several examples will be given to illustrate the general set-up. Example 1 has no other purpose than to facilitate the comprehension of the rather abstract notions and notations used in the preceding section. Some of the other examples are meant to give hints to possible applications of the theory.

§ 4.2. Example 1. Let T be the set of all positive and negative integers and 0.

$$T = \{\ldots -2, -1, 0, 1, 2 \ldots\}$$

$\zeta = \zeta(\tau)\big|_{\tau \in T}$ is a serially independent process, i.e. for any $\tau_1 \ldots \tau_q$ belonging to T, $\zeta(\tau_1) \ldots \zeta(\tau_q)$ are stochastically independent. The probability distribution of $\zeta(\tau)$ is given by

$$\Pr[\zeta(\tau) = z] = \binom{\alpha(\tau)}{z} \pi^z (1-\pi)^{\alpha(\tau)-z}$$

for $z = 0, 1, 2, \ldots \alpha(\tau)$.

π can assume any value in the interval [0,1]. For any $\alpha = \alpha(\tau)\big|_{\tau \in T}$, π generates Ω_α by running through the interval [0, 1]. The space A consists of all functions $\alpha(\tau)$ of the form

$$\alpha(\tau) = 1 \qquad \tau \leq t,\ \tau \in T$$

$$\alpha(\tau) = \alpha \qquad \tau > t,\ \tau \in T$$

where α runs through the values 1, 2, ... ∞. (There is no confusion in using the same symbol for the functional form $\alpha = \alpha(\tau)\big|_{\tau \in T}$ and the values α since there is a one to one correspondence between these two entities. We shall therefore, for the sake of linguistic simplicity, pretend that A consists of the numbers 1, 2, 3, ... ∞).

We write further $\zeta(\tau) = Z_\tau$

We choose $V(Y, \alpha) = 1$ if $Z_{t+1} = Z_{t+2} = \ldots = Z_{t+\rho} = 1$

$V(Y, \alpha) = 0$ otherwise

ρ is known. We have observed

$$X = \{Z_{t-n+1}, Z_{t-n+2} \ldots, Z_t\}$$

The probability distribution for X is easily written down. The class of all possible distributions Ω is generated by varying π. (Since there is a one-to-one correspondence we shall, in analogy with what we did with A, redefine Ω as the set of all points in the

interval [0, 1]). We can interpret the situation in the following manner. We have made n throwings with a coin with unknown probability π of getting heads. We intend to make ρ sets of α throwings. We earn a large sum if, in each set, we make 1 head and $\alpha - 1$ tails, otherwise we lose. We have freedom to choose how many throwings α we can make in each set, but it must be the same for all sets. What is the best choice of α. We want to show that this problem can be reduced to a very definite way of estimating the probability π on the basis of the past n throwings. The expected future utility is

$$EV = (\alpha \pi (1 - \pi)^{\alpha - 1})^{\rho} .$$

For given π the best choice of α is the greatest integer $\leq \frac{1}{\pi}$ if $\frac{1}{\pi}$ is not an integer and $0 < \pi < 1$. If $\frac{1}{\pi}$ is an integer and $0 < \pi < 1$ then both $\frac{1}{\pi}$ and $\frac{1}{\pi} - 1$ are best. If $\pi = 0$ or 1 all choices of α are equally good.

We then have

$\bar{A}(\pi) = \{1, 2, 3 \ldots \infty\}$ if $\pi = 0$ or 1.

$\bar{A}(\pi) = \left\{\frac{1}{\pi} - 1, \frac{1}{\pi}\right\}$ if $\frac{1}{\pi}$ is an integer and $\pi < 1$.

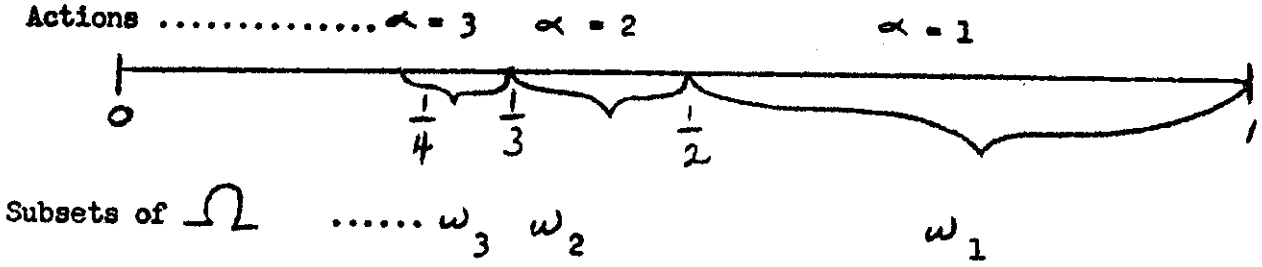
$\bar{A}(\pi) = \left[\frac{1}{\pi}\right]$ if $\frac{1}{\pi}$ is not an integer and $\pi > 0$.

([a] means the greatest integer $\leq a$)

At this point, it is worth mentioning what might be considered a very "natural" way of solving the problem. Take as an estimate π the relative frequence π^* of heads in the n throwings. This estimate is what most statisticians, I believe, would consider as the "best". (It is the estimate with the smallest variance among all unbiased estimates and it is a maximum likelihood estimate.) After having estimated π choose $\alpha = \left[\frac{1}{\pi^*}\right]$. This is a typical "two-step procedure." Is it the best?

In order to answer that question let us proceed as proposed in §3. Using the definition of ϕ we find that ϕ, in this case, consists of sets ω_ν, $\nu = 1, 2, 3 \ldots \infty$ where each ω_ν consists of all the points in the closed intervals $\left[\frac{1}{\nu + 1}, \frac{1}{\nu}\right]$ and the

points 0 and 1. Assumption 4 is fulfilled because $\sum_{\nu=1} \omega_\nu = [0, 1] = \Omega$. (Assumptions 1 - 3 are obviously fulfilled.) Assumption 5 is also fulfilled because $\bigcap_{\pi \in \omega} \bar{A}(\pi) = A = \{\nu + 1, \nu\} \cdot \{\nu\} \cdot \{\nu, \nu - 1\} = \{\nu\}$; i.e. the best action to take if $\pi \in \omega_\nu$ is $\alpha = \nu$.



The loss in utility which we suffer if we believe that $\pi \in \omega_\nu$ whereas π is the true probability is

$$W(\pi, \omega_\nu) = \left(\left[\frac{1}{\pi}\right]\pi(1-\pi)^{\left[\frac{1}{\pi}-1\right]}\right)^\rho - (\nu\pi(1-\pi)^{\nu-1})^\rho .$$

In other words, our prediction problem has now been reduced to a pure statistical inference problem about π on the basis of n observations. Φ consists of $\omega_1, \omega_2 \ldots$ and the weight function is $W(\pi, \omega_\nu)$.

§ 4.3. Example 2. This example is meant to illustrate how the case of "passive point prediction" is a special case of the set-up in § 3 and that consequently this case can be reduced. To a "pure" statistical inference problem in the Wald sense. It also illustrates that the weight function may depend on the sample point X.

Let T be the same as in Example 1, and let ε_τ be a serially independent process with $E(\varepsilon_\tau) = 0$ and $E(\varepsilon_\tau^2) = \sigma^2 < \infty$ for all $\tau \in T$ and let the distribution function for ε_τ be independent of τ. $\zeta(\tau)$ is a scalar and defined by

$$\zeta(\tau) + a_1 \zeta(\tau - 1) + \ldots + a_p \zeta(\tau - p) = \varepsilon_\tau$$

for all $\tau \in T$. $\Omega_\alpha = \Omega_0$, which is independent of α, is generated by varying $\theta = (a_1, \ldots a_p, \sigma)$ in the p + 1 dimensional space such that $\sigma > 0$. A and $\alpha(\tau)$ is the same as in example 1 except that α may assume any real value. At time t we want to predict according to the utility function

$$V(Y, \alpha) = -(\zeta(t+1) - \alpha)^2 .$$

The observable random variable is

$$X = (\zeta(t-n+1), \ldots, \zeta(t)) .$$

The expected future utility is then

$$EV = -\sigma^2 - [a_1 \zeta(t) + \ldots + a_p \zeta(t-p+1) + \alpha]^2 .$$

Again, for the sake of linguistic simplicity, we redefine Ω_α and Ω as the set of all θ. $\tilde{A}(\phi_X, X)$ consists of one element, namely

$$-[a_1 \zeta(t) + \ldots + a_p \zeta(t-p+1)] .$$

It is easily seen that each set ω of ϕ consists of all $\theta = (a_1 \ldots a_p, \sigma)$ for which $a = (a_1 \ldots a_p)$ is specified and σ runs through all values > 0. We are led to point estimation of $a = (a_1 \ldots a_p)$ regardless of σ. We have

$$\sup EV = \sigma^2 .$$

The weight function is then

$$W(a_1 a_1^* X) = [(a_1 - a_1^*)\zeta(t) + \ldots + (a_p - a_p^*)\zeta(t-p+1)]^2 .$$

Our problem has now been reduced to the following. On the basis of observations of $\zeta(\tau)$ for $\tau = t-n+1, \ldots t$ to find point estimates of the coefficients $a_1 \ldots a_p$ using the weight function $W(a, a^*, X)$. The predictor $f(X)$ for $\zeta(t+1)$ is then

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

where $a_1^* \ldots a_p^*$ are functions of X. This expression for $f(X)$ does not restrict the possible choices of the form of $f(X)$. Using the miminax principle to the point estimation problem defined above is therefore the same as the following "point prediction problem." Find the functional form f^0 which minimizes

$$\sup_\theta \int [\zeta(t+1) - f(X)]^2 \, d\mu .$$

Returning to the statistical inference problem, let us consider the case where $p = 1$. Then, on intuition, for any reasonable form of a_1^*, a_1^* and $\zeta(t)$ are "almost" independent for n large and the risk function is therefore

$$EW = E(a_1^* - a_1)^2 \int (t)^2 \sim \frac{\sigma^2}{1-a_1^2} E(a_1^* - a_1)^2.$$

§ 4.4. Example 3. Let T just contain two points of time, 0 and 1.

$$T = \{0, 1\}$$

$\int(0)$ and $\int(1)$ are independent. $\int(0)$ consists of p + q independent components each representing the amount of crop on different plots of equal size. The geographic location of the p first plots are randomly chosen among the p + q plots. Fertilizer I is applied to the p first plots, fertilizer II to the remaining plots; t = 0 is the point of time for prediction and $\int(0) = X$ is observed. The cumulative distribution function F_X for $\int(0) = X = (X_1 \ldots X_p, X_{p+1}, \ldots X_{p+q})$ is given by

$$F_X(x_1 \ldots x_{p+q}) = \prod_{i=1}^{p} G(x_i - \lambda_1) \prod_{i=p+1}^{p+q} G(x_i - \lambda_2)$$

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

$Y = \int(1)$ is a scalar and represents the amount of crop on a plot we plan to cultivate. The cumulative distribution function F_Y for Y is

$$F_Y(y) = \alpha H(y - \lambda_1) + (1 - \alpha) H(y - \lambda_2)$$

where H is a distribution function. $\alpha = 1$ if fertilizer I is used and $\alpha = 0$ if fertilizer II is used on the plot we plan to cultivate. $A = \{0, 1\}$.

Let $\mathcal{G}$ and $\mathcal{H}$ be two classes of univariate distribution functions. Ω_α is generated by letting (λ_1, λ_2) run through all pairs of real numbers and letting G run through the class $\mathcal{G}$, and H run through the class $\mathcal{H}$. Ω is generated by letting (λ_1, λ_2) run through all pairs of real numbers and letting G run through $\mathcal{G}$.

Both fertilizers are equally expensive and we want to maximize the crop. We can, therefore, let

$$V(Y, \alpha) = Y \ .$$

Which fertilizer shall we choose? Which is the optimum choice, $\alpha = 0$ or $\alpha = 1$? The expected future crop (utility) is

$$EV(Y, \alpha) = \alpha \lambda_1 + (1-\alpha) \lambda_2 + \int z \, d H(z).$$

We then get

$$\bar{A}(F_X) = \{1\} \qquad \text{if } \lambda_1 > \lambda_2$$

$$\bar{A}(F_X) = \{0, 1\} = A \qquad \text{if } \lambda_1 = \lambda_2$$

$$A(F_X) = \{0\} \qquad \text{if } \lambda_1 < \lambda_2.$$

It is now seen that ϕ consists of two sets ω_1 and ω_o.

$$\omega_1 = [F_X \mid \lambda_1 \geq \lambda_2], \qquad \omega_o = [F_X \mid \lambda_1 \leq \lambda_2].$$

We now find

$$W(F_X, \omega) = \begin{cases} 0 & \text{if } \omega = \omega_1 \text{ and } \lambda_1 \geq \lambda_2 \\ \lambda_1 - \lambda_2 & \text{if } \omega = \omega_o \text{ and } \lambda_1 \geq \lambda_2 \\ \lambda_2 - \lambda_1 & \text{if } \omega = \omega_1 \text{ and } \lambda_1 \leq \lambda_2 \\ 0 & \text{if } \omega = \omega_o \text{ and } \lambda_1 \leq \lambda_2 \end{cases}$$

Let $\omega = d(X)$ be the statistical decision function. Let R be the set of all X for which $d(X) = \omega_1$, i.e. we accept ω_1 if $X \in R$. Then

$$W(F_X, d(X)) = \begin{cases} 0 & \text{if } X \in R \text{ and } \lambda_1 \geq \lambda_2 \\ \lambda_1 - \lambda_2 & \text{if } X \notin R \text{ and } \lambda_1 \geq \lambda_2 \\ \lambda_2 - \lambda_1 & \text{if } X \in R \text{ and } \lambda_1 \leq \lambda_2 \\ 0 & \text{if } X \notin R \text{ and } \lambda_1 \leq \lambda_2 \end{cases}$$

Let p(F) be the power function for the test

$$p(F) = \Pr\{X \in R\}$$

The risk function is now

$$r(F_X, R) = EW = \begin{cases} (1 - p(F))(\lambda_1 - \lambda_2) & \text{if } \lambda_1 \geq \lambda_2 \\ p(R)\,(\lambda_2 - \lambda_1) & \text{if } \lambda_1 \leq \lambda_2 \end{cases}$$

We are then led to the following statistical inference problem.

<u>We want to test</u> $\lambda_1 \geq \lambda_2$ <u>against</u> $\lambda_1 \leq \lambda_2$, <u>i.e. one composite hypothesis against</u>

another. The weight function is $W(F_X, \omega)$.

It is seen from the risk function that, loosely speaking, we want the first and second kind of error small.

§ 4.5. Example 4. This is an example from the field of econometrics. It illustrates how, in a certain situation, if a monopolist is a profit maximizer and if he wants to minimax the loss in profit due to incomplete information about the "dynamics of the situation," then this leads him to be interested in estimating the elasticity of demand, and the manner in which this elasticity should be estimated is uniquely determined. This example also illustrates how Assumption 3 is essentially an assumption about identifiability.

Let $T = \{t_1, t_2, \ldots, t_n = t, t+1\}$.

Let further $\mathcal{E} = \mathcal{E}(\tau) \mid \tau \in T$ be a serially independent process where $\mathcal{E}(\tau)$ has two components $\mathcal{E}_{1\tau}$ and $\mathcal{E}_{2\tau}$ and where $E\,\mathcal{E}_{i\tau} = 0$ for $i = 1, 2$. The distribution function of $\mathcal{E}(\tau)$ is independent of τ and all unknown parameters mentioned below. Let the class of all distribution functions for $\mathcal{E}(\tau)$ be $\mathcal{E}$. Let $\zeta(\tau) = (z_{1\tau}, z_{2\tau})$ have a distribution function defined by the relations

$$\mathcal{E}_{1\tau} = z_{1\tau} - bz_{2\tau} - c$$

$$[\mathcal{E}_{2\tau} - z_{1\tau} + \beta(z_{2\tau} - \alpha_{1\tau}) + \gamma]\alpha_{2\tau} + (1 - \alpha_{2\tau})(\alpha_{1\tau} - z_{2\tau}) = 0$$

for all $\tau \in T$.

$\alpha_{2\tau} = 1$ for $\tau \in T_t$ and 0 otherwise.

Let $t_1, t_2, \ldots t_n$ (all belonging to T_t) be the points of time when $\zeta(\tau)$ is observed. $\alpha_{1\tau}$ for $\tau = t_1, t_2 \ldots t_n$ are known numbers.

$\alpha_{1\tau} = \alpha$ for $\tau = t + 1$.

$\alpha(\tau) = (\alpha_{1\tau}, \alpha_{2\tau})$. A is generated by letting α run through all positive numbers. Ω_α is generated by letting (b, c, β, γ) attain all values in the four-dimensional space for which $\beta < -1$ and the distribution function of $\mathcal{E}(\tau)$ run through $\mathcal{E}$. Ω is generated in the same manner.

$z_{1\tau}$ and $z_{2\tau}$ can be interpreted as the logarithm of quantity sold and price obtained for a commodity at time τ. $\alpha_{1\tau}$ for $\tau \leq t$ is $-\log(1-\rho_{\tau})$ where ρ_{τ} is a sales tax imposed at time τ on the commodity. For $\tau \leq t$ we have the free market situation

$$\varepsilon_{1\tau} = z_{1\tau} - b\, z_{2\tau} - c$$
$$\varepsilon_{2\tau} = z_{1\tau} - \beta(z_{2\tau} - \alpha_{1\tau}) + \gamma$$

At time $\tau = t+1$, $\alpha_{1\tau}$ a price fixation. We have a monopoly situation,

$$\varepsilon_{1\tau} = z_{1\tau} - b\alpha - c.$$

On the basis of this knowledge about $\varphi(\tau)$ and observations of $(z_{1\tau}, z_{2\tau})$, $\tau = t_1 \ldots t_n$ the monopolist wants to know what price α to fix when he is a profit maximizer and wants to minimize his loss in profit. Let the cost of producing a quantity Q be AQ + B where A and B are known. The utility in this case is the profit, i.e.

$$V(Y, \alpha) = e^{z_{1t+1}}(e^{\alpha} - A) - B.$$

The expected future utility is then

$$EV = e^{b\alpha + c}(e^{\alpha} - A)\, E(e^{\varepsilon_{1\tau}}) - B.$$

We then get

$$\bar{A}(\varphi_X) = \left\{\log A\, \frac{b}{b+1}\right\}$$

Since there is a one-to-one correspondence between b and $\log A \frac{b}{b+1}$, each ω consists of all φ_X for which b is specified. That means that what we want is a point estimate of b and we are not interested in the other parameters. Let b^* be an estimate of b. The expected loss in profit due to estimating b to be b^* is

$$A^{b+1}\left[\left(\frac{b}{b+1}\right)^b\left(-\frac{1}{b+1}\right) - \left(\frac{b^*}{b^*+1}\right)^b\left(-\frac{1}{b^*+1}\right)\right]e^c\, E(e^{\varepsilon_{1\tau}}).$$

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

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

The model for (Z_{1t_i}, Z_{2t_i}), $i = 1, 2 \ldots n$ is identified when and only when the following relation is <u>not</u> fulfilled.

$$\alpha_{1t_1} = \alpha_{1t_2} = \ldots = \alpha_{1t_n}.$$

If this relation <u>is</u> fulfilled, then there exist, as is well known, different pairs of values of (b, c) leading to the same distribution function for (Z_{1t_i}, Z_{2t_i}), $i = 1,2\ldots n$. Let (b_1, c_1) and (b_2, c_2) be two such pairs of values. Then the distribution function for Z_{1t+1} is given by the stochastic equation

$$\varepsilon_{1t+1} = Z_{1t+1} - b_i\alpha - c_i$$

which gives a different distribution function for $i = 1$ and $i = 2$. With the same probability distribution for $(Z_{1t_i}\ Z_{2t_i})$ $i = 1, 2, \ldots, n$, there are therefore several possible probability measures for ξ. Assumption 3 is not fulfilled. This shows how the assumption about identifiability and Assumption 3 are linked together and consequently, how identifiability is linked to the problem of prediction.

§5. The Role of the Identification Problem in the General Set-up

<u>§5.1</u>. In this section, we shall demonstrate how, in a special case, Assumption 3 is the same as asserting that the structure explaining the observable random variable X is identified.

<u>§5.2</u>. Suppose that the elements of Ω_α are in a one-to-one correspondence to the elements θ of a subset in the Euclidian space. Then μ can be written $\mu(S; \alpha_1\ \theta)$. Since to each element of Ω_α, there corresponds a probability measure φ_X, we may write $\varphi_X(S \mid \theta)$, where, due to Assumptions 1 and 2, the dormain of θ is independent of α.

Suppose now that Assumption 3 is fulfilled. Then θ must be identified. For suppose that θ_1 and θ_2 were such that $\varphi_X(S \mid \theta_1) = \varphi_X(S \mid \theta_2)$ for all S, then corresponding to

this probability measure for X there corresponds two measures for φ, $\mu(S \mid \alpha_1 \theta_1)$ and $\mu(S \mid \alpha_1 \theta_2)$, contrary to Assumption 3. If, on the other hand, θ is identified, then to each φ_X there corresponds just one θ and to each θ of course, just one μ, i.e. Assumption 3 is fulfilled.

<u>If Ω_α form a parametric set of probability measures with parameter θ, then a necessary and sufficient condition for Assumption 3 to be fulfilled is that all components of θ are identified in the model Ω for the observable random variable X.</u>

This formally links the problem of identification to the problem of prediction.

<u>§ 5.3</u>. Assumption 3 leads you to complete identification which is too restrictive for many purposes. In Example 3, for instance, Assumption 3 is not fulfilled and the structure is not completely identified. The only requirement about the process we need for prediction in that example, is identification of λ_1 and λ_2.

In the general case the following assumption could substitute Assumption 3.

<u>Assumption 3'</u>. Let us divide Ω in equivalence classes. Any two probability measures $\varphi_X^{(1)}$ and $\varphi_X^{(2)}$ are equivalent if for any x

$$\bar{A}(\varphi_X^{(1)}, x) = \bar{A}(\varphi_X^{(2)}, x).$$

We now require that to any equivalence class of φ_X, there corresponds just one μ.

In Example 3 , this assumption leads to the requirement that (λ_1, λ_2) is identified and in Example 4, it leads to the requirement that the elasticity of demand b is identified.

APPENDIX

Notations

$x \in A$ means, the element x belongs to the set A.

$\sup_{x \in A} f(x)$ denotes the least upper bound and $\inf_{x \in A} f(x)$ denotes the greatest lower bound of all values of f(x) corresponding to $x \in A$.

$(x_1, x_2, \ldots x_p)$ denotes a vector with components $x_1, x_2, \ldots x_p$.

$S \times R$ is the set of all pairs of elements (X, y) such that $X \in S$ and $y \in R$.

$f(x)\Big|_{x \in A}$ denotes the functional form of the function f(x) defined in the dormain A.

$[x \mid r_x]$ is the set of all x for which the restriction r_x is fulfilled.

$S - R$ is the set of all elements belonging to S but not to R.

$\Pr\{r_x\}$ is the probability measure for the set $[x \mid r_x]$.

$\Pr\{r_x \mid Y = y]$ and $E(X \mid Y = y)$ is the conditional probability of r_x and the expected value of X, given that $Y = y$.

$A \cdot B$ denotes the intersection (product set) of A and B.

$A + B$ is the union of A and B.

$\prod_i A_i$ is the intersection of $A_1, A_2, \ldots$.

$\sum_i A_i$ is the union of $A_1, A_2, \ldots$.

$\prod_{r_A} A$ and $\sum_{r_A} A$ denotes the intersection and the union of all sets A fulfilling a relation r_Z.

$A = 0$, where A is a set, means that A is empty (void).

$\{x_1, x_2 \ldots\}$ denotes the set containing $x_1, x_2 \ldots$.

$[a, b]$ is the set $[x \mid a \leq x \leq b]$.

$[a]$ is the greatest integer $\leq a$.

$x \notin A$ denotes x does not belong to A.

REFERENCES

[1] Trygve Haavelmo: "The Probability Approach in Econometrics," Econometrica, Vol. 12 (1944), Supplement.

[2] Mann and Wald:

[3] Lundberg's Theory of Risk, Reference not available in library.

[4]. A. Wald: "Statistical Decision Functions," Annals of Mathematical Statistics, Vol. XX, No. 2 (1949).

[5]. Norbert Wiener: "Extrapolation, Interpolation and Smoothing of Stationary Time Series," New York and Lond, 1949.

[6]. Monograph 10. To be published by Cowles Commission for Research in Economics.

[7]. A. Kolmogoroff: "Grumdbegriffe der Wahrscheinlichkeitsrecknung." Berlin (1933).

ERRATA TO COWLES COMMISSION DISCUSSION PAPER: STATISTICS NO. 344

page 2, line 1 from above: "[2, page]" read "[2, page 173]"

page 6, line 15 from below: "if" read "of."

page 7, line 10 from above: "functions of may" read "functions of 2 may"

page 7, line 11 from below: " is such that" read "Ω_α is such that."

page 9, line 1 from below: "C" read "α"

page 10, line 1 from below: "F_x" read "φ_x"

page 16, line 10 from below: "... reduced. To ..." read "...reduced to ..."

page 17, line 11 from below: " = t - n + 1" read "τ = t - n + 1"

page 18, line 1 from above: "$\sim$" read "$\approx$" (approximately equal to)

page 21, line 7 from above: "$\alpha_1 \tau$ a" read "$\alpha_1 \tau$ is a"

page 22, line 5 from below: "$\mu(S; \alpha_1 \theta)$" read "$\mu(S, \alpha, \theta)$."

page 22, line 3 from below: "dormain" read "domain."

page 23, line 1 and 2 from above: "$\mu(S \mid \alpha_1 \theta_1)$ and $\mu(S \mid \alpha_1 \theta_2)$" read "$\mu(S; \alpha, \theta_1)$ and $\mu(S; \alpha_1 \theta_2)$."

page 23, Assumption 3' should be: "Let us divide $\sum_\alpha \Omega_\alpha$ in equivalence classes. Any two probability measures $\mu^{(1)}$ and $\mu^{(2)}$ are equivalent if for any x

$$\tilde{A}(\mu^{(1)}, x) = \tilde{A}(\mu^{(2)}, x)$$

where A $(\mu^{(i)}, x)$ is the set of best actions corresponding to $\mu^{(i)}$ and x. We now require that all μ which lead to the same φ_x are equivalent."

Appendix, line 8 from above: "dormain" read "domain."

Appendix, line 6 from below: "r_x" read "r_A".

Reference, line 4 from above insert: "On the statistical treatment of linear stochastic difference equations." Econometrica, Volume 11 (1943).