BULL. AUSTRAL. MATH. SOC. 62A99, 62P20, 62P25, 82A05
VOL. 9 (1973), 391-410.

Macroscopic descriptions of microscopic phenomena

Peter D. Finch

Some problems in the behavioural and physical sciences arise in the context of an incomplete knowledge of the fine detail of underlying practical situations. This paper presents a general mathematical framework for the discussion of such problems. This framework provides an algebraic language for the discussion of ecological analysis in the social sciences, aggregation in economics and macroscopic descriptions in statistical physics. Here, however, only the mathematical framework is presented; detailed applications will be presented elsewhere.

1. Introduction

Let $x_1, x_2, \ldots, x_n$ be sample values of n independent random variables $X_1, X_2, \ldots, X_n$ which have a common but unknown distribution. A standard problem of statistical inference concerns the description of that unknown distribution on the basis of the information provided by the sample values. Here we consider the more general problem which arises when one wants to describe the unknown distribution on the basis of the information provided by the value $\xi(x_1, x_2, \ldots, x_n)$ of some function of sample values rather than by the whole sample itself. In the language of the title of this paper the whole sample $(x_1, x_2, \ldots, x_n)$ constitutes the microscopic data whereas the value $\xi(x_1, x_2, \ldots, x_n)$ constitutes the macroscopic or ecological summary of those data. Practical situations

Received 2 July 1973.

involving such summaries arise in a number of contexts and two simple examples will suffice to motivate the development of the general mathematical framework.

Our first example goes back to Robinson [3], an early but important paper in the development of ecological analysis. Robinson considered the extent to which correlation between colour and literacy in the United States was affected by grouping the data into regional zones. For our purposes the mathematical aspects of immediate interest can be formulated in the following way. Let Ω be a finite population of people which is divided into disjoint geographical regions $A_1, A_2, \ldots, A_k$. For each person ω in Ω let $X(\omega) = (A(\omega), L(\omega), R(\omega))$ where $A(\omega)$ is the region to which he belongs, $L(\omega)$ is 0 or 1 according as he is or is not literate and $R(\omega)$ is 0 or 1 according as he is White or Negro. Let $A = \{A_1, A_2, \ldots, A_k\}$, $\underline{2} = \{0, 1\}$ and write X for the cartesian product $A \times \underline{2} \times \underline{2}$. Let Ω_n be the n-fold cartesian product of Ω with itself and for each $j = 1, 2, \ldots, n$ define $X_j : \Omega_n \to X$ by the equations

$$X_j(\omega_1, \omega_2, \ldots, \omega_n) = X(\omega_j) .$$

Suppose that one takes an ordered random sample of size n with replacement from the population Ω. Under such a sampling procedure $X_1, X_2, \ldots, X_n$ are independent random variables with the same distribution, namely the relative frequency distribution of X over Ω. If $x_1, x_2, \ldots, x_n$ are the respective sample values then one can seek to make inferences from them about the unknown distribution of X, that is the joint relative frequency distribution of A, L and R. Suppose, however, that, either in principle or for reasons of economy, the whole sample is not available but one knows only that summary of it which gives, for each of the geographical regions in question, the number of people in the sample who belong to that region together with the number of those who are illiterate and the number who are Negro. In other words if $x_j = (a_j, l_j, r_j)$ in X, $j = 1, 2, \ldots, n$, are the sample values then the summary in question replaces the n-tuple $(x_1, x_2, \ldots, x_n)$ of triples by a k-tuple of triples, namely

$$\xi(x_1, x_2, \ldots, x_n) = (\phi_1(x_1, x_2, \ldots, x_n), \ldots, \phi_k(x_1, x_2, \ldots, x_n))$$

where, for each $i = 1, 2, \ldots, k$,

$$\phi_i(x_1, x_2, \ldots, x_n) = \left(\sum_{j=1}^{n} \delta_i(a_j), \sum_{j=1}^{n} \delta_i(a_j)l_j, \sum_{j=1}^{n} \delta_i(a_j)r_j\right)$$

and for each a in A , $\delta_i(a)$ is 1 or 0 according as a is or is not A_i . Our basic problem concerns the extent to which one can describe the unknown distribution of X in terms of the information provided by the summary $\xi(x_1, x_2, \ldots, x_n)$ rather than in terms of that provided by the whole sample $(x_1, x_2, \ldots, x_n)$.

Our second example concerns aggregation in economics as treated, for instance, by Theil [4]. With each member ω of a finite set Ω there are associated real-valued microquantities $B(\omega)$, $A_1(\omega), \ldots, A_m(\omega)$ so that one has $m + 1$ real valued functions $B : \Omega \to \mathrm{R}$ and $A_k : \Omega \to \mathrm{R}$, $k = 1, 2, \ldots, m$. One's interest is in the joint relative frequency distribution of B and $A_1, A_2, \ldots, A_m$ over Ω , in particular one is interested in the way B depends on $A_1, A_2, \ldots, A_m$. In economics B is called the endogenous microvariable, $A_1, A_2, \ldots, A_m$ are called the exogenous microvariables and it is not unusual to represent the sought for dependence by the microequations

$$B(\omega) = \sum_{k=1}^{m} B_k(\omega)A_k(\omega) + U(\omega) , \quad \omega \in \Omega ,$$

where $U(\omega)$ is a disturbance term which characterises the departure from linearity. However for our present purposes we may ignore this particular type of functional dependence. Write

$$X(\omega) = (B(\omega), A_1(\omega), \ldots, A_m(\omega))$$

and let X be the $(m+1)$-fold Cartesian product of R with itself so that $X : \Omega \to$ X . As in the last example we can, by means of a suitable sampling procedure, introduce independent random variables $X_1, X_2, \ldots, X_n$ each of them having the distribution of X over Ω . As before we can pose the standard inference problem about the distribution of X over Ω when one knows the sample values $x_1, x_2, \ldots, x_n$ but our present interest is in that aspect of the aggregation problem in economics which concerns the extent to which one can describe the distribution of X over Ω, in other words the joint distribution of the microvariables, in terms of certain aggregate values derived from the sample. Thus suppose that, for each $j = 1, 2, \ldots, n$, $x_j = (b_j, a_{j1}, \ldots, a_{jm})$ in X are the sample values and that the set $(x_1, x_2, \ldots, x_n)$ consisting of the n sample $(m+1)$-tuples is summarised by the single $(m+1)$-tuple

$$\xi(x_1, x_2, \ldots, x_n) = \left(\sum_{j=1}^{n} b_j, \sum_{j=1}^{n} a_{j1}, \ldots, \sum_{j=1}^{n} a_{jm}\right) .$$

Once again our basic problem concerns the extent to which one can describe the distribution of X in terms of the information provided by the summary $\xi(x_1, x_2, \ldots, x_n)$ rather than in terms of that provided by the whole sample $(x_1, x_2, \ldots, x_n)$.

We emphasise that the two examples are only introduced to provide motivation; more detailed studies of these and related problems will be published elsewhere. It is, however, worthwhile pointing out that in practice one is sometimes dealing with the whole population rather than with a random sample and that even if one does have a sample it may or may not have been taken randomly and it may well have been taken without replacement rather than with replacement. Although it is possible to deal with such situations by arguments like those used below it is, for our present purposes, more convenient to regard the mathematical framework as providing a conceptual model in terms of which such situations may be discussed. In terms of that model one asks how one would describe the distribution of some vector-valued quantity X over the finite population Ω if it were the case that all one could know was a certain summary of a random sample. With this model in mind, and to avoid inessential mathematical complexity, we restrict our discussion to random variables taking on only a finite number of values.

In the next section we introduce the concept of a summary function in an abstract way and derive some results for later use.

2. Summary functions

For any set T we write T_n for the n-fold Cartesian product of T with itself and put

$$T_* = \bigcup_{n=1}^{\infty} T_n .$$

Let X be a non-empty set. A subset S of X_* will be said to be exact when $S = T_*$ for some non-empty subset T of X. Let M be a non-empty set. An M-valued summary function in X is defined to be a function $\xi : X_* \to M$ which has an exact domain. A summary function is said to be universal when its domain is X_*. It is necessary to distinguish between a universal summary function ξ and $\xi|S$, the restriction of ξ to an exact subset of X_*. Of course $\xi|S$ is a summary function which agrees with ξ on its domain, but its domain is different from that of ξ. If η is a summary function and ξ is a universal summary function such that $\xi|\text{dom}\eta$ is η we say that ξ is a universal extension of η; even when such an extension exists it may not be uniquely determined by η.

Let $\xi : X_* \to M$ be an M-valued summary function in X; for each positive integer n there is an induced mapping from X_n to M; namely $\xi_n : X_n \to M$ where $\xi_n = \xi|X_n$ is the restriction of ξ to X_n. Thus a summary function is a compact way of talking about a particular algebraic structure on X, for ξ_n can be thought of as an M-valued partial n-ary operation in X, there being one such operation for each positive arity. If ξ is universal then these operations are everywhere defined on X and if, in addition, $M = X$ the universal summary function determines a particular kind of universal algebra carried by the set X. Indeed with the algebraic interpretation in mind we often write $\xi_n x_1 x_2 \ldots x_n$ instead of $\xi_n(x_1, x_2, \ldots, x_n)$ or $\xi(x_1, x_2, \ldots, x_n)$ whenever $(x_1, x_2, \ldots, x_n)$ is in the domain of ξ_n. Conversely if, for each positive integer n, ξ_n is an M-valued partial n-ary operation in X we can define $\xi : X_* \to M$ by decreeing that $\xi|X_n = \xi_n$ and if ξ so defined does have an exact domain it is a summary function in the sense defined above. Indeed summary functions often arise in this way in

https://doi.org/10.1017/S0004972700043409 Published online by Cambridge University Press

practice.

With the algebraic interpretation in mind we define a character of the M-valued summary function ξ to be a complex-valued function χ on M such that

(2.1) $$\text{codom}\xi \subseteq \text{dom}\chi \subseteq M$$

and

(2.2) $$\chi(\xi_n x_1 x_2 \ldots x_n) = \chi(\xi_1 x_1)\chi(\xi_1 x_2) \ldots \chi(\xi_1 x_n)$$

for each positive integer n and each $(x_1, x_2, \ldots, x_n)$ in the domain of ξ_n. Characters always exist, thus $\chi(m) \equiv 0$ on M and $\chi(m) \equiv 1$ on M are instances of characters, the former of these is called the trivial character and the latter is called the unit character. If ξ is universal than (2.2) holds for any $x_1, x_2, \ldots, x_n$ in X. If η is a universal summary function, λ is a character of η and S is exact then $\xi = \eta|S$ is a summary function and $\lambda|\text{codom}\xi$ is a character of ξ, but a general character χ of ξ is not necessarily the restriction of a character of η because it is only required that (2.2) hold for $(x_1, x_2, \ldots, x_n)$ in the domain of ξ and not for all $(x_1, x_2, \ldots, x_n)$ in X_*.

If ξ is an M-valued summary function and $\phi : M \to M$ has a domain which contains the codomain of ξ then $\phi \circ \xi$ is also an M-valued summary function. Moreover if ϕ is a bijection on its domain then the characters of $\phi \circ \xi$ are of the form $\chi \circ \phi^{-1}$ where χ is a character of ξ.

Let ξ be an M-valued summary function in X and let K be a non-empty finite subset of X such that K_* is contained in the domain of ξ. Let χ be a character of ξ and write

(2.3) $$K_\chi = \sum_{x \in K} \chi(\xi_1 x) .$$

A non-negative character χ is said to be normed on K when $K_\chi = 1$. If χ is normed on K then

$$\sum_{x_* \in K_n} \chi(\xi(x_*)) = 1 ,$$

moreover this equation holds for all $n \geq 1$ only if χ is normed.

For each x_* in X_* let $n(x_*)$ denote the n for which $x_* = (x_1, x_2, \ldots, x_n)$ belongs to X_n. We say that ξ is separative on K_* when

$$\xi(x_*) = \xi(y_*) \Rightarrow n(x_*) = n(y_*)$$

for all x_*, y_* in K_*. When this is so we can attach to each m in the codomain of $\xi|K_*$ the common arity of the x_* in K_* such that $\xi(x_*) = m$; let this common arity be denoted by $\nu(m)$ so that

$$\nu(\xi_n x_1 x_2 \ldots x_n) = n , \quad (x_1, x_2, \ldots, x_n) \in K_* .$$

Suppose that ξ is separative on K_* and let χ be a non-trivial non-negative character of $\xi|K_*$. Define $\chi' : M \to \mathrm{R}^+$ with a domain which is the codomain of $\xi|K_*$ by decreeing that χ' is given on its domain by the equation

$$\chi'(m) = K_\chi^{-\nu(m)}\chi(m) .$$

Then one verifies easily that χ' is a non-negative character of $\xi|K_*$ which is normed on K. In the case χ is itself normed on K one has, of course, that χ' is the restriction of χ to the codomain of $\xi|K_*$; thus in the case of separative summary functions all the normed characters can be obtained by this normalisation procedure.

Note that if $\xi : X_* \to M$ is not separative then it may be replaced by the separative summary function $\xi' : X_* \to Z^+ \times M$ defined on its domain, which is the same as that of ξ, by the equations

$$\xi'(x_*) = (n(x_*), \xi(x_*)) .$$

A particularly important special case occurs when ξ is an X-valued universal summary function with the properties

(i) $\xi_1 x = x$ for each x in X, and

(ii) for each integer $n \geq 2$ and any $x_1, x_2, \ldots, x_{n+1}$ in X,

$$\xi_2 x_1 (\xi_n x_2 \ldots x_{n+1}) = \xi_{n+1} x_1 x_2 \ldots x_{n+1} = \xi_2 (\xi_n x_1 \ldots x_n) x_{n+1} .$$

https://doi.org/10.1017/S0004972700043409 Published online by Cambridge University Press

Then ξ_2 is a semigroup operation on X, and writing it multiplicatively we find that

$$\xi_n x_1 x_2 \ldots x_n = x_1 x_2 \ldots x_n ,$$

where the expression on the right is the semigroup product of $x_1, x_2, \ldots, x_n$. The characters of ξ are just the complex-valued functions χ on X which have the property

$$\chi(xy) = \chi(x)\chi(y)$$

for any x and y in X; in other words they are characters of the semigroup X.

3. Surrogate probabilities

Let X be a non-empty set and let X be a random variable which takes on only a finite number of values in X with non-zero probability. The probability distribution of X is a function $P : X \to R^+$ such that

$$D = \{x : x \in X \ \& \ P(x) \neq 0\}$$

is finite and

$$\sum_X P(x) = 1 .$$

More generally, for each positive integer n, let $X_1, X_2, \ldots, X_n$ be a finite sequence of random variables each of which takes on only a finite number of values in X with non-zero probability. The joint probability distribution of $X_1, X_2, \ldots, X_n$ is a function $P_n = X_n \to R^+$ such that

$$D_n = \{(x_1, x_2, \ldots, x_n) : P_n(x_1, x_2, \ldots, x_n) \neq 0\}$$

is finite and

$$\sum_{X_n} P_n(x_1, x_2, \ldots, x_n) = 1 .$$

In what follows we will suppose that for each $n \geq 1$ the random variables $X_1, X_2, \ldots, X_n$ are mutually independent and that each of them has the distribution of X. Then D_n is the n-fold cartesian product of D and

https://doi.org/10.1017/S0004972700043409 Published online by Cambridge University Press

$$\forall(x_1, x_2, \ldots, x_n) \in D_n : P_n(x_1, x_2, \ldots, x_n) = \prod_{j=1}^{n} P(x_j) .$$

Since we wish to consider finite sequences $(X_1, X_2, \ldots, X_n)$ for arbitrary n it is convenient to define $P_* : X_* \to R^+$ by decreeing that $P_*|X_n = P_n$ as defined above. Then

$$D_* = \{x_* : x_* \in X_* \ \& \ P_*(x_*) \neq 0\}$$

is an exact subset of X_*, for each $n \geq 1$,

$$\sum_{x_* \in D_n} P_*(x_*) = 1$$

and, because of independence,

$$\forall x_* \in X_* : P_*(x_*) = \prod_{j=1}^{n(x_*)} P(x_j) .$$

A standard problem of statistical inference is that of "estimating" the function P on the basis of particular sample values $x_* = (x_1, x_2, \ldots, x_n)$. As indicated in the introduction we are interested in the more general problem of "estimating" P on the basis of the value $\xi(x_*)$ of some function ξ of sample values rather than on the basis of those sample values themselves. However the use of the term "estimating" raises controversial questions concerning "best" estimation procedures which we wish to avoid. To do so we remark that the practical problem is simply that we do not know the function P and so we are unable to calculate the functions P_n, $n \geq 1$; in other words we do not know the function P_*. However in the absence of this knowledge we want to use a surrogate for the function P_* so that, for each $n \geq 1$, we can calculate a surrogate probability of obtaining sample values $x_1, x_2, \ldots, x_n$ in a realisation of the n random variables $X_1, X_2, \ldots, X_n$. By use of the words "surrogate probability" rather than "estimated probability" we wish to emphasise the deputizing role an estimate of a probability distribution is required to play and, at this stage of our investigation, to pay less attention to the more controversial questions which arise when one asks the extent to which one surrogate is

https://doi.org/10.1017/S0004972700043409 Published online by Cambridge University Press

"better" than another in respect of the way it does play that role.

Motivated by the preceding considerations we define a surrogate function for P_* to be any function $Q_* : X_* \to \mathrm{R}^+$ which has domain X_* and is such that

(3.1) $$x_* \notin D_* \Rightarrow Q_*(x_*) = 0 \ ,$$

(3.2) $$\forall n \geq 1 : \sum_{x_* \in D_n} Q_*(x_*) = 1 \ ,$$

and

(3.3) $$\forall x_* \in X_* : Q_*(x_*) = \prod_{j=1}^{n(x_*)} Q_*(x_j) \ ,$$

where $n(x_*)$ is the n for which $x_* = (x_1, x_2, \ldots, x_n)$ is in X_n .

We say that $Q_*(x_*)$ is the surrogate probability for $P_*(x_*)$ and that, for each $n \geq 1$, $Q_n = Q_*|X_n$ is the surrogate distribution for P_n . Note that surrogate probabilities, like the probabilities for which they deputize, are non-negative quantities. The condition (3.1) ensures that sample values x_* which occur with zero probability are assigned zero surrogate probability, whereas condition (3.2) ensures that Q_n like P_n , for which it deputizes, sums to unity over D_n . Finally, condition (3.3) asserts "surrogate independence", namely that the surrogate joint probabilities $Q_n(x_1, x_2, \ldots, x_n)$ are to be calculated from the individual surrogate probabilities $Q_1(x_1), Q_1(x_2), \ldots, Q_1(x_n)$ in accordance with the assumed independence of the random variables $X_1, X_2, \ldots, X_n$. Note that if $Q : X \to \mathrm{R}^+$ is any probability distribution on X such that $Q(x) = 0$ for x not in D and we define $Q_1 = Q$ and Q_* by (3.3) then (3.1) and (3.2) are satisfied; conversely all surrogates Q_* arise in this way. Thus our definition of a surrogate requires no more than that a surrogate function Q_* arises in that way from some such probability distribution Q on X .

Suppose now that we wish to determine surrogate functions Q_* which take into account the fact that all we know about any set of sample values

https://doi.org/10.1017/S0004972700043409 Published online by Cambridge University Press

x_* is $\xi(x_*)$ where ξ is some summary function in X. More precisely let $\xi : X_* \to M$ be an M-valued summary function in X whose domain contains D_* and which is separative on D_*. To take account of the summary function ξ we observe that if x_* and x'_* are two sets of sample values in D_* for which $\xi(x_*) = \xi(x'_*)$ then there is no experimental datum which provides grounds for distinguishing between x_* and x'_*, and hence there are no grounds for distinguishing between $Q_*(x_*)$ and $Q_*(x'_*)$; it being implicit here that the summary function ξ provides all of the available information. It seems plausible therefore to require that

(3.4) $$\forall x_*, x'_* \in D_* : \xi(x_*) = \xi(x'_*) \Rightarrow Q_*(x_*) = Q_*(x'_*) .$$

Surrogate functions Q_* for which (3.4) holds will be said to be ξ-based.

Suppose then that the surrogate function Q_* is ξ-based. It follows from (3.4) that there is a function $\lambda : M \to \mathbb{R}^+$ with domain

$$\text{dom}\lambda = \text{codom}(\xi|D_*) ,$$

such that, for m in the domain of λ,

$$\lambda(m) = Q_*(x_*)$$

for any x_* in D_* such that $\xi(x_*) = m$. In other words, for each $n \geq 1$ and any $x_1, x_2, \ldots, x_n$ in D one has

$$Q_n(x_1, x_2, \ldots, x_n) = \lambda(\xi_n x_1 x_2 \ldots x_n) .$$

Substitution into (3.3) gives

$$\lambda(\xi_n x_1 x_2 \ldots x_n) = \lambda(\xi_1 x_1)\lambda(\xi_1 x_2) \ldots \lambda(\xi_1 x_n)$$

for each $(x_1, x_2, \ldots, x_n)$ in D_*, whereas (3.2) with $n = 1$ gives

$$\sum_{x \in D} \lambda(\xi_1 x) = 1 .$$

In other words λ must be a normed character of the summary function $\xi|D_*$ and

$$\forall x_* \in D_* : Q_*(x_*) = \lambda(\xi(x_*)) .$$

For x_* not in D_*, $Q_*(x_*)$ is zero because of (3.1). Recalling the fact that ξ is separative on D_* and the results of Section 2 we state

https://doi.org/10.1017/S0004972700043409 Published online by Cambridge University Press

PROPOSITION 3.1. *Let* ξ *be an* M*-valued summary function in* X *which is separative on* D_* *and let* Q_* *be a* ξ*-based surrogate function; then*

$$(3.5) \qquad Q_*(x_*) = \begin{cases} D_\chi^{-n(x_*)}\chi(\xi(x_*)) , & x_* \in D_* , \\ 0 & , \text{otherwise}, \end{cases}$$

where χ *is a non-trivial non-negative character of* $\xi|D_*$ *and* D_χ *is given by* (2.3) *with* K *replaced by* D .

Since X is a subset of X_* we may substitute any x belonging to X in place of x_* in (3.5) to give

$$(3.6) \qquad Q(x) = \begin{cases} D_\chi^{-1}\chi(\xi_1 x) , & x \in D , \\ 0 , & x \notin D , \end{cases}$$

where $Q = Q_1$ is a ξ-based surrogate probability distribution which deputizes for P . We say that Q is a macroscopic description of the distribution P based on the summary function ξ .

It should be noted that a summary function has, in general, more than one non-trivial non-negative character so that there will be several macroscopic descriptions based on the same summary function. This non-uniqueness plays an important role. A macroscopic description is a surrogate probability distribution of a particular functional form which involves unknown parameters. Different values of these parameters correspond to different characters and so determine different descriptions. In conventional terminology the problem of the "best" choice of character is the problem of the "best" estimate of the corresponding parameter values.

In subsequent discussions we place the emphasis on the macroscopic description Q rather than on the surrogate function Q_* because the latter is easily expressed in terms of the former. Indeed suppose that

$$(3.7) \qquad D = \{s_1, s_2, \ldots, s_m\}$$

and for each $x_* = (x_1, x_2, \ldots, x_n)$ in X_* and each $j = 1, 2, \ldots, m$ let $x_*(s_j)$ be the number of times s_j occurs in the sequence x_*. Then

$$\forall x_* \in D_* : Q_*(x_*) = \phi_1^{n_1}\phi_2^{n_2} \ldots \phi_m^{n_m}$$

where $n_j = x_*(s_j)$ and $\phi_j = Q(s_j)$.

The simplest example of a macroscopic description is obtained by taking the separative summary function ξ to be the identity map on X_*. This is the standard case of statistical theory in which the whole sample is available. The characters of $\xi|D_*$ satisfy the equations

$$\chi(x_1, x_2, \ldots, x_n) = \chi(x_1)\chi(x_2) \ldots \chi(x_n)$$

for any $x_1, x_2, \ldots, x_n$ in D. Thus if D is given by (3.7),

$$\chi(x_1, x_2, \ldots, x_n) = \prod_{j=1}^{m} [\chi(s_j)]^{x_*(s_j)} .$$

It follows that the non-trivial non-negative characters are determined in terms of m non-negative parameters $\chi_1, \chi_2, \ldots, \chi_m$, namely, $\chi_j = \chi(s_j)$, not all of which are zero. The corresponding macroscopic description is given by

$$Q(s_j) = \chi_j/(\chi_1 + \chi_2 + \ldots + \chi_m) , \quad j = 1, 2, \ldots, m$$

and the surrogate function Q_* is given by

$$Q_*(x_*) = \prod_{j=1}^{m} [\chi_j/(\chi_1 + \chi_2 + \ldots + \chi_m)]^{x_*(s_j)}$$

for each x_* in D_*. There remains, of course, the problem of "estimating" the parameters $\chi_1, \chi_2, \ldots, \chi_m$ when one does have a particular set of sample values y_*, say. In this particular case the simple-minded and obvious thing to do is to take

$$\chi_j = y_*(s_j) , \quad j = 1, 2, \ldots, n ,$$

so that the corresponding macroscopic description is just the sample distribution of the observation at hand.

https://doi.org/10.1017/S0004972700043409 Published online by Cambridge University Press

In practical problems one usually deals with universal summary functions ξ and although ξ in (3.6) is only required to be a non-negative character of $\xi|D_*$ it is convenient to restrict our macroscopie descriptions to those derived from the non-negative characters of ξ and we shall adopt this restriction in the discussion which follows.

4. Summary functions and sufficiency

It is worthwhile noting the following connection between macroscopic descriptions and the concept of sufficiency. In a sense made more precise below a separative summary function is a sufficient statistic for any of the macroscopic descriptions to which it leads.

Let $\xi : X_* \to M$ be a universal M-valued summary function which is separative on D_* . Let χ be a fixed non-trivial non-negative character of ξ and let the macroscopic description obtained from χ by (3.6) be denoted by $Q(\cdot|\chi)$ so that

$$Q_*(x_*|\chi) = \begin{cases} \chi(\xi(x_*))D_\chi^{-n(x_*)} , & x_* \in D_* , \\ \\ 0 , & x_* \notin D_* . \end{cases}$$

In the discussion which follows χ plays the role of the parameter in text-book discussions of sufficiency.

Let m belong to the codomain of $\xi|D_*$, then

$$Q_*(\xi^{-1}(m)|\chi) = \sum_{x_*:\xi(x_*)=m} Q_*(x_*|\chi)$$

is the surrogate probability attaching to the set of sample values x_* which have summary m . Thus

$$Q_*(\xi^{-1}(m)|\chi) = N(\xi^{-1}(m))\chi(m)D_\chi^{-\nu(m)} ,$$

where $N(\xi^{-1}(m))$ is the number of elements x_* such that $\xi(x_*) = m$ and $\nu(m)$ is their common arity. Replacing m by $\xi(x_*)$ we obtain

$$Q_*(x_*|\chi) = Q_*(\xi^{-1}\xi(x_*)|\chi)\,[N(\xi^{-1}\xi(x_*))]^{-1} .$$

This equation exhibits the sufficiency of the summary function in respect of the macroscopic descriptions derived from it since the second factor on the right does not depend on the particular character χ .

It is the sufficiency of the summary function in respect of the distributional form of the macroscopic descriptions based on it that gives meaning to the use of standard procedures for the estimation of the parameters in question. A detailed discussion of the estimation problem will be published elsewhere.

5. Macroscopic descriptions based on linear aggregation

Suppose that X is a commutative semigroup with identity, the semigroup operation being denoted by + and the identity by 0 . Let Z_+ be the set of positive integers and let the universal summary function $\xi : X_* \to Z_+ \times X$ be defined by

$$(5.1) \qquad \xi(x_1, x_2, \ldots, x_n) = (n, x_1 + x_2 + \ldots + x_n) .$$

We refer to the operations performed by this summary function as linear aggregation, it is clearly separative.

The characters of ξ are the complex-valued functions χ defined on $Z_+ \times X$ which have the property

$$(5.2) \qquad \chi(n, x_1 + x_2 + \ldots + x_n) = \chi(1, x_1)\chi(1, x_2) \ldots \chi(1, x_n) ,$$

for any positive integer n and any $x_1, x_2, \ldots, x_n$ in X . Now (5.2) implies that

$$\chi(n, x) = [\chi(1, 0)]^{n-1}\chi(1, x) , \quad (n, x) \in Z_+ \times X .$$

It is easily verified that if χ is non-trivial we must have $\chi(1, 0) \neq 0$ and so writing

$$\chi'(x) = \chi(1, x)/\chi(1, 0) ,$$

we obtain

$$\chi(n, x) = a^n\chi'(x) , \quad (n, x) \in Z_+ \times X ,$$

where $a = \chi(1, 0) \neq 0$ and χ' is a non-trivial character of the semigroup X . Thus the macroscopic descriptions based on linear

https://doi.org/10.1017/S0004972700043409 Published online by Cambridge University Press

aggregation are given by

$$(5.3) \qquad Q(x) = \begin{cases} \chi(x) / \sum_{y \in D} \chi(y) , & x \in D , \\ \\ 0 , & x \notin D , \end{cases}$$

where χ is a non-trivial non-negative character of the semigroup X .

In many practical applications X arises in the following way. For each $i = 1, 2, \ldots, k$, T_i is a commutative semigroup with identity and X is the cartesian product of $T_1, T_2, \ldots, T_k$ with the natural semigroup operation derived from those in the component semigroups. Thus if $x = (t_1, t_2, \ldots, t_k)$ and $x' = (t_1', t_2', \ldots, t_k')$ are in X then

$$x + x' = (t_1+t_1', t_2+t_2', \ldots, t_k+t_k') .$$

In such a case the characters χ of X can be shown to be of the form

$$\chi(x) = \chi_1(t_1)\chi_2(t_2) \ldots \chi_k(t_k) ,$$

where $x = (t_1, t_2, \ldots, t_k)$ is in X and, for each $i = 1, 2, \ldots, k$, χ_i is a character of the semigroup T_i .

By way of illustration let h_i , $i = 1, 2, \ldots, k$, be positive numbers and suppose that

$$T_i = \{nh_i : n \in Z\} ,$$

where the semigroup operation in each T_i is real number addition. The characters of T_i are of the form

$$\chi_i(t) = b_i^t , \quad t \in T_i ,$$

where b_i is a real number. Thus the characters of X are given by the expressions

$$\chi(x) = b_1^{t_1} b_2^{t_2} \ldots b_k^{t_k}$$

for each $x = (t_1, t_2, \ldots, t_k)$ in X . The non-trivial non-negative

https://doi.org/10.1017/S0004972700043409 Published online by Cambridge University Press

characters correspond to the choice of positive $b_1, b_2, \ldots, b_k$ and, with such a choice, if D is given by (3.7), where

(5.4) $$s_j = \left(s_{j1}, s_{j2}, \ldots, s_{jk}\right) , \quad j = 1, 2, \ldots, m ,$$

then the macroscopic descriptions are given by expressions of the form

(5.5) $$Q\left(s_j\right) = b_1^{s_{j1}} b_2^{s_{j2}} \ldots b_k^{s_{jk}} \left[\sum_{i=1}^{m} b_1^{s_{i1}} b_2^{s_{i2}} \ldots b_k^{s_{ik}}\right]^{-1} .$$

6. Maximum entropy distributions

Jaynes, [1] and [2], has indicated a formal development of statistical mechanics based on an information-theoretic principle of entropy maximisation. In a notation suitable for comparison with our results his method may be formulated in the following way.

Let $\{s_1, s_2, \ldots, s_m\}$ be a finite set and let g_i , $i = 1, 2, \ldots, k$, be $k < m$ real-valued functions defined on that set. Suppose that γ_i , $i = 1, 2, \ldots, k$, are k given real numbers. Jaynes showed that the probability distribution p on the set $\{s_1, s_2, \ldots, s_m\}$ which maximised the information-theoretic entropy

$$S_I = - \sum_{j=1}^{m} p\left(s_j\right) \log p\left(s_j\right) ,$$

subject to the constraints

$$\sum_{j=1}^{m} g_i\left(s_j\right) p\left(s_j\right) = \gamma_i , \quad i = 1, 2, \ldots, k ,$$

is given by

(6.1) $$p_0\left(s_j\right) = \left[\Pi\left(\lambda_1, \lambda_2, \ldots, \lambda_k\right)\right]^{-1} \exp\left\{- \sum_{i=1}^{k} \lambda_i g_i\left(s_j\right)\right\} ,$$

where

$$\Pi\left(\lambda_1, \lambda_2, \ldots, \lambda_k\right) = \sum_{j=1}^{m} \exp\left\{- \sum_{i=1}^{k} \lambda_i g_i\left(s_j\right)\right\}$$

and the $\lambda_1, \lambda_2, \ldots, \lambda_k$ are real numbers determined by the constraints, namely,

$$(6.2) \quad \gamma_i = -\frac{\partial}{\partial \lambda_i} \log \prod(\lambda_1, \lambda_2, \ldots, \lambda_k) \ , \quad i = 1, 2, \ldots, k \ .$$

Jaynes interpreted this result as providing a constructive criterion for determining probability distributions on the basis of partial knowledge. Noting that this criterion led to expressions formally equivalent to those of statistical mechancis he argued that in the resulting subjective statistical mechanics the usual rules are justified independently of experimental verification because, whether or not the results agree with experiment, they represent the best estimate that could have been made on the basis of the information available. For Jaynes the partial knowledge, on the basis of which one is required to determine the distribution p_0, is provided by the available information. This is supposed to be specified by the quantities γ_i, $i = 1, 2, \ldots, k$, which are interpreted as average values of the functions g_i, $i = 1, 2, \ldots, k$, respectively. Thus the problem considered by Jaynes is essentially the determination of a probability distribution in terms of certain known average values. This problem is similar to the one considered in the last section where one determined the form of a macroscopic description in terms of certain linear aggregates. To highlight this similarity we recast (5.5) in the form (6.1).

Introduce functions $g_i : D \to T_i$ defined by

$$g_i(s_j) = s_{ji} \ , \quad i = 1, 2, \ldots, k \ ; \quad j = 1, 2, \ldots, m \ ,$$

where the s_{ji} are defined by (3.7) and (5.4). For each $i = 1, 2, \ldots, k$ write

$$\lambda = -\log b_i \ ;$$

then (6.1) becomes (5.5). Now the result of measurement is given by (5.1) and the right-hand side of that equation can be written, in the present instance, as a vector whose ith component is

https://doi.org/10.1017/S0004972700043409 Published online by Cambridge University Press

$$G_i = \sum_{j=1}^{m} g_i(s_j)x_*(s_j) \quad , \quad i = 1, 2, \ldots, k \ ,$$

where $x_* = (x_1, x_2, \ldots, x_n)$ is the observation in question. Thus equation (6.2) is analagous to estimating the parameters b_i by equating the surrogate mean values

$$\sum_{j=1}^{m} Q(s_j)g_i(s_j)$$

to the corresponding quantities $n^{-1}G_i$; these quantities are, of course, just averages over the observation at hand.

It follows from the formal similarity to the maximum entropy distributions that one can develop statistical mechanics in a systematic way through the concept of a macroscopic description. In such a development statistical mechanics becomes explicitly a surrogate statistical description of microscopic phenomena which is based on macroscopic measurement. However it is not a subjective theory, on the contrary it is empirically based in the following sense. Theory cannot tell us which summary functions will lead to results in agreement with experiment. Indeed one has to experiment to find out which summary functions do provide useful macroscopic descriptions of microscopic phenomena, useful in the sense that they do agree reasonably well with the results of experimentation. The same empirical basis underlies the use of macroscopic descriptionr in other fields of enquiry.

References

[1] E.T. Jaynes, "Information theory and statistical mechanics", *Phys. Rev. A* 106 (1957), 620-630.

[2] E.T. Jaynes, "Information theory and statistical mechanics", *Statistical physics*, 181-218 (Brandeis University Summer Institute Lectures in Theoretical Physics, 1962, 3. W.A. Benjamin, New York, Amsterdam, 1963).

[3] W.S. Robinson, "Ecological correlations and the behavior of individuals", *Amer. Sociological Rev.* 15 (1950), 351-357.

[4] H. Theil, *Linear aggregation of economic relations* (Contributions to Economic Analysis, 7. North-Holland, Amsterdam, 1954).

Department of Mathematics,
Monash University,
Clayton,
Victoria.