INFORMATION AND QUESTIONNAIRES IN STATISTICAL INFERENCE*

by

George T. Duncan

Technical Report No. 140

University of Minnesota
Minneapolis, Minnesota

June, 1970

Submitted as a Thesis to the Faculty of the Graduate School of the University of Minnesota in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

*The typing and duplication of this report was supported by National Science Foundation Grant NSF-GP-9556.

INFORMATION AND QUESTIONNAIRES

IN STATISTICAL INFERENCE

A THESIS

SUBMITTED TO THE FACULTY OF THE GRADUATE SCHOOL

OF THE UNIVERSITY OF MINNESOTA

By

George T. Duncan

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

December, 1970

George T. Duncan
497 words

ABSTRACT

Information and Questionnaires

in Statistical Inference

A charging scheme based on the resolution of questions strikes a new direction from the approach of Claude Picard, Théorie des Questionnaires, Gauthier-Villars, Paris (1965). The relationship between questionnaire theory and noiseless coding theory is explored. Graph theoretic methods are used to obtain results valid for codes in which words are constructed from arbitrary mixtures of alphabets, as well as arborescence questionnaires, i.e., those having representation as rooted, directed trees. A charge of $\log d$ for each resolution d question is justified by an equity principle. Using this charging scheme an extended noiseless coding theorem shows that the average charge for a heterogeneous questionnaire is bounded below by the Shannon entropy. This result is shown to hold for both finite and countable state spaces. The decision theoretic problem of choosing a questionnaire to resolve a finite state space is examined. Certain admissibility and essentially complete class results are obtained which indicate the structure of optimal heterogeneous questionnaires. In particular it is shown that, for an essentially complete class of questionnaires, charges for state determination are ordered inversely to state probabilities. The regions of minimum charge are shown to be convex. If the state space has a finite number, m, of elements, then an essentially complete class of questionnaires has an average charge depending

only on the $(m - 2)$ largest state probabilities. An initial resolution m question gives the minimax questionnaire, while equally likely state probabilities give a least favorable prior distribution and $\log m$ is the lower value of the game.

A dynamic programming approach is used to provide an algorithm for finding an optimal questionnaire. Approximations to the dynamic programming algorithm are proposed and evaluated.

A charging scheme for a lattice questionnaire is presented which maintains the Shannon entropy as a lower bound on average questionnaire charge. The fact that this lower bound can be attained allows a characterization of the Shannon entropy in terms of average questionnaire charge to be developed.

Certain information theoretic results based on the Shannon entropy function are extended to results about uncertainty functions, as defined by DeGroot (Ann. Math. Statist. 33 (1962) 404-419). Results of Renyi (Studia Scientarium Mathematicarum Hungarica 2 (1967) 249-256) concerning data reduction and sufficiency are generalized. Countable state space results are achieved through a version of Jensen's inequality which is valid for a function from sequence space. Payment schedules for a forecaster which allow no profit in dishonesty and promote diligence are studied. The relationship between uncertainty functions which are Bayes risk functions and payment schedules which emphasize the value of information are studied. Information in an observable random variable X about a random parameter θ is defined as the average reduction in uncertainty about θ given X. Minimum average questionnaire charge is examined

as an uncertainty function. Questionnaire information is compared to Shannon information. By simultaneously determining a sufficiently large number of parameter realizations, the questionnaire information per parameter realization may be made arbitrarily close to the Shannon information per parameter realization.

Approved
R J Buehler

I dedicate this thesis to my wife

Mary

for her patience, assistance, and smiles.

ACKNOWLEDGEMENT

The author expresses his sincere gratitude to Professor Robert J. Buehler for his encouragement and guidance in the preparation of this dissertation. In addition, the interesting discussions with Professor William Sudderth, the enthusiasm and knowledge of Professor Milton Sobel, and the typing skill of Mr. Gerald DuChaine are appreciated.

TABLE OF CONTENTS

Chapter I

Introduction and Summary

The present knowledge of a Bayesian decision maker is reflected in his probability distribution over the possible states of nature. The decision maker will order the actions available to him according to their Bayes risk. Unfortunately, it often happens that no action has an acceptably small Bayes risk, and therefore the decision maker is unwilling to choose any action. In such a situation, the decision maker is advised to modify his probability distribution through one or more of three techniques.

First, he may accumulate further information in a statistical fashion through experimentation. Chapter IV of this paper, continuing in the tradition of DeGroot (1962), will extend certain information theoretic results based on the Shannon (1948) entropy function to results about uncertainty functions. In particular it will generalize results of Rényi (1967) which are concerned with data reduction and sufficiency. The generalizations to a countable state space are achieved through the use of a version of Jensen's inequality valid for a function from sequence space, R^{∞}.

Second, the decision maker may employ a forecaster who is then given the responsibility of producing a more satisfactory probability measure. Payment schedules for the forecaster are studied, in Chapter IV, which allow "no profit in dishonesty" and "promote diligence." The first property is related to "keeping

a forecaster honest," introduced by McCarthy (1956). Attention is also given to payment schedules which emphasize the value of information to the decision maker, a concept introduced by Marschak (1959).

Third, the decision maker may have available a sure sequential procedure or strategy for separating the state space until the true state is found. Picard (1965) calls such a procedure a questionnaire. Depending on the field in which it is applied, a questionnaire may be called a diagnostic schedule, a trouble-shooting routine, a taxonomic key, a weighing design, a search scheme, or even a "Twenty Questions" game strategy. The majority of this paper is devoted to the development of a theory of questionnaires in which the questioner is allowed complete freedom in the resolution of the questions to be used at any stage of the questioning. A charge will be incurred depending on the nature of the questionnaire, and, in particular, (most often) on the resolution of the questions asked. This strikes a new direction from the approach of Picard and others, such as Petolla (1969) and Dubail (1967).

Chapter II emphasizes the relationship between questionnaire theory and noiseless coding theory. In this chapter a specific charge of $\log d$ for each resolution d question is justified by appealing to an equity principle. Graph theoretic methods are used to generalize Kraft's (1949) theorem to obtain a result valid for codes where words are constructed from arbitrary mixtures of alphabets, as well as arborescence questionnaires, i.e., those

having representation as rooted, directed trees. This allows an extended noiseless coding theorem to be proved which provides that the average charge for a heterogeneous questionnaire is bounded below by the Shannon entropy. A condition for equality is given which connects a state probability with the number of questions of each resolution required to determine that state. The only charge based on the resolution d of a question which permits this theorem is $\log d$. It is also shown that equality is attained between Shannon entropy and average questionnaire charge if and only if each question is Shannon efficient, i.e., partitions the state space into sets of equal probability. The Shannon lower bound theorem is generalized to a countable state space. Some discussion is given of the continuous state space in terms of ϵ-entropy. A charge dependent theory of lattice questionnaires is developed which allows a characterization of the Shannon entropy in terms of minimum average questionnaire charge.

Chapter III examines the decision theoretic problem of choosing a questionnaire to resolve a finite state space. The states are ordered so that $p_1 \geq p_2 \geq \ldots \geq p_m$ where $p = (p_1, p_2, \ldots, p_m)$ is the probability vector over the state space. A questionnaire is called admissible if for all such probability vectors no questionnaire is preferred to it and strictly preferred for some p. The explicit form of the optimal questionnaires are obtained for $m = 3$ and $m = 4$. The regions of optimality are shown to be convex in the probability simplex for any m. Certain admissibility and

essentially complete class theorems are obtained which indicate the structure of optimal heterogeneous questionnaires. It is shown that there exists an essentially complete class of questionnaires with state charges ordered inversely to state probabilities. Further, the set of questionnaires whose average charge depends only on p through $p_1, p_2, \ldots, p_{m-2}$ form an essentially complete class. The questionnaire consisting of an initial resolution m question is shown to be minimax, $p^* = (\frac{1}{m}, \frac{1}{m}, \ldots, \frac{1}{m})$ is least favorable, and $\log m$ is the lower value of the game. A dynamic programming approach is used to provide an algorithm for finding an optimal questionnaire. An essentially complete class result substantially reduces the number of partitions which must be examined in using this algorithm. Approximations to the dynamic programming solution are considered.

In Chapter IV the information in an observable random variable X about a random parameter θ is defined as the average reduction in uncertainty about θ given X. Minimum average questionnaire charge is examined as an uncertainty function. Questionnaire information is compared to Shannon information. By simultaneously determining a sufficiently large number of parameter realizations, the questionnaire information per parameter realization may be made arbitrarily close to the Shannon information per parameter realization.

Payments to a forecaster are considered which are commensurate with the value of the forecast to a questioner.

It is noted that the choice of a questionnaire has implications for an individual's subjective probability.

Chapter II

State Determination and Noiseless Coding

1. Fundamentals.

1.1. Introduction.

It is commonly recognized among workers in information theory that the instantaneous codes of noiseless communication theory correspond to schemes of questioning which anticipate unambiguous, truthful replies. Implicitly, this is evident from Barnard's (1951) treatment of a weighing design problem; it is made somewhat more explicit, in the same context, by Kerridge (1961). In fact, such standard textbooks on information theory as Ash (1965) use this correspondence to illustrate, with simple examples, their material on coding.

A scheme for specifying the questioning procedure might be described by a variety of labels. These can include a diagnostic schedule, a trouble-shooting routine, a taxonomic key, a "Twenty Questions" game strategy, a search scheme, or,as above, a weighing design, all depending on the field of application. A generic term which might encompass all of these is questionnaire. A questionnaire deals with a finite or countably infinite state space. The task of a questionnaire is to single out one state which has some unique characteristic.

An introduction to the subject of questionnaires is given by Claude Picard in his book Théorie des Questionnaires (1965). He uses some of the results of communication theory, in particular,

the Huffman (1952) coding scheme, to obtain a "best" questionnaire in a certain class. The intimate relationship between questionnaire theory (in the homogeneous case where each question has the same number of possible responses) and coding theory is spelled out in more detail in further work by Picard (1969). Loosely speaking, this correspondence identifies one symbol from an alphabet containing exactly d characters with one question having d possible responses.

The fundamental noiseless coding theorem of communication theory was first presented by Shannon (1948); it has a questionnaire-theoretic analogue. This has been more or less evident to many researchers who have attempted to apply information theory to a variety of fields. Examples include the work on group testing by Sobel (1960), the brief discussion of search theory by Campbell (1968), and the mathematical treatment of one aspect of taxonomy by MacDonald (1952).

Looking at the problem of state determination from the viewpoint of a questionnaire suggests a generalization of the noiseless coding theorem which is proved in this chapter. (The generalization is in a different direction from that of Billingsley (1961).) The fundamental notions of questionnaires and noiseless coding are developed in this chapter to provide sufficient background for the theorem.

1.2. Definition of a Questionnaire.

Picard (1965) treats a questionnaire in a very useful graph-theoretic manner. This formalizes the quite natural representations

which can be found throughout the literature, and as early as Shannon and Weaver (1949). The graph theory used by Picard is standard and may be found in such books as Flament (1963).

It is within the spirit of modern mathematics to follow a slightly different approach and construct a set-theoretic foundation for questionnaire theory and then to demonstrate the extensive and illuminating interplay with graph theory. First, rather informally, a questionnaire may be thought of in terms of three components: (1) a state space Θ containing a finite or countable number of elements, (2) a countable set ω of arbitrary symbols used purely for convenience, and (3) an operator Q which acts in a particular way on the subsets of the state space.

Thus a questionnaire may be treated as a triple (Θ, ω, Q). The elements of Θ are called states and denoted by θ, while subsets of Θ are denoted, generically, by Θ^*. When Q operates on Θ^*, the result is a family whose elements are subsets of Θ^* plus, possibly, elements chosen from ω. The union of those image sets which are subsets of Θ^* must equal Θ^*. The set containing the empty set is not in any of the image sets. (This means that no node corresponds to the empty set.)

A graph-theoretic representation is then determined: first identify the state space, Θ, with a point, to be called the root. Then establish a point (node) for each set in the image families under Q. Allow a directed edge to join any particular node with each of the nodes corresponding to sets in the image family. (The graph-theoretic terminology used here will be primarily that

of Ore (1962).) It should be noted that the root has the distinction of being the only node which has no edge directed into it. There are no isolated nodes in the graph.

Now, given a graph-theoretic representation, the questionnaire (Θ, ω, Q) is also determined. Thus, in this context, one may speak of graphs or operators on sets interchangeably.

Figure 2.1 is an example of the graph-theoretic representation of a valid questionnaire for resolving $\Theta = \{\theta_1, \theta_2, \ldots, \theta_7\}$.

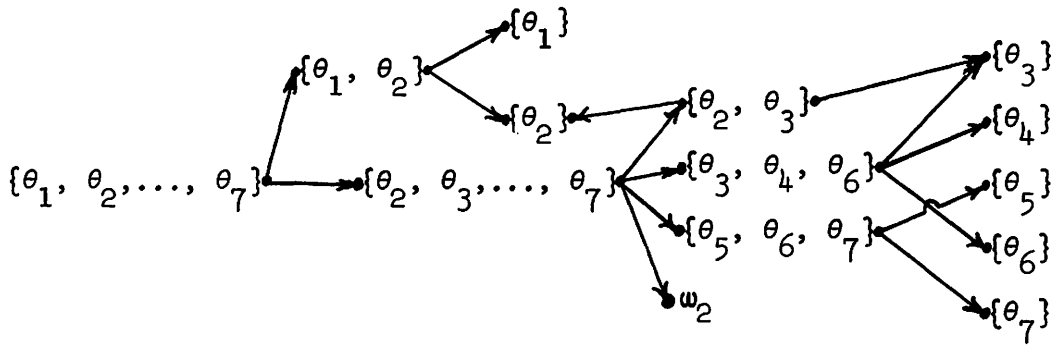


Figure 2.1

This is a quite general type of questionnaire. Notice first that each question does not necessarily partition the set on which it acts. Second, perhaps because of the nature of the questioning "device" or "format," not all "responses" may be "informative" about the true state. Hence the introduction of the ω-nodes.

Attention will be focused on a fixed state space with several operators, Q, defined on it. Thus, it will usually cause no confusion to use the same symbol, Q, to denote, simultaneously, the questionnaire, the operator, and the graph representation.

With this as background, we will now attempt to make the foregoing precise:

Let $\Omega = \Theta \cup \omega$ be the universal set, where Θ is a finite or countable set, called the state space, and ω is an arbitrary countable set of symbols $\{\omega_1, \omega_2, \ldots\}$ with the property that $\Theta \cap \omega = \emptyset$. If A is an arbitrary set, $|A|$ will denote the cardinality of A.

Definition 2.1. A question, q, is a mapping from a particular $\Theta^* \subset \Theta$ into $2^{2^{\Theta^*}} \cup \omega$. It is required that

(i) $$\bigcup_{\Theta_\alpha^* \in (q\Theta^*) \cap 2^\Theta} \Theta_\alpha^* = \Theta^*,$$

and

(ii) if $\Theta_\alpha^* \in q\Theta^*$, then $\{\emptyset\} \notin \Theta_\alpha^*$.

The function q may be said to be the question at Θ^*.

Definition 2.2. An answer, a, to a question, q at Θ^*, is a mapping from $q\Theta^*$ to one of its elements.

Definition 2.3. A questionnaire, Q, is the extension of q to the domain of all subsets of Θ. Therefore, a questionnaire, Q, is a mapping from 2^Θ into $2^{2^\Theta} \cup \omega$ with three specific stipulations: If we define Γ by $\Gamma = Q\Theta^* \cap 2^\Theta$, then

(i) $\Gamma \subset 2^{\Theta^*}$,

(ii) $$\bigcup_{\Theta_\alpha^* \in \Gamma} \Theta_\alpha^* = \Theta^*,$$

and

(iii) if $\Theta_\alpha^* \in \Gamma$, then $\{\emptyset\} \notin \Theta_\alpha^*$.

Definition 2.4. The operator Q^{-1} maps 2^{Θ} into $2^{2^{\Theta}}$ and is defined by

$$(1.2.1) \qquad Q^{-1}\Theta^* = \{\Theta^*_\alpha : Q\Theta^*_\alpha = \Theta^*\}.$$

Note that Q^{-1}, when applied to a node of the graph, gives its direct antecedants. It will actually only be used when Q is such that the right hand side of (1.2.1) is either a singleton or empty, except in Definition 2.5 to follow.

Definition 2.5. A questionnaire, Q, is valid if $Q^{-1}\{\theta_i\}$ is not empty for $i = 1, 2, \ldots$.

Definition 2.6. A state, θ_i, is said to be determined at stage k_i if k_i is the smallest integer such that $Q^{-k_i}\{\theta_i\} = \Theta$.

Definition 2.7. A questionnaire, Q, is said to be an arborescence questionnaire if the sets in $Q\Theta^*$ are disjoint for all $\Theta^* \subset \Theta$.

The reason for the choice of the term arborescence is that such questionnaires have a graph representation as a rooted, directed tree. There is precisely one edge directed into each node beyond the root. Since the graph is connected, there is then exactly one path to each of the terminal nodes.

Definition 2.8. A questionnaire is called a lattice questionnaire if it is not an arborescence questionnaire.

Definition 2.9. A question, q at Θ^*, is said to have resolution d if the sets in $q\Theta^*$ are disjoint and $|q\Theta^*| = d$.

Definition 2.10. A questionnaire will be called homogeneous if each question has the same resolution; it will be called

heterogeneous if the question resolutions may be different.

Note that a questionnaire which is either homogeneous or heterogeneous must be an arborescence questionnaire.

2. Arborescence Questionnaires.

2.1. General.

The class of questionnaires which have arborescence representation has many interesting connections with the instantaneous codes of communication theory. This section will be devoted to an exploration of these connections.

Figure 2.2, below, gives a typical homogeneous questionnaire in rooted, directed tree form.

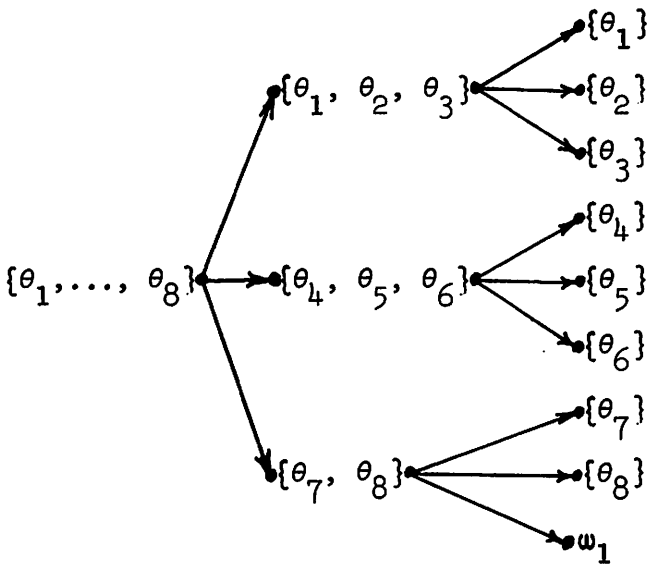


Figure 2.2

Here all questions are of resolution 3 and would often be called ternary questions.

On the other hand, a heterogeneous questionnaire might appear as in Figure 2.3.

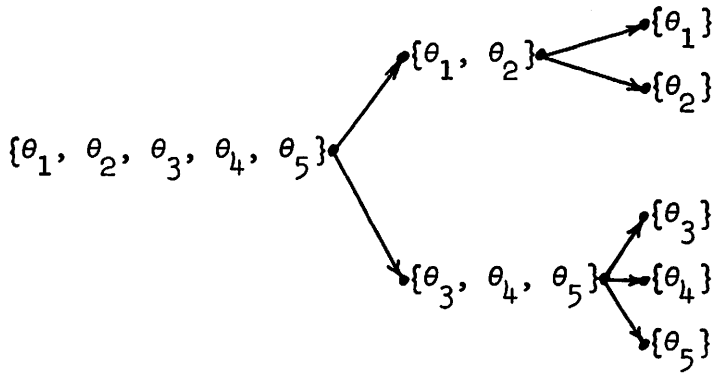


Figure 2.3

Notice that this second questionnaire includes both resolution 2 (binary) and resolution 3 (ternary) questions.

A useful set of terminology is provided by viewing a questionnaire as an asexual family tree and using botanical and geneological terms interchangeably. Metaphors will be mixed whenever it seems convenient.

With this representation, each node of the tree can be associated either with a set from a particular partition of Θ (finite or countably infinite) or an element from the set ω. The questionnaire will be valid if and only if it contains $\{\theta_i\}$ $i = 1, 2, \ldots$ as nodes.

The operator Q, when applied to a node, yields its offspring. Conversely, the operator, Q^{-1}, will yield the father of a node.

2.2. Noiseless Coding and Homogeneous Questionnaires.

In order to demonstrate the connection between questionnaire theory and noiseless coding theory, it is useful to have available some of the basic terminology of coding similar to that presented by Ash (1965). A finite collection of code characters is given. It is called the alphabet and code words are formed from it by juxtaposition. The code is uniquely decipherable if every finite sequence of code characters corresponds to, at most, one message. A sequence A is a prefix of a sequence B if B may be written as AC for some sequence C. Then a code having the property that no code word is a prefix of another code word is called instantaneous.

An instantaneous code is specified by a dictionary which prescribes one or more code words for each message which might be sent. A graph-theoretic representation can be constructed from the dictionary by reading each code word from left to right. As the first character is read, edges are established emanating from a root; the number of edges is equal to the number of characters in the alphabet used. This process is repeated from the farther node of the edge actually indicated by the code word. When the word has been read, the terminal node is identified with the message. This construction is continued for each of the code words in the dictionary. Appropriate connections are made in the graph if more than one code word corresponds to a single message. If there is a one-to-one correspondence between code words and messages, the resulting graph will be an arborescence; but whatever the form

of the resulting graph, there will exist a questionnaire with this graph-theoretic representation. Note that some terminal nodes may not correspond to any message. Such terminal nodes might be identified with elements from the set ω.

In communication theory, there is a very compelling frequency basis for assigning a probability measure to the message space and designing a coding scheme to meet optimality criteria accordingly. Within the context of questionnaire theory, there is usually a nonrepeatability of circumstances, as is often the case in search theory, which calls for a subjective interpretation of any probability measure over the state space. The effective design of a questionnaire depends on this probability measure.

Definition 2.11. The symbol $p = (p_1, p_2, \ldots)$ will denote the probability vector over the state space such that $P(\theta_i$ is the true state$) = p_i$, $i = 1, 2, \ldots$.

A basic quantity used in coding theory is the Shannon entropy.

Definition 2.12. The Shannon entropy, $H(p)$, is defined by

$$(2.2.1) \qquad H(p) = - \sum_{i=1}^{m} p_i \log_2 p_i$$

where $p = (p_1, \ldots, p_m)$ is a probability vector, and, in order to maintain continuity, $0 \log 0$ is taken to be 0.

Picard (1969) essentially gives the immediate translation of the noiseless coding theorem as stated by Ash (1965). This might be stated in questionnaire-theoretic language in the following form:

"Let Q be a valid homogeneous questionnaire of resolution d for determining θ contained in $\Theta = \{\theta_1, \ldots, \theta_m\}$ and N(Q) be the random number of questions required. Then for each given probability vector p,

$$(2.2.2) \qquad E_p N(Q) \geq -\frac{1}{\log_2 d} \sum_{i=1}^{m} p_i \log_2 p_i = \frac{H(p)}{\log_2 d} = -\sum_{i=1}^{m} p_i \log_d p_i .$$

Equality is obtained iff $p_i = d^{-n_i}$, where n_i is the number of questions required under Q to determine θ_i."

Ash further notes (in this translated form) that there exists a questionnaire of resolution d with

$$(2.2.3) \qquad -\sum_{i=1}^{m} p_i \log_d p_i \leq E_p N(Q) < -\sum_{i=1}^{m} p_i \log_d p_i + 1,$$

or

$$(2.2.4) \qquad H(p) \leq \log_2 d \; E_p N(Q) < H(p) + \log_2 d.$$

This theorem suggests the possibility of a generalization to arbitrary heterogeneous questionnaires; appropriate charges need to be made based on the resolution of the questions asked.

2.3. Charges for Heterogeneous Questionnaires.

There are two basic approaches to the study of heterogeneous questionnaires. The first, which Picard (1965) and, following him, Petolla (1966, 1969) take, is to consider the number of questions of each resolution as being fixed. An optimization problem is then to assemble these questions into a valid questionnaire which will have minimum average length. The second approach,

which is followed here, is to attempt to determine an appropriate charge for a question based on its resolution. The optimization problem will involve minimizing the average charge for a valid questionnaire where complete freedom is allowed in the choice of the resolution of the questions used. (A completely general formulation of the charge for state resolution is presented in Parkhomenko (1969) and Petolla (1969). That approach is not followed in this chapter. However, many of the results of Chapter III will be valid for a general charge formulation.)

In considering this problem of charge determination, it is first of all evident that questions of higher resolution should incur higher charges. Thus, if $c(d)$ denotes the charge for a resolution d question, we would require $c(d) \geq c(d')$ if $d \geq d'$. To obtain a more precise determination, the following principle may be invoked:

It is desired to determine relative charges for device A and device B; there is assumed available an ordinal performance criterion which relates A and B. Equity requires that if device A accomplishes no less than device B, the charge for A should not be less than the charge for B. Symbolically, this amounts to choosing a charge function $C(\cdot)$ such that,

(2.3.1) $A \succ B$ implies $C(A) \geq C(B)$

where $A \succ B$ indicated that A dominates B according to the performance criterion.

Application of this principle to questionnaires requires a very definite specification of the charging system. First, charges will be assessed for each question that is actually utilized in the questionnaire (this depends on which state is true, naturally). Second, the charge shall only depend on the resolution of the question asked and not, in particular, on the question's position in the questionnaire.

It is also necessary to have a precise notion of the performance criterion to be employed. Since the charging scheme determines charges in terms of question resolution, it is reasonable to compare the performance of two homogeneous questionnaires, one of resolution 2, i.e., composed entirely of binary questions, and one of resolution d, where d is an arbitrary integer greater than 2. Their performance will be evaluated in terms of the ratio of their average lengths. Each is set the task of resolving a state space, where one questionnaire's job may be bigger than another in the sense that its state space may contain more elements. The comparison will be made under conditions most favorable to each. For homogeneous questionnaires, these are known by the usual noiseless coding theorem (see section 2.2) to occur when each state has equal probability.

If there are 2^m states in Θ, the true state can be found after exactly m questions. That this result is the best obtainable is confirmed by the usual noiseless coding theorem. On the other hand, if there were d^n states in Θ, the true state could be found after exactly n resolution d questions. Thus, if

positive integers m and n could be found so that $2^m = d^n$, the appropriate relative charge for the resolution d question would be $\frac{m}{n} = \log_2 d$. Since, when d is not a power of two, $\log_2 d$ will be irrational, it will usually not be possible to find such an m and n. So, in general, proceed by a straightforward procedure in the Dedekind cut spirit and note

$$(2.3.2) \qquad \sup_{m,n} A(m, n) = \log_2 d = \inf_{m,n} B(m, n)$$

where

$$(2.3.2a) \qquad A(m, n) = \{ \frac{m}{n} : \frac{m}{n} \leq \log_2 d, \ m \text{ and } n \text{ positive integers} \}$$

and

$$(2.3.2b) \qquad B(m, n) = \{ \frac{m}{n} : \frac{m}{n} \geq \log_2 d, \ m \text{ and } n \text{ positive integers} \}.$$

Then given $\epsilon > 0$, we can choose (m_0, n_0) and (m_1, n_1) so that $\frac{m_0}{n_0} \in A(m, n)$ and $\frac{m_1}{n_1} \in B(m, n)$ while

$$(2.3.3) \qquad \log_2 d - \frac{\epsilon}{2} \leq \frac{m_0}{n_0} \leq \log_2 d \leq \frac{m_1}{n_1} \leq \log_2 d + \frac{\epsilon}{2} .$$

Now with $\frac{m_0}{n_0} \in A(m, n)$, a state space with d^{n_0} states can be resolved in n_0 resolution d questions. Also, a state space with 2^{m_0} states can be resolved with m_0 binary questions. But 2^{m_0} is less than d^{n_0}. Thus, $\frac{m_0}{n_0}$ is too small a charge for a resolution d question. Similarly, it is argued that $\frac{m_1}{n_1}$ is too large a charge. Thus

$$(2.3.4) \qquad \log_2 d - \frac{\epsilon}{2} \leq c(d) \leq \log_2 d + \frac{\epsilon}{2} .$$

Then, since ϵ is arbitrary, take

(2.3.5) $c(d) = \log_2 d$

as the charging scheme satisfying the stated principle. Results which are valid for an arbitrary base logarithm will be stated in the form "log d"; if the base is important it will be given, as in "$\log_2 d$." A choice of base 2 for the logarithm establishes a convenient charge of 1 unit for a binary question.

2.4. Noiseless Coding and Heterogeneous Questionnaires. Finite State Case.

This section will generalize a theorem of Kraft (1949) to heterogeneous questionnaires. An extended noiseless coding theorem can then be proved which provides that the average charge for a heterogeneous questionnaire is bounded below by the Shannon entropy. A condition for equality in terms of the state probabilities and the number of questions of each resolution required is given. The log d charging scheme is shown to be the only one yielding this theorem. It is further proved that the Shannon lower bound is obtained by a valid questionnaire iff each question is Shannon efficient, i.e., partitions the state space into sets of equal probability. It is shown that there exists a questionnaire with average charge strictly bounded above by the Shannon entropy plus one, and that this is the best upper bound available in general.

Suppose that the state space, Θ, is assumed finite so that $|\Theta| = m < \infty$. Let the valid questionnaire, Q, be given. Terminal nodes, γ_i $(i = 1,\ldots, r)$, of Q may be identified with $\{\theta_i\}$ $(i = 1,\ldots, m)$ or ω_i $(i = m + 1,\ldots, r)$. Now a count can be

made of the number of questions of each resolution, d, required to reach γ_i. This quantity can be denoted by n_{id}.

Definition 2.13. If k_i is the smallest integer such that

$$Q^{-k_i}\gamma_i = \Theta,$$

then

$$(2.4.1) \qquad n_{id} = \sum_{j=1}^{k_i} \chi_{\{d\}}(|QQ^{-j}\gamma_i|), \chi_{\{d\}}(x) = \begin{cases} 1 & d = x \\ 0 & \text{otherwise} \end{cases}.$$

Then the average charge for Q can be expressed as

$$(2.4.2) \qquad E_p C(Q) = \sum_{i=1}^{m} \sum_{d=1}^{\infty} p_i n_{id} \log_2 d.$$

Note that for $|\Theta| = m$, consideration may actually be restricted (with no loss in terms of average charge) to questions of, at most, resolution m.

The usual noiseless coding theorem can now be generalized to provide bounds on the average charge for an arbitrary heterogeneous questionnaire. The first step is to obtain a generalization of the "only if" part of a theorem due to Kraft (1949) (Kraft's theorem is Theorem 2.3.1 in Ash (1965)):

Theorem 2.1.

If a questionnaire Q is valid and uses precisely n_{id} resolution d questions to determine θ_i (i = 1,..., m), then

$$(2.4.3) \qquad \sum_{i=1}^{m} \prod_{d=1}^{\infty} d^{-n_{id}} \leq 1.$$

Proof:

The questionnaire Q will determine, through equation (2.4.1) a vector $w = (w_1, w_2, \ldots)$ where

$$(2.4.4) \qquad w_d = \max_{i=1,\ldots,m} n_{id} \quad (d = 1,2,\ldots).$$

(Note that the maximum need only be taken over indices, i, associated with terminal nodes which are assigned to θ-values. This is because, by the definition of a questionnaire, ω-terminal nodes are brothers of terminal nodes which are assigned to θ-values and hence use the same number of questions of each resolution.)

The basic strategy of the proof is to extend the graph Q to a graph $\overline{Q}$ so that $\overline{Q}$ uses precisely w_d resolution d questions in reaching any of its terminal nodes. This can be done by noting the deficiency in resolution d questions at each γ_i, namely, $w_d - n_{id}$. A tree is then constructed from each γ_i, which repairs this deficiency by successively constructing questions of the appropriate resolution. As an example of this,consider the following graph, Q:

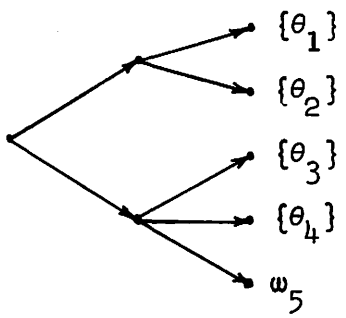


Figure 2.4

This would be extended to the graph, $\overline{Q}$, with the following form:

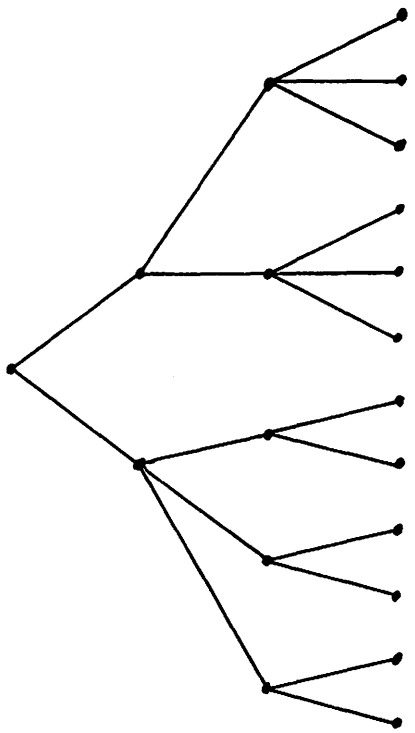


Figure 2.5

It is now desired to determine the total number of terminal nodes of the graph, $\overline{Q}$. Fix an index i. The terminal nodes of $\overline{Q}$ which are descendents of γ_i all may be reached after exactly w_d resolution d questions. The order in which these questions are asked will depend on i. Thus, in the example, descendents of $\{\theta_1\}$ which are terminal nodes of $\overline{Q}$ are reached by a (2, 2, 3) pattern for the resolution of questions, while descendents of ω_5 which are terminal nodes of $\overline{Q}$ are reached after a (2, 3, 2) pattern. Now, for any fixed pattern, a combinatorial tree could be constructed illustrating the multiplication principle and thereby (using commutativity) having $\prod_{d=1}^{\infty} d^{w_d}$ terminal nodes. Since this is true for any pattern, $\overline{Q}$ must have $\prod_{d=1}^{\infty} d^{w_d}$ terminal nodes.

If $\{n_{id}: d = 1, 2, \ldots\}$ gives the number of questions of each resolution required to determine θ_i by Q, then $\prod_{d=1}^{\infty} d^{w_d - n_{id}}$ terminal nodes will have been added to produce $\bar{Q}$ provided $w_d \neq n_{id}$ for some d; otherwise no terminal nodes will have been added. Let C be the set of all indices, i ($i = 1, \ldots, m$), such that $\{\theta_i\}$ is identified with a terminal node of $\bar{Q}$; let C^c be the set of all indices, i, such that $\{\theta_i\}$ is not identified with a terminal node of $\bar{Q}$. Then since $\{\theta_i\}$ cannot be a descendent of $\{\theta_j\}$, the following accounting inequality holds:

$$(2.4.5) \qquad \sum_{i \in C^c} \prod_{d=1}^{\infty} d^{w_d - n_{id}} + |C| \leq \prod_{d=1}^{\infty} d^{w_d}.$$

But then since $i \in C$ requires $w_d = n_{id}$, equation (2.4.5) can be written as

$$(2.4.6) \qquad \sum_{i \in C^c} \prod_{d=1}^{\infty} d^{w_d - n_{id}} + \sum_{i \in C} \prod_{d=1}^{\infty} d^{w_d - n_{id}} \leq \prod_{d=1}^{\infty} d^{w_d}.$$

Hence dividing by $\prod_{d=1}^{\infty} d^{w_d}$ in (2.4.6),

$$(2.4.7) \qquad \sum_{i \in C^c} \prod_{d=1}^{\infty} d^{-n_{id}} + \sum_{i \in C} \prod_{d=1}^{\infty} d^{-n_{id}} \leq 1.$$

Therefore inequality (2.4.3) follows. □

It is interesting to note that, unlike the situation in the homogeneous case, the converse to the theorem is false, i.e., it is not necessarily possible to construct a valid questionnaire from $\{n_{id}\}$ merely because they satisfy (2.4.3). As an example of this consider: $m = 3$ with $n_{12} = 1 = n_{23} = n_{32} = n_{33}$ and $n_{13} = 0 = n_{22}$. Then equality is obtained in (2.4.3) since $2^{-1}3^{0} + 2^{0}3^{-1} + 2^{-1}3^{-1} = 1$. But no questionnaire of the form given exists.

A restatement of the previous theorem is possible in coding theory terminology. In this context, a message is coded as a finite sequence of characters. Suppose that one extends the usual convention that each character is chosen from an alphabet with a total of d characters and allows the characters to be chosen arbitrarily from distinct alphabets having $1, 2, 3, \ldots$ characters each. Thus a word may be mixed with regard to the alphabet employed. Now given a particular word, θ_i, one can count the number of characters from each alphabet actually used. This quantity might be called n_{id}. Then we have the following restatement of Theorem 2.1:

Corollary 2.1.

Denote by α_d an alphabet with precisely d characters. Let n_{id} $(i = 1, \ldots, m;\ d = 1, 2, \ldots)$ be the number of characters from α_d which are employed in the i^{th} code word. Then an instantaneous code must satisfy inequality (2.4.3).

Now the questionnaire in its graph-theoretic form may have terminal nodes which are not assigned to θ-values (instead they are assigned ω-values). In terms of coding theory, Fano (1961) describes a code with this representation as not being complete. If all terminal nodes are assigned to θ-values then equality will prevail in (2.4.3), according to Corollary 2.2 to follow.

Definition 2.14. A questionnaire Q (whether arborescence or lattice) is said to be adapted to a state space $\mathfrak{B}$ if Q maps $2^{\mathfrak{B}}$ into $2^{2^{\mathfrak{B}}}$.

Corollary 2.2.

Let Q be a questionnaire operator. Then Q is adapted to Θ if and only if

$$(2.4.8) \qquad \sum_{i=1}^{m} \prod_{d=1}^{m} d^{-n_{id}} = 1.$$

Proof:

Only if Q has no elements from ω in its image may each terminal node in the questionnaire be classified as either a θ_i terminal node of the original graph or as a descendent of such a node. Then equality will hold in (2.4.5). Therefore (2.4.8) is affirmed by the argument of the proof of Theorem 2.1. □

Based on Theorem 2.1, a generalized noiseless coding theorem can be obtained which guarantees that the average charge for a valid heterogeneous questionnaire is bounded below by the Shannon entropy.

Theorem 2.2.

Let $\Theta = \{\theta_1, \ldots, \theta_m\}$ be a finite state space and $p = (p_1, \ldots, p_m)$ be a probability vector. If Q is a valid heterogeneous questionnaire and $C(Q)$ is the random charge based on $\log d$ for each resolution d question, then

$$(2.4.9) \qquad H(p) \leq E_p C(Q).$$

Equality is attained if and only if $n_{id} = 0$ for all $d > m$ and

$$(2.4.10) \qquad p_i = \prod_{d=2}^{m} d^{-n_{id}} \quad (i = 1, \ldots, m),$$

where n_{id} is the number of resolution d questions specified by Q to determine θ_i.

Proof:

The proof makes use of a basic inequality of information theory (derivable from Jensen's inequality) and Theorem 2.1.

The basic inequality is

$$(2.4.11) \quad -\sum_{i=1}^{m} p_i \log p_i \leq -\sum_{i=1}^{m} p_i \log q_i$$

where

$$(2.4.11a) \quad \sum_{i=1}^{m} p_i = \sum_{i=1}^{m} q_i = 1$$

and

$$(2.4.11b) \quad p_i, q_i \geq 0 \quad (i = 1, \ldots, m).$$

Equality is obtained if and only if

$$(2.4.12) \quad p_i = q_i \quad \text{for all} \quad i = 1, \ldots, m.$$

Now define q_i $(i = 1, \ldots, m)$ to satisfy (2.4.11a, b) by

$$(2.4.13) \quad q_i = \prod_{d=1}^{\infty} d^{-n_{id}} / \sum_{i=1}^{m} \prod_{d=1}^{\infty} d^{-n_{id}}.$$

But then

$$(2.4.14) \quad -\sum_{i=1}^{m} p_i \log q_i = \sum_{i=1}^{m} \sum_{d=1}^{\infty} p_i n_{id} \log d + \left(\sum_{i=1}^{m} p_i\right) \log\left(\sum_{i=1}^{m} \prod_{d=1}^{\infty} d^{-n_{id}}\right)$$

using basic properties of the logarithm.

Theorem 2.1 shows that the last term on the right hand side of equation (2.4.14) is nonpositive. Thus inequality (2.4.11) requires that

$$(2.4.15) \quad -\sum_{i=1}^{m} p_i \log p_i \leq \sum_{i=1}^{m} \sum_{d=1}^{\infty} p_i n_{id} \log d$$

which gives equation (2.4.9) by Definitions 2.11 and 2.12.

Now if (2.4.10) is satisfied and $n_{id} = 0$ for $d > m$, then

$$(2.4.16) \quad -\sum_{i=1}^{m} p_i \log p_i = -\sum_{i=1}^{m} p_i \log \prod_{d=2}^{m} d^{-n_{id}} = \sum_{i=1}^{m} \sum_{d=2}^{m} p_i n_{id} \log d$$

$$= \sum_{i=1}^{m} \sum_{d=1}^{\infty} p_i n_{id} \log d = E_p C(Q).$$

On the other hand if equality is obtained in (2.4.9) then we use the "only if" part of equation (2.4.12) and by equation (2.4.14),

$$(2.4.17) \quad \log \left(\sum_{i=1}^{m} \prod_{d=1}^{\infty} d^{-n_{id}} \right) = 0$$

so

$$(2.4.18) \quad \sum_{i=1}^{m} \prod_{d=1}^{\infty} d^{-n_{id}} = 1.$$

Suppose there existed $d \geq m + 1$ such that $n_{id} > 0$. But since by (2.4.18) Q is adapted to $\mathcal{A}$, then $|\mathcal{A}| \geq m+1$ which is a contradiction of the definition of $\mathcal{A}$. Therefore (2.4.10) holds. □

Theorem 2.2 establishes the Shannon entropy as a lower bound on the average charge for a questionnaire. The lower bound will be called the Shannon lower bound. Determined also is a condition in terms of the number of questions of each resolution required by the questionnaire for the Shannon lower bound to be met.

Corollary 2.2 and the proof of Theorem 2.2 give the immediate corollary:

Corollary 2.3.

A questionnaire Q has an average charge which meets the Shannon lower bound for some probability vector if and only if Q is adapted to ®.

Theorem 2.2 allows a further justification to the use of the charge log d for each question of resolution d.

Corollary 2.4.

Let the charge for a resolution d question be c(d). Then

$$(2.4.19)\quad \sum_{i=1}^{m}\sum_{d=2}^{m} p_i n_{id} c(d) \geq H(p)$$

with equality iff $p_i = \prod_{d=2}^{m} d^{-n_{id}}$ is equivalent to

$$(2.4.20)\quad c(d) = \log d.$$

Proof:

Sufficiency is established by Theorem 2.2. For necessity, consider any arbitrary integer d' greater than 2. Then for $m = d'$ and $p_i = \frac{1}{d'}$, $H(p) = \log d'$. Now there exists a valid questionnaire with

$$(2.4.21)\quad n_{id} = \begin{cases} 1 & d = d' \\ 0 & d \neq d' \end{cases} \quad \text{for all } i.$$

So

$$(2.4.22)\quad \sum_{i=1}^{d'}\sum_{d=2}^{d'} p_i n_{id} c(d) = \sum_{i=1}^{d'} \frac{1}{d'} c(d'),$$

which must equal log d'. Thus, c(d') = log d'. □

A result which parallels Theorem 2.2 to some extent is presented in Dubail (1967). There, in order to study the heterogeneous questionnaire case, a concept of generalized entropy is introduced. It is shown that this generalized entropy provides a lower bound on the average length of a questionnaire. Equality is attained as in Theorem 2.2. Now Theorem 2.2 and the result of Dubail are trivially equivalent for homogeneous questionnaires. In the simplest example of a heterogeneous questionnaire, it can be shown that Theorem 2.2 implies Dubail's result but not conversely. No more general correspondence between the two results has been shown.

An amusing exposition by R. T. Cox (1961) of the game of "Twenty Questions" suggests a different outlook on the problem of equality between entropy and average questionnaire charge. (Games of this type have been called "taxonomic games" by Tribus, Shannon, and Evans (1966).) In this particular game, only questions which allow "yes-no" responses are allowed in the quest for revelation of the "true state." That is, only resolution 2 or binary questions are allowed. Here it is generally (for example, Bartlett (1951)) considered to be good strategy to ask questions which "split the state space in half." More formally, one would seek to ask questions which partition the state space into two sets, each of equal probability. Also, in many problems, e.g., ranking or tournament problems, it is the inability to make such splits which causes the difficulty. References here include Burge (1958), Ford and Johnson (1959), and Hadian (1969).

In our more general situation, it is useful to have the following definition:

Definition 2.15.

A resolution d question is called Shannon efficient if it partitions Θ into d sets of equal probability.

In terms of this definition, we deduce the following result:

Theorem 2.3.

The Shannon lower bound is attained by a valid questionnaire Q if and only if each question is Shannon efficient.

Proof:

We shall make use of the well-known defining property of Shannon entropy, the generalized grouping axiom, which states that for a d-fold partition of Θ into $\Theta_1, \ldots, \Theta_d$:

$$(2.4.22) \quad H(p) = H(q) + \sum_{i=1}^{d} q_i H(p^{(i)})$$

where

$$(2.4.22a) \quad q = (q_1, \ldots, q_d)$$

with

$$(2.4.22b) \quad q_i = P(\Theta_i) \ (i = 1, \ldots, d),$$

and $p^{(i)}$ is the vector whose components give the conditional probability that the true state is any particular state of Θ_i given that the true state is in Θ_i.

First note that Q contains subquestionnaires, Q^{Θ_i} $(i = 1, \ldots, d)$, which resolve Θ_i $(i = 1, \ldots, d)$, respectively. Also,

$$(2.4.23) \quad E_p C(Q) = \log d + \sum_{i=1}^{d} q_i E_{p^{(i)}} C(Q^{\Theta_i}).$$

Now, by Theorem 2.2,

$$(2.4.24) \quad H(p^{(i)}) \leq E_{p^{(i)}} C(Q^{\Theta_i}) \quad (i = 1, \ldots, d).$$

Hence,

$$(2.4.25) \quad \sum_{i=1}^{d} q_i H(p^{(i)}) \leq \sum_{i=1}^{d} q_i E_{p^{(i)}} C(Q^{\Theta_i}).$$

Further,

$$(2.4.26) \quad H(p) \leq E_p C(Q).$$

Therefore equality can hold in (2.4.26) only when

$$(2.4.27) \quad H(q) = \log d.$$

But (2.4.27) requires

$$(2.4.28) \quad q_1 = q_2 = \ldots = q_d.$$

Thus the first question of Q must be Shannon efficient.

The same argument can be applied to the first question in each of the state spaces, Θ_i. Continuing this process it is shown that for equality to hold in (2.4.26) every question must be Shannon efficient.

An iterative expansion of (2.4.22) shows that the converse is valid, i.e., if every question is Shannon efficient, equality must hold in (2.4.26). □

It is well-known (see equation 2.2.4) in information theory that there exists a binary code (a homogeneous questionnaire of resolution 2) whose average length (charge) is strictly bounded

above by the Shannon entropy plus 1. Thus it is immediate that there exists a heterogeneous questionnaire with the same upper bound, i.e.,

$$\inf_{Q} E_p C(Q) < H(p) + 1. \tag{2.4.29}$$

It is of interest to consider whether the bound (2.4.29) can be improved upon in the heterogeneous case. The answer to this is, in general, no. This can be seen from the following argument: Consider the case when there are m states with $p_1 = \frac{1}{2}$ and $p_2 = \ldots = p_m = \frac{1}{2(m-1)}$. Here the Shannon lower bound is achieved by a questionnaire, Q, which has the tree representation of Figure 2.6.

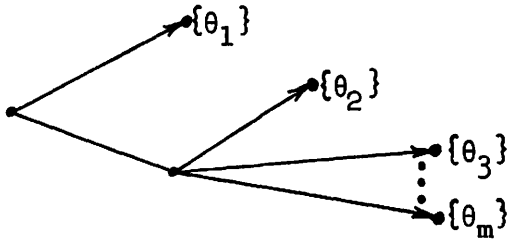


Figure 2.6

This follows from Theorem 2.3 or by direct computation. Therefore, Q has an average charge which is no higher than any other questionnaire for that probability vector, p. Further, it is quite clear and can be shown that Q will be best in this sense for any probability vector of the form

$$p_1^{\varepsilon} = 1 - (m-1)\varepsilon \quad \text{and} \quad p_2^{\varepsilon} = \ldots = p_m^{\varepsilon} = \varepsilon$$

where

$$\varepsilon \leq \frac{1}{2(m-1)}. \tag{2.4.30}$$

Then $E_{p^\epsilon} C(Q) = 1 + (m-1)\epsilon \log (m-1)$ and

$$(2.4.31) \quad H(p^\epsilon) = (1 - (m-1)\epsilon) \log \frac{1}{1-(m-1)\epsilon} + (m-1)\epsilon \log \frac{1}{\epsilon} .$$

Consider taking the limit in (2.4.30) and (2.4.31) as $\epsilon \to 0$. Obtained is

$$(2.4.32) \quad \lim_{\epsilon \to 0} E_{p^\epsilon} C(Q) = 1$$

and

$$(2.4.33) \quad \lim_{\epsilon \to 0} H(p^\epsilon) = 0.$$

Therefore,

$$(2.4.34) \quad \lim_{\epsilon \to 0} E_{p^\epsilon} C(Q) = \lim_{\epsilon \to 0} H(p^\epsilon) + 1.$$

Thus, ϵ may be chosen so that the average charge for the best questionnaire is arbitrarily close to the Shannon entropy plus 1.

2.5. Countable State Case.

It is desirable to generalize the results of the previous section to the case when the state space is countably infinite. Here one complication presents itself--the countable version of Shannon entropy is an infinite series which may diverge to $+\infty$. Nevertheless, a countable variety of noiseless coding theorem is available as

Theorem 2.4.

Suppose the state space, $\mathcal{S}$, is countably infinite. Then the average charge for a valid questionnaire is never less than the

countable Shannon entropy. If the entropy is finite, there exists a valid questionnaire with average charge not greater than the entropy plus 1. Thus we have

$$(2.5.1) \qquad H(p) \leq \inf_{Q} E_p C(Q) \leq H(p) + 1$$

where

$$(2.5.2) \qquad H(p) = - \sum_{i=1}^{\infty} p_i \log p_i .$$

Proof:

Define

$$(2.5.3) \qquad H^M(p) = - \sum_{i=1}^{M} p_i \log p_i - \delta_M \log \delta_M$$

where

$$(2.5.3a) \qquad \delta_M = p_{M+1} + p_{M+2} + \ldots .$$

We note that

$$(2.5.4) \qquad H^M(p) \leq \inf_{Q_M} E_p C(Q_M) < H^M(p) + 1$$

where Q_M denotes any questionnaire determining the true state among $\{\theta_1\}, \ldots, \{\theta_M\}, \{\theta_{M+1}, \ldots\}$. Now

$$(2.5.5) \qquad - \sum_{i=1}^{M} p_i \log p_i \to H(p) \text{ (finite or } +\infty) \text{ as } M \to \infty,$$

and

$$(2.5.6) \qquad \delta_M \log \delta_M \to 0 \text{ as } M \to \infty, \text{ so that}$$

$$(2.5.7) \qquad H^M(p) \to H(p) \text{ as } M \to \infty.$$

Let us first suppose that $H(p)$ is finite. Then by taking limits as $M \to \infty$ in (2.5.4), we obtain

$$(2.5.8) \qquad H(p) \leq \gamma \leq H(p) + 1$$

where $\gamma = \lim_{M \to \infty} \inf_{Q_M} E_p C(Q_M)$. Now we wish to establish the existence of a questionnaire resolving $\circledast$ with an average charge arbitrarily close to γ. First define

$$(2.5.9) \qquad \delta_M^{(k)} = (\tfrac{1}{2})^{k-1} \delta_M \quad (k = 1, 2, \ldots),$$

and choose positive integers, $M^{(k)}$ $(k = 1, 2, \ldots)$, such that

$$(2.5.10) \quad \text{(i)} \quad M = M^{(1)},$$

$$\text{(ii)} \quad M^{(k)} < M^{(k+1)},$$

$$\text{(iii)} \quad p_{M^{(k)}+1} + \ldots + p_{M^{(k+1)}} \leq \delta_M^{(k)}$$

and

$$\text{(iv)} \quad p_{M^{(k)}+1} + \ldots + p_{M^{(k+1)}+1} > \delta_M^{(k)}.$$

Since we are dealing with finite sets in each case, we can let Q_M^*, $Q^*_{M^{(1)}, M^{(2)}}$, $Q^*_{M^{(2)}, M^{(3)}}$, ... be the best questionnaires for determining the true state among $\{\theta_1, \ldots, \theta_{M^{(1)}}\}$, $\{\theta_{M^{(1)}+1}, \ldots, \theta_{M^{(2)}}\}$, $\{\theta_{M^{(2)}+1}, \ldots, \theta_{M^{(3)}}\}, \ldots$, respectively. Then Q_M^* can be extended to determine the true state in all of $\circledast$ in the following manner:

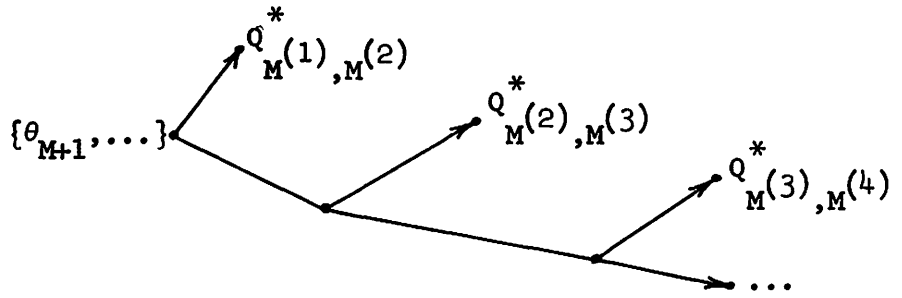


Figure 2.7

If we call this extended questionnaire, $Q_{(M)}^*$, we note that the additional charge for $Q_{(M)}^*$ over Q_M^* is

$$(2.5.11) \qquad \sum_{k=1}^{\infty} q_k(M)(k + E_p C(Q^*_{M^{(k)},M^{(k+1)}}))$$

where

$$(2.5.12) \qquad q_k(M) = (p_{M^{(k)}+1} + \ldots + p_{M^{(k+1)}}) \quad (k = 1, 2, \ldots).$$

Now we note that expression (2.5.11) can be no bigger than

$$(2.5.13) \qquad \sum_{k=1}^{\infty} k\delta_M^{(k)} + \sum_{k=1}^{\infty} q_k(M)(- \sum_{i=M^{(k)}+1}^{M^{(k+1)}} \frac{p_i}{q_k(M)} \log \frac{p_i}{q_k(M)}) + \sum_{k=1}^{\infty} q_k(M).$$

This result follows from (2.4.29) and (2.5.10). Then by (2.5.9) expression (2.5.13) may be rewritten as

$$(2.5.14) \qquad \delta_M(1 + \sum_{k=1}^{\infty} k(\tfrac{1}{2})^{k-1}) + (- \sum_{k=1}^{\infty} \sum_{i=M^{(k)}+1}^{M^{(k+1)}} p_i \log \frac{p_i}{q_k(M)}),$$

which in turn equals

$$(2.5.15) \qquad 5\delta_M + (- \sum_{i=M}^{\infty} p_i \log p_i) + \sum_{k=1}^{\infty} \sum_{i=M^{(k)}+1}^{M^{(k+1)}} p_i \log q_k(M).$$

Now as $M \to \infty$, the first term of (2.5.15) clearly goes to zero while the second term must also go to zero since it is the tail of a convergent series ($H(p)$ is assumed finite). Consider the third term of (2.5.15). We make use of the basic inequality of information theory (2.4.11) to write

$$(2.5.16) \qquad \sum_{k=1}^{\infty} \sum_{i=M^{(k)}+1}^{M^{(k+1)}} \frac{p_i}{\delta_M} \log \frac{q_k(M)}{\delta_M} \leq \sum_{k=1}^{\infty} \sum_{i=M^{(k)}+1}^{M^{(k+1)}} \frac{p_i}{\delta_M} \log \frac{p_i}{\delta_M}.$$

But this is equivalent to

$$(2.5.17)\qquad \sum_{k=1}^{\infty} \sum_{i=M^{(k)}+1}^{M^{(k+1)}} p_i \log q_k(M) \leq \sum_{i=M}^{\infty} p_i \log p_i.$$

Again using the fact that the entropy has been assumed finite, we have the third term of (2.5.15) going to zero as $M \to \infty$.

Therefore, we have demonstrated the existence of a positive integer M so that the questionnaire $Q_{(M)}^{*}$ has an average charge arbitrarily close to $\lim_{M\to\infty} \inf_{Q_M} E_p C(Q_M) = \gamma$. Hence from (2.5.8) we have established the conclusion, (2.5.1), to our theorem.

Suppose now that $H(p) = +\infty$. We have that

$$(2.5.18)\qquad H^M(p) \leq \inf_{Q_M} E_p C(Q_M) < \inf_{Q} E_p C(Q)$$

where Q is any questionnaire to determine $\theta \in \Theta$. But

$$(2.5.19)\qquad H^M(p) \uparrow H(p) = +\infty;$$

so

$$(2.5.20)\qquad \inf_{Q} E_p C(Q) = +\infty. \ \square$$

Consider two specific examples in the countable state case.

i. Geometric probability case. Suppose the probabilities are given by

$$(2.5.21)\qquad p_i = p^{i-1}q \ (i = 1, 2, \ldots;\ 0 < p < 1;\ q = 1 - p)$$

so that

$$(2.5.22)\qquad H(p) = -\sum_{i=1}^{\infty} p_i \log p_i = -\left\{ \frac{p}{q} \log p + \log q \right\}.$$

Note that for $p = q = \frac{1}{2}$, we have $H(p) = 2$ which is the limit of truncated entropies and is attainable by a questionnaire.

ii. Poisson probability case. Here we let

$$(2.5.23) \quad p_i = \frac{\lambda^i e^{-\lambda}}{i!} \quad (i = 0, 1, \ldots; \lambda \geq 0).$$

Then

$$(2.5.24) \quad H(p) = -\lambda \log \lambda + \lambda \log e + \sum_{i=0}^{\infty} \frac{\lambda^i e^{-\lambda}}{i!} \log i! \ .$$

This entropy is finite since the series converges by the ratio test.

2.6. Continuous State Case.

Conceptually, it is often useful to formulate problems so that the state space, Θ, is continuous. Then the true state may be treated as the realization of a continuous random variable, θ, with range, say, the real line, R^1. Note that considerably more structure has been imposed on the state space than in previous cases. Analogous to the Shannon entropy in the finite or countable state case, we can define the differential entropy or Wiener entropy (Wiener, 1948) as

$$(2.6.1) \quad H(\theta) = -\int p(\theta) \log p(\theta) d\theta,$$

where $p(\cdot)$ is a probability density function (with respect to Lebesque measure) of the random variable, θ.

As has been pointed out by Kolmogorov (1965) and Moran (1951), the differential entropy is fundamentally of a different character than the entropy that has been previously considered here. It

does not admit a direct combinatorial interpretation in terms of questionnaires.

Further, from a practical point of view, it is not necessary to know the exact state, θ, which is operational in a particular situation, but only a value which is "sufficiently close" to θ. Therefore, we will assume that Θ is a metric space with a metric, ρ, in order to deal with this mathematically. Kolmogorov (1956), again, provides the clue on how to handle this problem by introducing the concept of ϵ-entropy. We will admit a second random variable, η, jointly measurable with θ. A basic requirement on η is that, for a preassigned ϵ,

$$(2.6.2) \qquad P(\rho(\eta - \theta) \leq \epsilon) = 1.$$

The interpretation is then clearcut. If we know the value of η, we shall, almost surely, know the value of θ within ϵ. In order to be back in our familiar combinatorial framework, where questionnaires are possible, we shall require η to be at most countably valued.

For $p(\cdot)$ fixed, we now define W_ϵ to be the set of joint probability measures of θ and η satisfying (2.6.2) where η is discrete. We further define

$$(2.6.3) \qquad H_\epsilon(\theta) = \inf_{\substack{\eta \\ (\theta,\eta)\in W_\epsilon}} I(\theta, \eta),$$

where $I(\theta, \eta)$ is the usual mutual information in η about θ (or vice-versa). The right hand side of (2.6.3) may be written in the alternative forms

(2.6.4) $\inf_{\eta} [H(\theta) - EH(\theta|\eta)] = H(\theta) - \inf_{\eta} EH(\theta|\eta)$

and

(2.6.5) $\inf_{\eta} [H(\eta) - EH(\eta|\theta)]$.

If we now require that given θ, η is almost surely constant, we will have

(2.6.6) $H(\eta|\theta = \theta_0) = 0$ for each θ_0.

Hence

(2.6.7) $\inf_{\eta} I(\theta, \eta) = \inf_{\eta} H(\eta)$.

This corresponds to some extent to the intuitive feeling that a "cheap" η (in terms of questionnaire charges) which meets the ϵ-requirements is one in which $I(\theta, \eta)$ is small.

3. Lattice Questionnaires.

3.1. Origin.

A lattice questionnaire has a graph representation showing more than one path leading to some terminal node. Thus there is at least one question which does not partition the set on which it acts. A lattice questionnaire adapted to $\Theta = \{\theta_1, \theta_2, \theta_3, \theta_4\}$ might appear as in Figure 2.8 below:

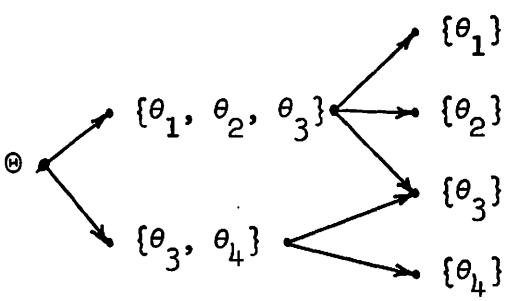


Figure 2.8

This class of questionnaires is introduced by Picard (1965, 1968).

The following examples are illustrative of situations in which lattice questionnaires might arise:

a. Addition algorithm.

Consider $x, y \in \{0, 1, 2, 3\}$. Write x and y in binary notation and use the usual addition algorithm to obtain $x + y$. The results of the steps of the algorithm can be shown by the lattice questionnaire in Figure 2.9 below:

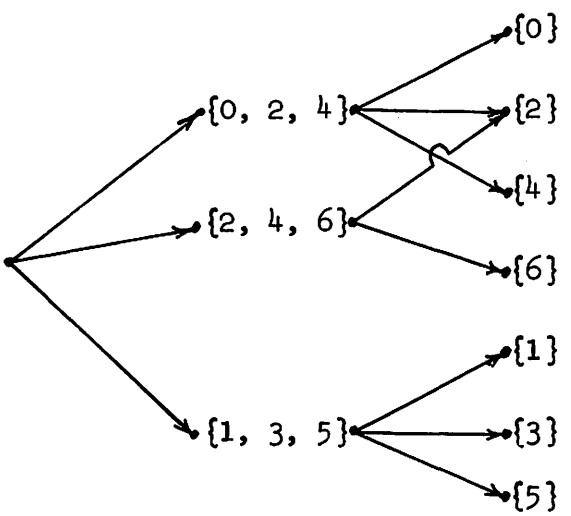


Figure 2.9

b. Uniform distributions.

Suppose a random variable is available which is distributed either uniform on $(0, 2)$ or uniform on $(1, 3)$. We denote the first case as state θ_1 and the second as state θ_2. If we continue random sampling until the true state is determined, the lattice questionnaire will have the form:

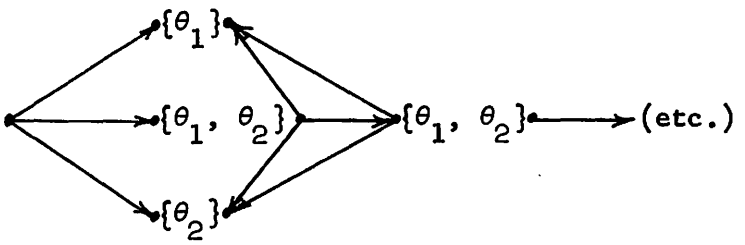


Figure 2.10

Here the true state would be determined only with probability one.

c. Taxonomy.

Osborne (1963) discusses the use of lattice questionnaires for biological classification purposes when there is some possibility of receiving the "wrong" answer to a question. He refers to these questionnaires as "reticulated keys" and examines their effectiveness under a restricted model for the errors.

d. Discrete search for a maximum.

Sought is an optimal setting of controllable variables. The response is assumed to be deterministic and unimodal. In this context, discussion of various single variable search schemes is contained in Wilde (1964). For illustration, consider a very special case in which 4 settings of the controllable variables are possible. Using the property of unimodality,there are 8 possible configurations of responses. The actual configuration is not of interest, but only which setting produces the highest response, i.e., the location of the mode. As an example, there is the search scheme represented in Figure 2.11 below, with the positions evaluated indicated in parentheses.

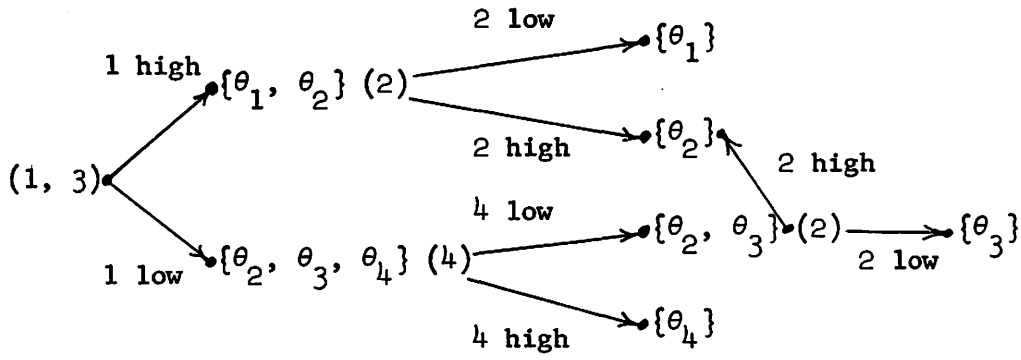


Figure 2.11

This is a lattice questionnaire.

3.2. Charging Scheme for a Lattice Questionnaire.

There is no obvious charging scheme for a lattice questionnaire. This section will propose a charge structure in this case which is intuitively satisfying but contains indeterminate elements. This flexibility will be exploited to obtain a charging scheme which will extend the validity of the Shannon lower bound result to arbitrary lattice questionnaires.

To every lattice questionnaire, there corresponds a unique arborescence questionnaire. Every path to a terminal node is allowed to lead to a distinct terminal node in the tree. Thus a number of artificial states are introduced, say m_i, for the i^{th} state. As an example, the lattice questionnaire given in Figure 2.12 below

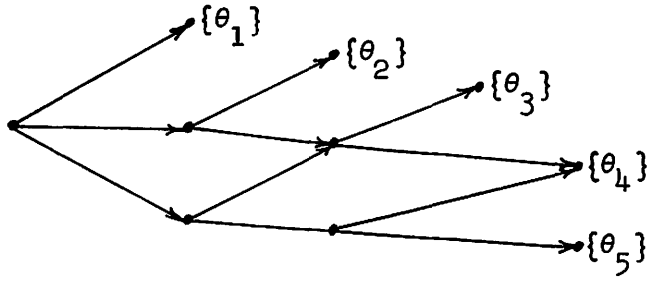


Figure 2.12

has the corresponding tree representation shown in Figure 2.13.

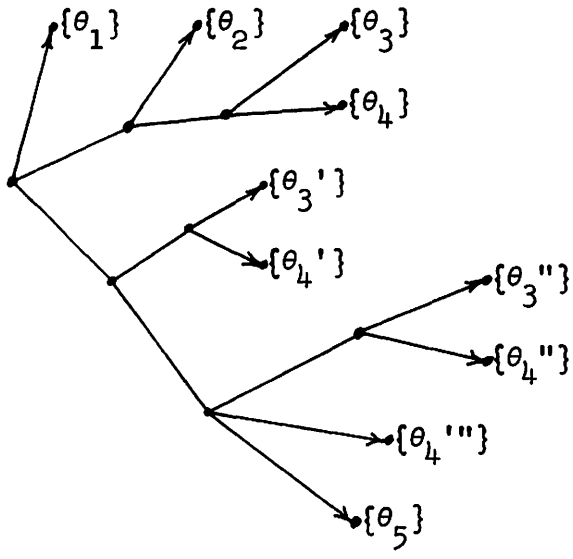


Figure 2.13

This section will concern itself with lattice questionnaires which are adapted to a finite state space $\Theta = \{\theta_1, \ldots, \theta_m\}$. Now suppose probability p_i is allotted to the i^{th} state. Then it is reasonable to assess the average charge of the arborescence questionnaire as

$$\sum_{i=1}^{m} \sum_{r=1}^{m_i} p_{ir}\, \varphi_{ir}$$

where the p_{ir} are restricted by

$$\sum_{r=1}^{m_i} p_{ir} = p_i$$

but for the time being are otherwise arbitrary. The $\{\varphi_{ir}\}$ are determined as the usual charges (based on $\log d$ for each resolution d question) along the r^{th} path to a terminal node

associated with state θ_i. Based on this representation, a charging scheme is sought for the lattice questionnaire.

The basic desideratum employed is the maintenance of the Shannon lower bound (Theorem 2.2) including the possibility of equality between average charge and entropy for some value of the probability vector. The simplest (but not the only) approach to this is to make a modification in the usual charge for the first question of the questionnaire.

Suppose the first question has an image family containing d sets. Then a corrected charge c (less than $\log_2 d$) is sought for the first question. The object is to obtain a modified charge structure which will give the Shannon lower bound. First, the following definition is useful:

Definition 2.16.

A probability vector $p^{(0)}$ and a charge structure $\{\varphi_{ir}\}$ are called compatible if

$$(3.2.1) \qquad -\log \frac{p_j^{(0)}}{p_k^{(0)}} = \varphi_j - \varphi_k \quad \text{for all} \quad j,k = 1,\ldots, m$$

where

$$(3.2.2) \qquad \varphi_i = \min_r \varphi_{ir} .$$

Note that given $\{\varphi_{ir}\}$ there exists a unique compatible vector, $p^{(0)}$.

Theorem 2.5.

Let the valid lattice questionnaire Q have the associated charge structure $\{\varphi_{ir}\}$. Let $p^{(0)}$ be the compatible probability

vector and let c be determined by

$$(3.2.3) \qquad c = (\log d - \sum_{i=1}^{m} p_i^{(0)} \varphi_i) + H(p^{(0)}),$$

where d is the resolution of the first question. Then a charge for determining the i^{th} state of

$$(\varphi_i - \log d) + c$$

will maintain the Shannon lower bound for every p and the questionnaire will be efficient at $p^{(0)}$.

Before proceeding to the proof of the theorem, consider some implications of this charging scheme. First, since by Theorem 2.2,

$$(3.2.4) \qquad \sum_{i=1}^{m} p_i^{(0)} \varphi_i \geq H(p^{(0)}),$$

it follows by the definition of c, equation (3.2.3), that

$$(3.2.5) \qquad c \leq \log d.$$

Thus c is a discounted charge, where the amount of discount depends on the nature of the lattice questionnaire.

Second, if, in fact, the questionnaire has arborescence representation, then

$$(3.2.6) \qquad -\log p_j^{(0)} = \varphi_j$$

follows from equation (3.2.1) and Corollary 2.3. In this case equality is obtained in (3.2.4). Thus <u>a fortiori</u> there is no discount in the charge for the first question.

Proof of Theorem 2.5:

Let p be an arbitrary probability vector. Then it is sufficient to show that

$$(3.2.7) \qquad \sum_{i=1}^{m} p_i(c - \log d + \varphi_i) \geq H(p),$$

with equality iff $p = p^{(0)}$.

This statement (3.2.7) is equivalent to

$$(3.2.8) \qquad \sum_{i=1}^{m} (p_i - p_i^{(0)})\varphi_i \geq H(p) - H(p^{(0)}),$$

by the definition of c. Now since $p^{(0)}$ is compatible with the charge structure,

$$(3.2.9) \qquad \log p_i^{(0)} + \varphi_i = \log p_k^{(0)} + \varphi_k \quad (i, k = 1, \ldots, m).$$

Now (3.2.9) implies

$$(3.2.10) \qquad \sum_{i=1}^{m} p_i \log p_i^{(0)} + \sum_{i=1}^{m} p_i\varphi_i = \log p_k^{(0)} + \varphi_k \quad (k = 1, \ldots, m),$$

which in turn implies

$$(3.2.11) \qquad \sum_{i=1}^{m} p_i \log p_i^{(0)} + \sum_{i=1}^{m} p_i\varphi_i = \sum_{k=1}^{m} p_k^{(0)} \log p_k^{(0)} + \sum_{k=1}^{m} p_k^{(0)}\varphi_k.$$

But (3.2.11) implies that

$$(3.2.12) \qquad \sum_{i=1}^{m} (p_i - p_i^{(0)})\varphi_i = - \sum_{i=1}^{m} p_i \log p_i^{(0)} - H(p^{(0)}).$$

Now

$$(3.2.13) \qquad - \sum_{i=1}^{m} p_i \log p_i^{(0)} - H(p^{(0)}) \geq H(p) - H(p^{(0)})$$

by the fundamental inequality of information theory. Therefore, the validity of (3.2.8) is affirmed and the theorem is proved. □

3.3. Characterization of Shannon Entropy.

Let C denote any class of questionnaires (arborescence or lattice) and let $C(Q)$ be the charge described in the preceeding section (this includes the charge for arborescence questionnaires described in Section 2.3 as a special case).

The function K_C defined by

$$(3.3.1) \qquad K_C(p) = \inf_{Q \in C} E_p C(Q)$$

is a continuous, concave, piecewise linear function which has points of tangency with the Shannon entropy function, $H(p)$. The number of points of tangency depends on the size of the class C. For each p which admits an efficient questionnaire, there is a point of tangency.

A natural question is whether an enlargement from the arborescence class of questionnaires to include the lattice class will provide a characterization of the Shannon entropy in terms of average charges for a questionnaire. The answer to this is affirmative.

Theorem 2.6.

If $\mathcal{L}$ is the class of all questionnaires (arborescence or lattice),

$$(3.3.2) \qquad H(p) = K_{\mathcal{L}}(p).$$

Proof:

Since $K_{\mathcal{L}}(p)$ is continuous and concave, it is sufficient to demonstrate that given a probability vector, p, with rational components, there exists a lattice questionnaire, Q, which is efficient. Now efficiency of Q requires that

$$(3.3.3) \qquad -\log \frac{p_j}{p_k} = \varphi_j - \varphi_k = \sum_{d=2}^{\infty} (n_{jd} - n_{kd}) \log d, \quad (j,k = 1,\ldots, m)$$

where it is recalled that n_{id} is the number of resolution d questions required by Q to determine θ_i. Then

$$(3.3.4) \qquad \frac{p_k}{p_j} = \prod_{d=2}^{m} d^{(n_{jd} - n_{kd})} \quad (j, k = 1,\ldots, m).$$

If p has rational components, then $\frac{p_k}{p_j}$ may be expressed as $\frac{r}{s}$ for some integral r and s. Therefore, for the sake of definiteness, the prime factorization theorem may be invoked to specify a unique $\{n_{jd} - n_{kd}\}$. This gives a consistent set of equations for $\{n_{kd}\}$. If $n_d = (n_{1d},\ldots, n_{md})$, the solutions will be of the form

$$(3.3.5) \qquad n_d = b + je \quad (j = 0, 1, 2,\ldots)$$

where e is the m-vector with each component one.

Then the lattice questionnaire, Q, can be constructed so as to satisfy the constraints. First ask an initial resolution m question. An arborescence questionnaire constructed from this "base" will then have m "primary" branches. The i^{th} branch will lead to a subtree reserved for determining the i^{th} state $(i = 1,\ldots, m)$. From the i^{th} node of the initial question,

construct the subtree so as to have the required number of questions of each resolution needed to determine θ_i. Identify one of the terminal nodes having the requisite number of questions of each resolution with θ_i. This process will leave some of the terminal nodes unidentified. Now in a cyclical manner, establish an edge from each such terminal node in the $(i + 1)$st subtree to the i^{th} node of the initial question ($i = 1,\ldots, m-1$); direct those in the first subtree to the m^{th} node of the initial question. The result of this procedure is a lattice questionnaire having the required charge structure for efficiency at p.

As an example of this construction, consider the case when $m = 2$ and $p_1 = .6$. Then

$$\frac{p_1}{p_2} = \frac{.6}{.4} = \frac{3}{2} .$$

Thus, using the prime factorization form,

$$n_{13} - n_{23} = 1$$

and

$$n_{12} - n_{22} = -1$$

while all other differences are zero. Therefore, it is possible to let $n_{13} = 1$, $n_{23} = 0$, $n_{12} = 1$, $n_{22} = 2$. The lattice questionnaire is then constructed in stages as follows:

Stage 1.

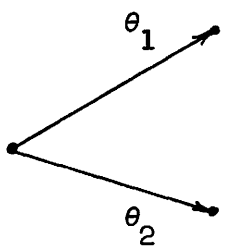


Figure 2.14a

Stage 2.

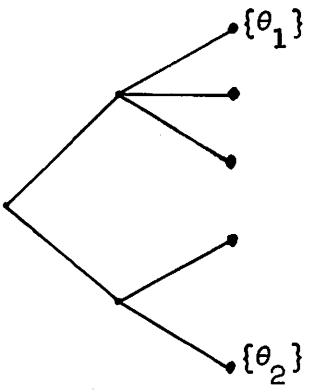


Figure 2.14b

Stage 3.

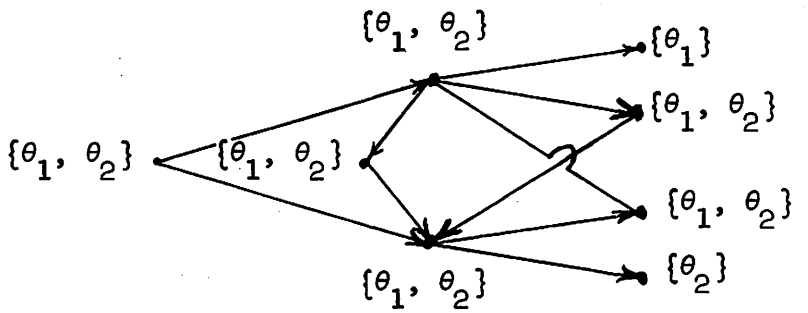


Figure 2.14c

This is a lattice questionnaire which is efficient for $p = (p_1, p_2) = (.6, .4)$. □

Chapter III

Optimal Heterogeneous Questionnaires

1. Motivation.

This chapter examines arborescence questionnaires exclusively. It is assumed that a charge φ_i is assessed when the state θ_i is found to be true. This charge φ_i is a function of the questionnaire Q which has been employed; thus it might be written $\varphi_i(Q)$. The charging scheme is at this level of generality for all of the results of this chapter unless it is necessary to restrict the class of charging schemes. The restrictions may take the form of making the charge depend only on the resolution of the questions used or as far as charging the specific amount $\log_2 d$ for each resolution d questions used. This last charging scheme is of course the one discussed in Chapter II.

The basic aim of the chapter is to attempt to find a "best" questionnaire. It is therefore useful to find essentially complete classes of questionnaires having readily identifiable characteristics. Further, it is desirable to have an algorithm for actually finding a "best" questionnaire. Such an algorithm will be developed in this chapter.

2. Comparison of Questionnaire Charges.

Suppose now that a charging scheme, i.e., a method of assessing charges for any questionnaire Q has been fixed. Thus for any questionnaire Q the charge for determining that state θ_i is true, the quantity $\varphi_i(Q)$, will be known.

We will assume that the individual designing the questionnaire, who might be called a questioner, wants to find a questionnaire which has a small average charge. The average would be computed according to the questioner's prior distribution on the state space.

Definition 3.1.

Let a probability vector, p, on Θ be given. Then the average charge for resolution of Θ by Q is defined by

$$E_p C(Q) = \sum_{i=1}^{|\Theta|} p_i \varphi_i. \tag{2.1}$$

If Θ is finite, we will write

$$|\Theta| = m. \tag{2.2}$$

In the sequel, it is assumed with no loss of generality that the states of Θ have been renumbered so that $p_1 \geq p_2 \geq \ldots$.

The concern of this section will be with the comparison of questionnaires in terms of their average charges. Thus if Q_1 and Q_2 are valid questionnaires, and p is a probability vector with $p_1 \geq p_2 \geq \ldots \geq p_m$, the following definitions prove helpful:

Definition 3.2.

Q_1 and Q_2 are called charge-equivalent at p iff

$$E_p C(Q_1) = E_p C(Q_2). \tag{2.3}$$

In this case we write

$$Q_1 \stackrel{p}{\sim} Q_2. \tag{2.4}$$

Definition 3.3.

Q_1 and Q_2 are called charge-equivalent iff

(2.5) $\quad Q_1 \tilde{p} Q_2$ for all p.

We shall denote this by

(2.6) $\quad Q_1 \sim Q_2$.

Definition 3.4.

Q_1 is said to be preferred at p to Q_2 iff

(2.7) $\quad E_p C(Q_1) \leq E_p C(Q_2)$.

This will be written

(2.8) $\quad Q_1 \underset{p}{\succeq} Q_2$.

And then analogously to Definition 3.3 we have

Definition 3.5.

Q_1 is said to be preferred to Q_2 iff

(2.9) $\quad Q_1 \underset{p}{\succeq} Q_2$ for all p.

This is denoted by

(2.10) $\quad Q_1 \succ Q_2$.

Now Definitions 3.2 and 3.4 provide a complete ordering of the valid questionnaires on ⊛ with respect to a fixed probability vector p. Definitions 3.3 and 3.5 yield a partial ordering of the valid questionnaires on ⊛.

Much of the previous work on questionnaire theory has been done with reference to the length of a questionnaire:

Definition 3.6.

The (average) length of Q is

$$E_p L(Q) = \sum_{i=1}^{|\Theta|} p_i k_i \tag{2.11}$$

where state θ_i is determined at stage k_i.

The desire to study the selection of a questionnaire on the basis of charges suggests the following definition:

Definition 3.7.

A valid questionnaire Q^* has minimum average charge with respect to p if

$$Q^* \underset{p}{\geq} Q \quad \text{for all} \quad Q$$

or

$$E_p C(Q^*) = \inf_Q E_p C(Q) \tag{2.12}$$

where Q is a valid questionnaire.

Definition 3.8.

A valid questionnaire Q^* is tight in a class $\mathcal{G}$ if

$$E_p L(Q^*) = \inf_{Q \in \mathcal{G}} E_p L(Q). \tag{2.13}$$

Definition 3.9.

A valid questionnaire Q^* is optimal with respect to p if it is tight in the class of questionnaires with minimum average charge with respect to p.

Definition 3.10.

A charging scheme $\{\varphi_i\}$ is said to be question based if a charge is made for each question used to determine θ_i as the true state; a question based charging scheme is called resolution increasing if the question charge, c, is a function only of the question resolution, d, and

(2.14) $\quad c(d) > c(d')$ if $d > d'$.

3. The Admissible Class of Questionnaires.

The task set is to resolve a finite state space Θ while sustaining the minimum average charge. Except in the trivial case when $m = 2$, there will be no one questionnaire which will be preferred uniformly in p. Nevertheless, for each fixed p, there will exist at least one questionnaire with minimum charge. The class of all such questionnaires over p is the one of interest.

Definition 3.11.

The class $\mathcal{Q}_m$ is the admissible class of questionnaires iff for each valid Q, $Q \in \mathcal{Q}_m$ is equivalent to the nonexistence of a valid Q^* with $Q^* > Q$ and

$$E_p C(Q^*) < E_p C(Q)$$

for some probability vector p with $p_1 \geq p_2 \geq \ldots$. Note that the restriction on p only amounts to a relabelling of the states.

Theorem 3.1.

Suppose the charging scheme is resolution increasing. Let (Θ, ω, Q) be a questionnaire. If Q is not adapted to Θ, then

Q is not admissible.

Proof:

Let $\tilde{Q}$ be the questionnaire which deletes all branches leading to nodes in ω in Q. Then $\tilde{Q} > Q$ with strict dominance for all p with nonzero components. □

4. Optimal Questionnaires when $m = 3$ and $m = 4$.

To give concreteness to the discussion of optimal questionnaires, it is useful to examine some special cases which are easy to work out explicitly. Suppose that the $\log_2 d$ charging scheme is used. In the $m = 2$ case there is clearly only one questionnaire of interest--the one calling for a single binary question. When $|\Theta| = m$ is larger, the number of questionnaires that must be examined can be considerably reduced by ordering the states so that $p_1 \geq p_2 \geq \ldots \geq p_m$. Then in the $m = 3$ case, there are two tight admissible questionnaires of interest. They have the arborescence representation given in Figure 3.1.

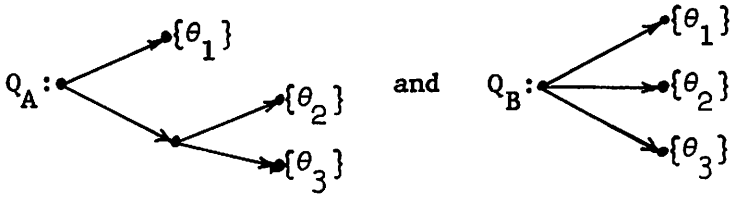


Figure 3.1

Their average charges are

$$E_p C(Q_A) = 2 - p_1 \tag{4.1}$$

and

$$E_p C(Q_B) = \log_2 3. \tag{4.2}$$

Thus $Q_A \geq_p Q_B$ provided $p_1 \geq 2 - \log_2 3 \approx .42$. The regions of optimality can be shown on an interval for p_1, or more suggestively by graphing $\inf_Q E_p C(Q)$ and $H(p)$ (where $p_2 = p_3 = \frac{1}{2}(1 - p_1)$) against p_1.

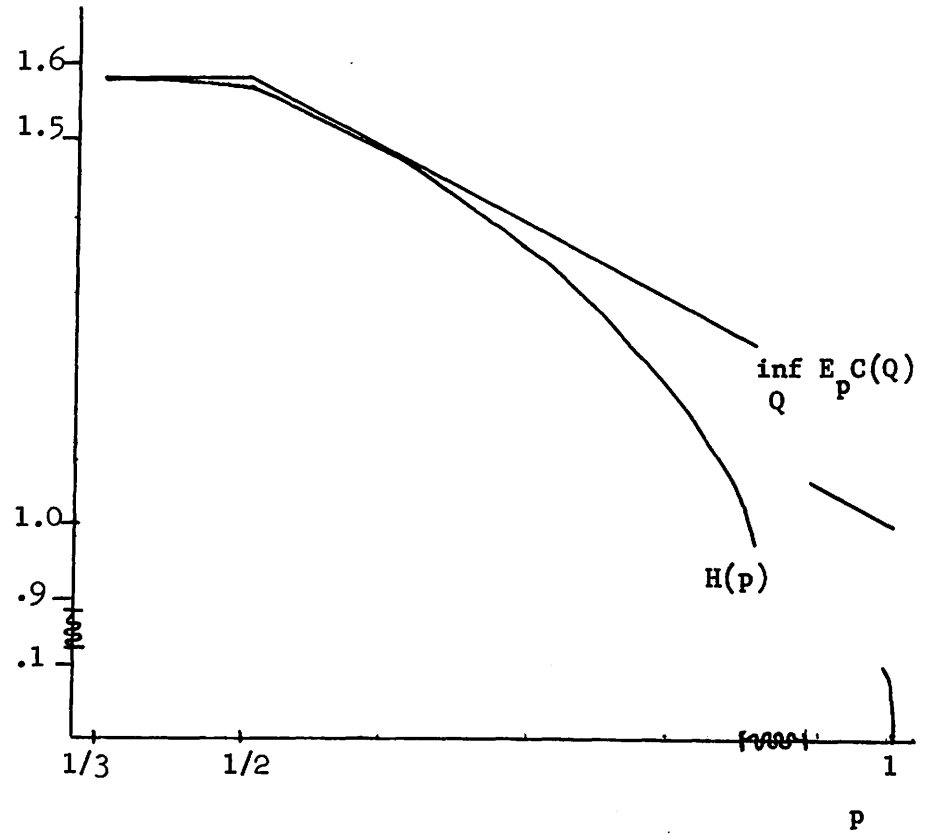
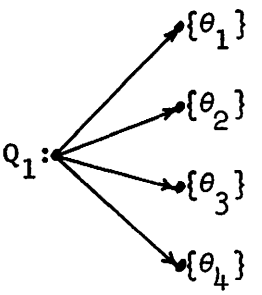


Figure 3.2

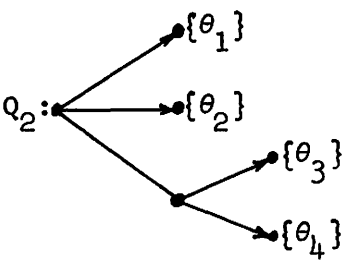
Notice that $\inf_Q E_p C(Q)$ is tangent to $H(p)$ at the two points $(p_1, H(p_1)) = (1/3, \log 3)$ and $(p_1, H(p_1)) = (1/2, 3/2)$. Naturally these two points correspond to $p = (1/3, 1/3, 1/3)$ and $p = (1/2, 1/4, 1/4)$, respectively, where Shannon efficient questionnaires are possible.

Going one step beyond the $m = 3$ case, it is possible to examine in detail the $m = 4$ case. Here an essentially complete class of tight questionnaires has precisely 4 elements. Their arborescence representation and average charge are shown below:



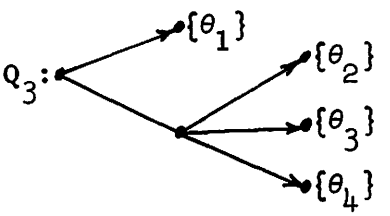
$$E_p C(Q_1) = 2$$

Figure 3.3a



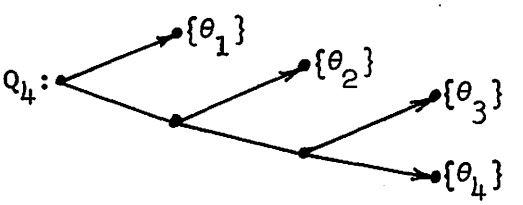
$$E_p C(Q_2) = 1 + \log 3 - (p_1 + p_2)$$

Figure 3.3b



$$E_p C(Q_3) = 1 + (1 - p_1)\log 3$$

Figure 3.3c



$$E_p C(Q_4) = 3 - 2p_1 - p_2$$

Figure 3.3d

Notice that the average charge for any questionnaire in this class is a function of p_1 and p_2 (remember $p_1 \geq p_2 \geq \ldots$ has been assumed). By comparison of these average charges, regions in the permissible subset of the (p_1, p_2) plane can be found where each of these questionnaires is optimal. These regions are shown in Figure 3.4 below with the points where Shannon efficient questionnaires are possible located.

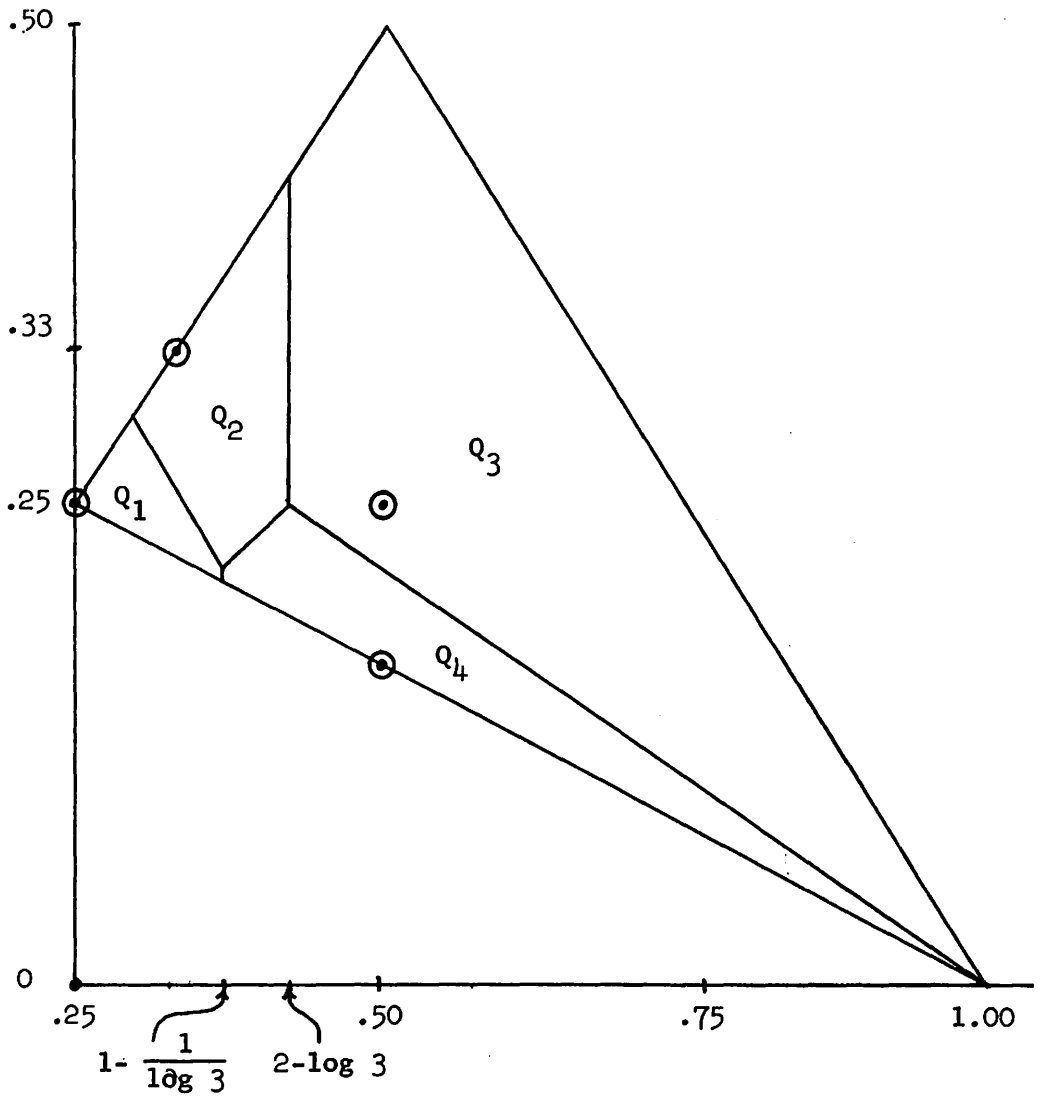


Figure 3.4

5. The Interchange Operation on a Questionnaire.

An existing questionnaire can sometimes be improved in terms of its expected charge property by interchanging the role of two states in the questionnaire. This can be considered as an operator on the questionnaire Q.

Definition 3.12.

The questionnaire $I_r^s Q$ is formed by changing the domain of Q from subsets of $\{\theta_1, \ldots, \theta_r, \ldots, \theta_s, \ldots\}$ to subsets of $\{\theta_1, \ldots, \theta_s, \ldots, \theta_r, \ldots\}$. The operator $I_r^s Q$ is said to interchange θ_r and θ_s. An interchange of θ_r and θ_s gives an expected charge which can be expressed by

$$(5.1) \qquad E_p C(I_r^s Q) = \sum_{i \neq r,s}^{|\Theta|} p_i \sum_{j=1}^{k_i} \log |QQ^{-j}\{\theta_i\}| + p_r \sum_{j=1}^{k_s} \log |QQ^{-j}\{\theta_s\}| + p_s \sum_{j=1}^{k_r} \log |QQ^{-j}\{\theta_r\}|.$$

Then (valid in fact for any charging scheme)

$$(5.2) \qquad E_p C(Q) - E_p C(I_r^s Q) = (p_r - p_s)(\varphi_s - \varphi_r).$$

6. Structure of Optimal Questionnaires.

The structure of optimal questionnaires, particularly in terms of their charges, is investigated in the theorems below. (It is assumed that the states have been renumbered so that $p_1 \geq p_2 \geq \ldots \geq p_m$; note that the results are independent of the specific log d charging scheme.)

Theorem 3.2.

Suppose $p_1 > p_2 > \ldots > p_m$. Then every admissible question-

naire has $\varphi_1 \leq \varphi_2 \leq \ldots \leq \varphi_m$. In general, there exists an essentially complete class of questionnaires with $\varphi_1 \leq \varphi_2 \leq \ldots \leq \varphi_m$.

Proof:

Let Q be an admissible questionnaire. Choose $r < s$ and suppose $p_r > p_s$. For the first conclusion of the theorem, it is sufficient to show $\varphi_r \leq \varphi_s$. Suppose, on the contrary, that $\varphi_r > \varphi_s$. Let $\tilde{Q}$ be the questionnaire which interchanges the role of θ_r and θ_s in Q. Then the following statements are equivalent:

$$(6.1) \qquad \varphi_r > \varphi_s$$

$$(6.2) \qquad (p_r - p_s)\varphi_r > (p_r - p_s)\varphi_s$$

$$(6.3) \qquad p_r\varphi_r + p_s\varphi_s > p_s\varphi_r + p_r\varphi_s$$

$$(6.4) \qquad E_p C(Q) > E_p C(\tilde{Q}).$$

But this last statement (6.4) contradicts the admissibility of Q.

To demonstrate the second conclusion to the theorem, note that if Q' is an arbitrary valid questionnaire with $\varphi_r > \varphi_s$ for $r < s$, the questionnaire Q^* which interchanges the role of θ_r and θ_s is preferred to Q' (as above with possible equality in (6.2) and thereafter). □

Theorem 3.3.

Suppose $|\Theta| = m < \infty$, and the charging scheme is question based. The set of questionnaires, $\mathcal{B}$, whose average charge depends on p only

through $p_1, p_2, \ldots, p_{m-2}$ forms an essentially complete class.

Proof:

Let Q be an admissible questionnaire whose charge depends on p_{m-1} and p_m, i.e., for some fixed $p^{(m-2)} = (p_1, \ldots, p_{m-2})$, there exists (p_{m-1}, p_m) and (p'_{m-1}, p'_m) such that

$$(6.5) \qquad E_{(p^{(m-2)},(p_{m-1},p_m))}^{C(Q)} \neq E_{(p^{(m-2)},(p'_{m-1},\, p'_m))}^{C(Q)}.$$

Then it must be shown that for each p there exists $Q_p \in \mathcal{B}$ such that $Q_p \underset{p}{\succeq} Q$. First it is noted that (6.5) implies

$$(6.6) \qquad p_{m-1}\varphi_{m-1} + p_m\varphi_m \neq p'_{m-1}\varphi_{m-1} + p'_m\varphi_m.$$

But θ_{m-1} and θ_m are not offspring of the same node (requiring $\varphi_{m-1} = \varphi_m$) since

$$(6.7) \qquad p_{m-1} + p_m = p'_{m-1} + p'_m .$$

Therefore there exists a state $\theta_\ell^* \in \{\theta_1, \ldots, \theta_{m-2}\}$ which is a sibling of θ_m since a node can never be an only child.

Let $\tilde{Q}$ be the questionnaire which interchanges the role of θ_ℓ and θ_{m-1} in Q, i.e.,

$$(6.8) \qquad \tilde{Q} = I_\ell^{m-1} Q.$$

Then

$$(6.9) \qquad E_{(p^{(m-2)},(p_{m-1},p_m))}^{C(\tilde{Q})} \leq E_{(p^{(m-2)},(p_{m-1},p_m))}^{C(Q)}$$

is equivalent to

(6.10) $$p_\ell\varphi_{m-1} + (p_{m-1} + p_m)\varphi_m \leq p_{m-1}\varphi_{m-1} + (p_\ell + p_m)\varphi_m$$

since $\varphi_\ell = \varphi_m$ because θ_ℓ^* and θ_m are brothers under Q and the charging scheme is question based. But (6.10) is equivalent to

(6.11) $$\varphi_m \geq \varphi_{m-1},$$

which is confirmed for an essentially complete class of questionnaires by Theorem 3.2. □

An examination of the results for the $m = 3$ and $m = 4$ case suggests an important property of the subset of the probability simplex where a given questionnaire Q^* is preferred to all others. These subsets might be termed <u>regions of minimum average charge</u>. Then, regardless of the charging scheme used, we have

<u>Theorem 3.4</u>.

Regions of minimum average charge are convex.

<u>Proof</u>:

Suppose Q^* has minimum average charge at $p^{(1)}$ and $p^{(2)}$. Then

(6.12) $$E_{p^{(1)}}C(Q^*) = \inf_Q E_{p^{(1)}}C(Q)$$

and

(6.13) $$E_{p^{(2)}}C(Q^*) = \inf_Q E_{p^{(2)}}C(Q).$$

Now choose $0 \leq \lambda \leq 1$. It is sufficient to show that

(6.14) $$E_{\lambda p^{(1)}+(1-\lambda)p^{(2)}}C(Q^*) = \inf_Q E_{\lambda p^{(1)}+(1-\lambda)p^{(2)}}C(Q).$$

But

$$(6.15)\qquad \inf_{Q} E_{\lambda p^{(1)}+(1-\lambda)p^{(2)}} C(Q) = \inf_{Q} \sum_{i=1}^{|\Theta|} (\lambda p_i^{(1)} + (1-\lambda)p_i^{(2)})\varphi_i(Q)$$

which in turn is no less than

$$(6.16)\qquad \lambda \inf_{Q} \sum_{i=1} p_i^{(1)}\varphi_i(Q) + (1-\lambda) \inf_{Q} \sum_{i=1}^{|\Theta|} p_i^{(2)}\varphi_i(Q)$$

$$= \lambda E_{p^{(1)}} C(Q^*) + (1-\lambda) E_{p^{(2)}} C(Q^*).$$

Further,

$$(6.17)\qquad E_{\lambda p^{(1)}+(1-\lambda)p^{(2)}} C(Q^*) = \sum_{i=1}^{|\Theta|} (\lambda p_i^{(1)} + (1-\lambda)p_i^{(2)})\varphi_i(Q^*)$$

$$= \lambda E_{p^{(1)}} C(Q^*) + (1-\lambda) E_{p^{(2)}} C(Q^*).$$

Therefore,

$$(6.18)\qquad E_{\lambda p^{(1)}+(1-\lambda)p^{(2)}} C(Q^*) \leq \inf_{Q} E_{\lambda p^{(1)}+(1-\lambda)p^{(2)}} C(Q).$$

Since the reverse inequality in (6.18) is obvious, equality holds in (6.14). □

Theorem 3.5.

Suppose $|\Theta| = m$. Then if the charging scheme is $\log_2 d$, the questionnaire Q^* consisting of an initial resolution m question is minimax and the distribution $p^* = (1/m, \ldots, 1/m)$ is least favorable and $\log m$ is the maximin or lower value of the game.

Proof:

Note that

(6.19) $H(p^*) = \log m$

and since a resolution m question has charge $\log m$, Theorem 2.2 gives

(6.20) $$\log m = \inf_{Q} E_{p^*} C(Q) = E_p C(Q^*).$$

Now

(6.21) $$E_p C(Q) = \sum_{i=1}^{m} p_i \varphi_i$$

and for an essentially complete class of questionnaires,

(6.22) $$0 < \varphi_1 \leq \varphi_2 \leq \ldots \leq \varphi_m,$$

by Theorem 3.2. Therefore since $p_1 \geq p_2 \geq \ldots \geq p_m$, $E_p C(Q)$ is maximized with respect to p when $p_1 = p_2 = \ldots = p_m$, i.e., when $p = p^*$. Thus

(6.23) $$E_{p^*} C(Q) \geq E_p C(Q)$$

which implies

(6.24) $$\inf_{Q} E_{p^*} C(Q) \geq \inf_{Q} E_p C(Q)$$

and then

(6.25) $$\inf_{Q} E_{p^*} C(Q) \geq \sup_{p} \inf_{Q} E_p C(Q).$$

But it is immediate that

$$\inf_{Q} E_{p^*} C(Q) \leq \sup_{p} \inf_{Q} E_{p} C(Q) \tag{6.26}$$

so equality obtains in (6.25). Therefore Q^* is minimax, p^* is least favorable, and log m is the maximin value of the game. □

7. Determining an Optimal Questionnaire.

7.1. Huffman Coding.

Suppose now that the charging scheme assesses a charge of $\log_2 d$ for each resolution d question used. Then within the class of homogeneous questionnaires of any fixed resolution d, an optimal questionnaire can be found using the well-known Huffman (1952) coding scheme. (Note that, for a homogeneous questionnaire, average charge minimization is equivalent to average length minimization.) This optimization procedure has been generalized by Picard (1965) to the class of heterogeneous questionnaires in which the number of questions of each resolution which may be used is fixed. Different costs are introduced by Petolla (1969). It is not clear that the Huffman procedure admits a generalization to optimization within the broader class of arbitrary resolution heterogeneous questionnaires. However, a dynamic programming procedure can be used to provide an algorithmic solution to the optimization problem.

7.2. Dynamic Programming Solution.

Let $\Theta = \{\theta_1, \theta_2, \ldots, \theta_m\}$. Suppose $\mathcal{P}$ is a (d-fold) partition of Θ with elements $\Theta_i (i = 1, \ldots, d)$. Thus

$$\bigcup_{j=1}^{d} \Theta_j = \Theta \quad \text{and} \quad \Theta_i \cap \Theta_j = \emptyset \quad \text{for} \quad i \neq j.$$

Further let q_j be the probability of Θ_j and $p_k^{(j)}$ be the normalized and renumbered probabilities of the states of Θ_j.

Then for any valid questionnaire Q,

$$(7.2.1) \qquad E_p C(Q) = \log d + \sum_{j=1}^{d} q_j E_p C(Q_j)$$

where Q_j resolves Θ_j. Then, writing

$$(7.2,2) \qquad K(p) = \inf_{Q} E_p C(Q),$$

equation (7.2.1) implies that

$$(7.2.3) \qquad K(p) = \inf_{\mathcal{P}, Q_1, \ldots, Q_d} [\log d + \sum_{j=1}^{d} q_j E_p C(Q_j)].$$

Therefore,

$$(7.2.4) \qquad K(p) = \inf_{\mathcal{P}} [\log d + \sum_{j=1}^{d} q_j \inf_{Q_1, \ldots, Q_d} E_p C(Q_j)],$$

and finally,

$$(7.2.5) \qquad K(p) = \inf_{\mathcal{P}} [\log d + \sum_{j=1}^{d} q_j K(p^{(j)})].$$

Established by this argument is the validity of the Principle of Optimality of Dynamic Programming (Bellman (1957)). Equation (7.2.5) provides the fundamental functional equation which allows a straightforward algorithmic determination of a minimum average charge questionnaire.

The function K has the following important properties:

Let $p = (p_1, \ldots, p_m)$. Then

(i) K is continuous and concave; K is piecewise linear;

(ii) if $m = 1$, $K(p) = 0$ for all p;

(iii) if $m = 2$, $K(p) = 1$ for all p;

(iv) $H(p) \leq K(p) < H(p) + 1$ for all p (Theorem 2.2); and

explicitly showing the dependence of K on m,

(v) $K_{m-1}(p) < K_m(p) \leq \log m$ for all p and $m \geq 2$.

Now when m is large the number of partitions which must be examined is very large, indeed. Therefore it is useful to be able to restrict the class of partitions which must be examined.

Theorem 3.6.

Let $Q\mathfrak{B} = \{\mathfrak{B}_1, \mathfrak{B}_2, \ldots, \mathfrak{B}_d\}$. Write $\mathfrak{B}_r \prec \mathfrak{B}_s$ if $\theta_i \in \mathfrak{B}_r$ and $\theta_j \in \mathfrak{B}_s$ implies $i < j$. If $p_1 \geq p_2 \geq \ldots \geq p_m$, then the class of questionnaires satisfying

$$\mathfrak{B}_1 \prec \mathfrak{B}_2 \prec \ldots \prec \mathfrak{B}_d \tag{7.2.6}$$

and

$$|\mathfrak{B}_1| \leq |\mathfrak{B}_2| \leq \ldots \leq |\mathfrak{B}_d|, \tag{7.2.7}$$

is essentially complete.

Proof:

By Theorem 3.2 the class, $\mathcal{S}_1$, of admissible questionnaires satisfying $\varphi_1 \leq \varphi_2 \leq \ldots \leq \varphi_d$ is essentially complete. Let $\mathcal{S}_2$ be the class of those questionnaires in $\mathcal{S}_1$ which also satisfy (7.2.6). It is now asserted that $\mathcal{S}_2$ is essentially complete.

To show this, choose $Q \in \mathcal{S}_1 - \mathcal{S}_2$. Then there exist θ_i and θ_j with $i < j$ so that $\theta_i \in \Theta_r$ and $\theta_j \in \Theta_s$ with $r > s$. It is asserted that $I_i^j Q$ has an average charge which is no greater than that of Q. Assuming this for the moment, then each pair of states which fail to conform to the strictures of (7.2.6) may be interchanged with no loss in average charge. Then there will exist an essentially complete class of questionnaires satisfying (7.2.6); therefore $\mathcal{S}_2$ will be essentially complete.

Thus it must be shown that

$$E_p C(Q) - E_p C(I_i^j Q) \geq 0. \tag{7.2.8}$$

But by (5.2) this is equivalent to

$$(p_i - p_j)(\varphi_j - \varphi_i) \geq 0, \tag{7.2.9}$$

which is affirmed since both factors are nonnegative.

Now to complete the proof it is asserted that if $Q \in \mathcal{S}_2$, then (7.2.7)holds. Suppose, on the contrary, that $Q \in \mathcal{S}_2$ but (7.2.7)fails. Then there exist $\Theta_r \prec \Theta_s$ but $|\Theta_r| > |\Theta_s|$. Since $Q \in \mathcal{S}_1$, if λ and μ are arbitrary probability vectors on Θ_r and Θ_s, respectively,

$$\sum_{i=1}^{|\Theta_r|} \lambda_i \varphi_{r_i} \leq \sum_{j=1}^{|\Theta_s|} \mu_j \varphi_{s_j}, \tag{7.2.10}$$

where φ_{r_i} and φ_{s_j} denote charges for state determination in Θ_r and Θ_s, respectively.

Let $\tilde{Q}$ be the questionnaire which replaces the subquestionnaire Q_s with Q_r and identifies terminal nodes lexicographically

until the states of Θ_s are exhausted and then assigns the remaining terminal nodes to elements from the set ω. Then since

$$(7.2.11)\qquad \sum_{i=1}^{|\Theta_s|} \lambda_i \varphi_{r_i} < \sum_{j=1}^{|\Theta_s|} \mu_j \varphi_{s_j},$$

$$(7.2.12)\qquad \tilde{Q} > Q \text{ (strictly)},$$

and hence Q is not admissible. But this is a contradiction of $Q \in S_2$. Hence if $Q \in S_2$, then (7.2.7) must hold. □

We now proceed to count the number of partitions which must be examined, i.e., the number of partitions which satisfy (7.2.6) and (7.2.7).The number of such partitions is the number of non-trivial unordered partitions of the integer m. This combinatorial problem is discussed in Hall (1967) and some typical values are given below:

m	4	5	10	25	50	100
p(m)-1	4	6	41	1957	204,225	190,569,291

Lehmer (1964) gives an asymptotic expression for $p(m)$ of the form

$$\frac{1}{4m\sqrt{3}} \exp[\pi(\frac{2m}{3})^{1/2}]$$

and also a systematic method for generating the partitions.

The results of this section, in particular equation (7.2.3), provide an algorithm for determining an optimal questionnaire for any m.

7.3. Optimality and Shannon Efficiency.

The Huffman coding scheme is essentially a backward optimization procedure. A very simple-minded forward procedure would attempt to apply Theorem 2.3 and construct Shannon efficient questions to the farthest possible stage. Picard (1965) has given an example to show that this procedure can fail. Nevertheless, an attempt in the direction of Shannon efficiency should be made--as is verified by the following theorem:

Theorem 3.7.

At any stage of questioning, the optimal question (partition of the conditional state space) must satisfy

$$(7.3.1) \quad (\log d) - 1 < H(q_1, \ldots, q_d) \leq \log d,$$

where a resolution d question is asked and the offspring have probabilities $q_1, \ldots, q_d$. This gives a necessary condition for optimality.

Proof:

It is sufficient to consider the first stage of questioning. By equation (2.4.29) of Chapter II

$$(7.3.2) \quad K(p) < H(p) + 1.$$

But then by equations (2.4.22) of Chapter II and (7.5),

$$(7.3.3) \quad \log d + \sum_{j=1}^{d} q_j K(p^{(j)}) < H(q_1, \ldots, q_d) + \sum_{j=1}^{d} q_j H(p^{(j)}) + 1.$$

This implies

$$(7.3.4) \qquad \sum_{j=1}^{d} q_j[K(p^{(j)}) - H(p^{(j)})] < H(q_1,\ldots, q_d) - \log d + 1.$$

But the left hand side of(7.3.4) is nonnegative by Theorem 2.2 and hence

$$(7.3.5) \qquad 0 < H(q_1,\ldots, q_d) - \log d + 1.$$

Then since the entropy is bounded above by log d, the conclusion to the theorem follows. □

7.4. Approximation to the Dynamic Programming Solution.

The dynamic programming solution requires that in order to make the optimal partition at the first stage of questioning one must go forward to the later stages and determine minimum costs of questioning. The procedure would be considerably simplified if the best first-stage partition could be based entirely on first-stage calculations. An approximation to the optimal procedure which allows this is available from the following considerations: Choose a partition of $\mathcal{X}$ so that

$$(7.4.1) \qquad K(p) = \log d + \sum_{j=1}^{d} q_j K(p^{(j)}).$$

Then since

$$(7.4.2) \qquad H(p^{(j)}) \leq K(p^{(j)}) < H(p^{(j)}) + 1 \quad (j = 1,\ldots, d),$$

$$(7.4.3) \qquad \log d + \sum_{j=1}^{d} q_j H(p^{(j)}) \leq \log d + \sum_{j=1}^{d} q_j K(p^{(j)})$$

$$< \log d + \sum_{j=1}^{d} q_j H(p^{(j)}) + 1.$$

This suggests choosing a question at each stage so that

(7.4.4) $$\log d + \sum_{j=1}^{d} q_j H(p^{(j)})$$

is minimized. But by equation (2.4.22) of Chapter II this equals

(7.4.5) $$\log d + H(p) - H(q_1, \ldots, q_d).$$

Therefore minimizing expression (7.3.2) is equivalent to minimizing

(7.4.6) $\log d - H(q_1, \ldots, q_d)$. This says that Shannon efficient questions are to be asked at each stage. As noted in the previous section, this procedure is not necessarily optimal. However in the $m = 3$ case it does produce optimal results. In the case of $m = 4$, the regions suggested by this procedure are given below, in Figure 3.5. Note that this procedure cannot be optimal since the regions are not convex (see Theorem 3.4). These approximate regions may be compared with the exact results shown in Figure 3.4.

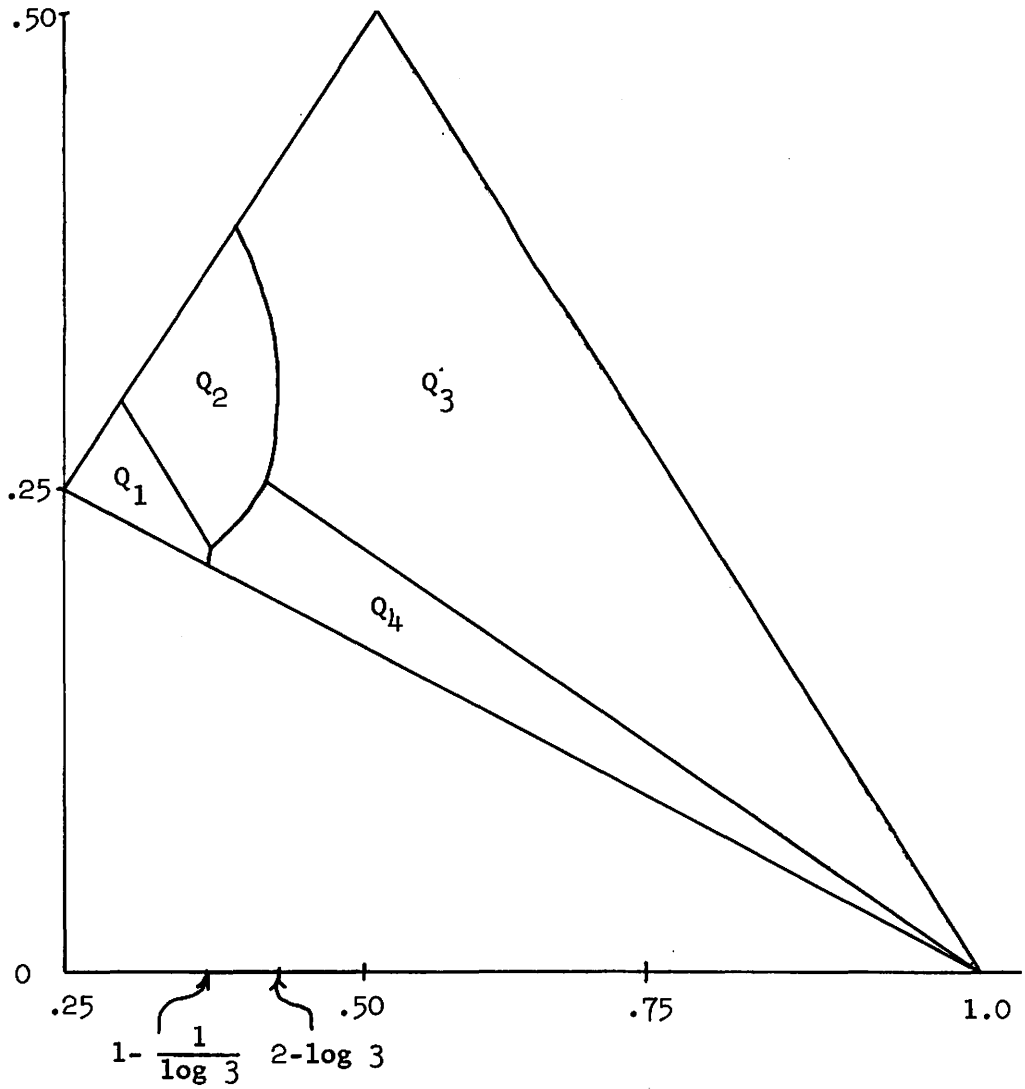


Figure 3.5

Chapter IV

Questionnaires, Uncertainty, and Statistical Inference

1. Information from Uncertainty Functions.

Consider the problem posed in Chapter I of selecting the true state in a state space. Chapters II and III have been devoted to a study of the sure procedure of a questionnaire for identifying this true state. In these past two chapters it was made quite clear that the average charge that the decision maker anticipates for this determination depends on his prior probability measure over the state space. If this measure is relatively diffuse, the average charge calculated from the subjective probability will tend to be large. As was indicated in Chapter I, the decision maker might then choose to collect data which would engender a more concentrated measure via Bayes' Theorem. This would then, among other things, allow a "cheaper" questionnaire to be produced. It is profitable to examine this process of data collection in quite general terms in the spirit of DeGroot (1962) and then specialize to the questionnaire case.

The decision maker has now chosen to be an experimenter whose initial uncertainty about the true value of θ in a finite or countably infinite state space is expressed concisely by $U(p)$, where U designates any one of a class of uncertainty functions and p is a probability vector. The restrictions placed on the class are that the function U be nonnegative, continuous and concave. (This definition is conformable with that of DeGroot (1970) rather than DeGroot (1962).)

An information measure based on the uncertainty function, U, can then be defined. The experimenter chooses to observe a random variable X whose sampling distributions are known.

Then the information content in X about θ (relative to U) is defined by

$$(1.1) \qquad I_U(\theta, X) = U(p) - EU(p(X))$$

where $p(x)$ is the posterior probability vector given $X = x$ and the expectation is computed according to the unconditional distribution of X.

DeGroot shows that (for a finite state space) the information content in X about θ is nonnegative. It follows that the widely used Shannon information is nonnegative since the Shannon entropy is an uncertainty function. The role of Shannon information in the design of experiments is discussed by Lindley (1956, 1957). It is important to realize that the appropriate formulation of information may well depend on the context within which it is employed. Therefore it is useful to examine the properties of information defined over a wide class of uncertainty functions.

An important tool in developing results about uncertainty functions is Jensen's inequality. A desire to have results for a countable state space motivates the following section.

2. A Generalized Jensen's Inequality.

Jensen's inequality is often stated for a convex function of a k-dimensional random vector. As given by Ferguson (1967, p. 76) it has the form: "Let $f(x)$ be a convex real-valued function

defined on a nonempty convex subset S of E_k, and let Z be a k-dimensional random vector with finite expectation EZ for which $P(Z \in S) = 1$. Then $EZ \in S$ and

(2.1) $\quad f(EZ) \leq Ef(Z)."$

We will find it useful to generalize this result to obtain a convexity inequality valid for a real-valued map from sequence space, R^∞. If $x \in R^\infty$, $x = (x_1, x_2, \ldots)$ where $x_i \in R^1$ $(i = 1, 2, \ldots)$.

The usual norm on R^∞ would be the ℓ_2 norm given by

$$\|x\|_{\ell_2}^2 = \sum_{i=1}^{\infty} x_i^2. \tag{2.2}$$

The metric associated with this norm induces the ordinary product topology on R^∞ and hence the usual Borel sets, $\mathcal{B}^\infty$, on R^∞.

It will be convenient to use a different norm defined by

$$\|x\| = \sum_{i=1}^{\infty} |x_i|/2^i. \tag{2.3}$$

This norm is <u>not</u> equivalent to the ℓ_2 norm since if $x = (\epsilon, \epsilon, \epsilon, \ldots)$, then, for $\epsilon > 0$,

$$\|x\| = \epsilon \tag{2.4}$$

while

$$\|x\|_{\ell_2}^2 = \sum_{i=1}^{\infty} \epsilon^2 = +\infty \quad \text{for every } \epsilon. \tag{2.5}$$

It will be useful to have two lemmas which establish the relationship between real-valued continuous functions on $(R^\infty, \|\cdot\|)$ and real-valued continuous functions on $(R^\infty, \|\cdot\|_{\ell_2})$.

Lemma 4.1.

Let $\mathcal{S}$ and $\mathcal{J}$ be the topologies which are induced on R^∞ by $\|\cdot\|$ and $\|\cdot\|_{\ell_2}$, respectively. Then $\mathcal{J} \subset \mathcal{S}$.

Proof:

Let $0 \in \mathcal{J}$. Then if $y \in 0$, there exists $\epsilon > 0$ such that

$$(2.6) \qquad \{x: \sum_{i=1}^{\infty} (y_i - x_i)^2 < \epsilon\} \subset 0.$$

Now $0 \in \mathcal{S}$ since there exists $\delta > 0$ such that

$$(2.7) \qquad \{x: \sum_{i=1}^{\infty} |y_i - x_i|/2^i < \delta\} \subset 0. \square$$

Lemma 4.2.

Denote the open sets of R^1 by $\mathcal{B}^*$. If $U:(R^\infty, \mathcal{J}) \to (R^1, \mathcal{B}^*)$ is continuous and $\mathcal{J} \subset \mathcal{S}$, then $U: (R^\infty, \mathcal{S}) \to (R^1, \mathcal{B}^*)$ is continuous.

Proof:

$0 \in \mathcal{B}^*$ implies $U^{-1}(0) \in \mathcal{J} \subset \mathcal{S}$. $\square$

The following two lemmas will affirm necessary convergence properties:

Lemma 4.3.

If $\{X_i\}_{i=1}^{\infty}$ is a uniformly integrable collection of random variables on a probability space $(\Omega, \mathcal{A}, P)$, then

$$(2.8) \qquad \sum_{i=1}^{\infty} E|X_i|/2^i < \infty$$

and hence

$$(2.9) \qquad \sum_{i=n+1}^{\infty} E|X_i|/2^i \to 0 \quad \text{as} \quad n \to \infty.$$

Proof:

Uniform integrability of $\{X_i\}_{i=1}^{\infty}$ asserts that given $\epsilon > 0$, there exists $c > 0$ such that

$$(2.10) \qquad \int_{\{|X_i| \geq c\}} |X_i| dP < \epsilon \qquad \text{for all} \quad i = 1, 2, \ldots.$$

Then

$$(2.11) \qquad E|X_i| < \epsilon + c \qquad \text{for all} \quad i = 1, 2, \ldots,$$

and therefore $\{E|X_i|\}_{i=1}^{\infty}$ is uniformly bounded.

But then (2.8) and (2.9) are affirmed. □

Lemma 4.4.

If $\{X_i\}_{i=1}^{\infty}$ is a uniformly integrable collection of random variables on a probability space, $(\Omega, \mathcal{G}, P)$, then

$$(2.12) \qquad \sum_{i=n+1}^{\infty} |X_i|/2^i \to 0 \quad \text{a.s.} \quad \text{as} \quad n \to \infty.$$

Proof:

Let

$$(2.13) \qquad S_n = \sum_{i=1}^{n} |X_i|/2^i.$$

Define $\mathcal{B}_n$ to be the minimal σ-field such that $X_1, \ldots, X_n$ are measurable. Then $\mathcal{B}_n \subset \mathcal{B}_{n+1}$ for $n = 1, 2, \ldots$.

Further

$$(2.14) \qquad E^{\mathcal{B}_n}(S_{n+1} - S_n) = E|X_{n+1}|/2^{n+1} \geq 0,$$

where $E^{\mathcal{B}_n}(\cdot)$ denotes the conditional expectation with respect to $\mathcal{B}_n$.

Therefore $\{S_n, \mathcal{B}_n\}_{n=1}^{\infty}$ is a submartingale.

Now

$$(2.15) \qquad \sup_n E|S_n| = \sup_n E \sum_{i=1}^{n} |X_i|/2^i = \sup_n \sum_{i=1}^{\infty} E|X_i|/2^i.$$

But this last expression in (2.15) is finite by Lemma 4.3.

Therefore by the submartingale convergence theorem (see, for example, Loève (1963; p. 393)), there exists an integrable random variable S_∞ such that

$$(2.16) \qquad S_n \to S_\infty \text{ a.s. as } n \to \infty.$$

Thus

$$(2.17) \qquad S_\infty - S_n \to 0 \text{ a.s. as } n \to \infty$$

and hence (2.12) is true. □

We are now in a position to prove a generalized Jensen's inequality.

Theorem 4.1.

Let $(\Omega, \mathcal{G}, P)$ be a probability space, $(R^1, \mathcal{B})$ be the real numbers with their Borel sets, and $(R^\infty, \mathcal{B}^\infty)$ be the space of sequences of real numbers together with the Borel sets induced by the ℓ_2 norm. Let $X:(\Omega, \mathcal{G}, P) \to (R^\infty, \mathcal{B}^\infty)$ and $X = (X_1, X_2, \ldots)$ where $X_i:(\Omega, \mathcal{G}, P) \to (R^1, \mathcal{B})$ $(i = 1, 2, \ldots)$. Suppose $\{X_i\}_{i=1}^{\infty}$ is a uniformly integrable collection of random variables and define $E(X) = (EX_1, EX_2, \ldots)$. Let $V:(R^\infty, \mathcal{B}^\infty) \to (R^1, \mathcal{B})$ be a continuous, concave function. Define

$$(2.18) \qquad \pi_{(n)}X = (X_1, X_2, \ldots, X_n, 0, 0, \ldots)$$

and suppose that $V(\pi_{(n)}X) \geq G(X)$ $(n = 1, 2, \ldots)$ where G is an integrable function.

Then

(2.19) $\quad EV(X) \leq V(EX).$

Proof:

Now

(2.20) $\quad \|\pi_{(n)}X - X\| = \sum_{i=n+1}^{\infty} |X_i|/2^i,$

which goes to zero a.s. as $n \to \infty$ by Lemma 4.4. Hence the continuity of V and Lemmas 4.1 and 4.2 allow us to write

(2.21) $\quad V(X) = \lim_n V(\pi_{(n)}X)$ a.s.

and then taking expectations,

(2.22) $\quad EV(X) = E \lim_n V(\pi_{(n)}X).$

But by Fatou's lemma,

(2.23) $\quad E(\lim_n V(\pi_{(n)}X)) \leq \lim_n \inf EV(\pi_{(n)}X).$

Now Jensen's inequality in R^n shows that

(2.24) $\quad EV(\pi_{(n)}X) \leq V(E\pi_{(n)}X)$

and hence

(2.25) $\quad \lim_n \inf EV(\pi_{(n)}X) \leq \lim_n \inf V(E\pi_{(n)}X).$

But

$$(2.26) \qquad \|E\pi_{(n)}X - EX\| = \sum_{i=n+1}^{\infty} |EX_i|/2^i \leq \sum_{i=n+1}^{\infty} E|X_i|/2^i$$

using the convexity of absolute value.

But

$$(2.27) \qquad \sum_{i=n+1}^{\infty} E|X_i|/2^i \to 0 \quad \text{a.s. as} \quad n \to \infty$$

by Lemma 4.3. Therefore, by continuity of V,

$$(2.28) \qquad \liminf_n V(E\pi_{(n)}X) = V(EX)$$

and combining this with (2.22), (2.23), and (2.25) we see that (2.18) is established. □

For our purposes the following corollary will suffice:

Corollary 4.1.

Let p be a random probability vector on a countable state space; let U be an uncertainty function. Then

$$(2.29) \qquad EU(p) \leq U(Ep).$$

Proof:

By definition the components of p are uniformly bounded and U is nonnegative. Therefore the conditions of Theorem 4.1 are met. □

It is noted that a result similar to Theorem 4.1 might be obtained by establishing a supporting hyperplane theorem in R^{∞} with a method of proof similar to that in Lemma B.1.2 of Karlin

(1959, p. 398). A proof similar to that of Ferguson (1967, p. 76) might then be attempted. Naturally it would be necessary to avoid induction on the dimension of the space.

3. Data Reduction.

Rényi (1962, 1965, 1967) has discussed problems of statistical inference with reference to Shannon information. His particular concern (1967) is with the question of sufficiency of a statistic. This aspect can also be examined for more general information formulations as we shall demonstrate.

First establish the following countably additive structure: Let $(\Omega, \mathcal{A}, P)$ be a probability space; let Θ be a discrete space and $(R^k, \mathcal{B}^k)$ be k-dimensional Euclidean space together with its Borel sets. Suppose X and $\tilde{\theta}$ are jointly distributed random variables so that $(X, \tilde{\theta})$ maps $(\Omega, \mathcal{A}, P)$ into $(R^k \times \Theta, \mathcal{B}^k \times 2^{\Theta})$.

Denote by $\mathcal{A}' \subset \mathcal{A}$ the minimal σ-field such that X is measurable. Allow $T(X)$ to be an arbitrary measurable map into $(R^k, \mathcal{B}^k)$. Note that T is a statistic in the usual sense. Let $\mathcal{A}'' \subset \mathcal{A}'$ be the minimal σ-field such that T is measurable.

Define the function $\chi_{\{\theta\}}$ by

$$\chi_{\{\theta\}}(\omega) = (\chi_{\{\theta_1\}}(\tilde{\theta}(\omega)), \chi_{\{\theta_2\}}(\tilde{\theta}(\omega)), \ldots) \tag{3.1}$$

where $\chi_{\{\theta_i\}}$ is the indicator function of the singleton set containing θ_i. Then taking expectations,

$$E\chi_{\{\theta\}} = (p_1, p_2, \ldots) \tag{3.2}$$

$$(3.3) \qquad E^{\mathcal{G}'}\chi_{\{\theta\}} = (p_1(X),\ p_2(X),\ldots)$$

and

$$(3.4) \qquad E^{\mathcal{G}''}\chi_{\{\theta\}} = (p_1(T(X)),\ p_2(T(X)),\ldots).$$

Each of these expectations is a map to the probability simplex.

Now let U be a concave, continuous map from the simplex to the nonnegative reals, i.e., U is an uncertainty function.

With this rather elaborate structure we can state the following lemma:

Lemma 4.5.

If U is an uncertainty function, then

$$(3.5) \qquad EU(E^{\mathcal{G}'}\chi_{\{\theta\}}) \leq EU(E^{\mathcal{G}''}\chi_{\{\theta\}}).$$

Proof:

First write

$$(3.6) \qquad EU(E^{\mathcal{G}'}\chi_{\{\theta\}}) = EE^{\mathcal{G}''}U(E^{\mathcal{G}'}\chi_{\{\theta\}}).$$

But since U is concave, apply the generalized Jensen's inequality (Corollary 4.1) for conditional expectations to obtain

$$(3.7) \qquad EE^{\mathcal{G}''}U(E^{\mathcal{G}'}\chi_{\{\theta\}}) \leq EU(E^{\mathcal{G}''}E^{\mathcal{G}'}\chi_{\{\theta\}}).$$

Now

$$(3.8) \qquad EU(E^{\mathcal{G}''}E^{\mathcal{G}'}\chi_{\{\theta\}}) = EU(E^{\mathcal{G}''}\chi_{\{\theta\}})$$

since $\mathcal{G}'' \subset \mathcal{G}'$ and the lemma is established. □

This lemma provides a basis for the assertion that "data reduction never increases information." The following theorem is more

general than the result (1.11) by Rényi (1967) in that U is an arbitrary uncertainty function and the state space is allowed to be countable rather than finite.

Theorem 4.2.

Let U be an uncertainty function. Then

$$I_U(\theta, T(X)) \leq I_U(\theta, X). \tag{3.9}$$

Proof:

Note that by definition

$$I_U(\theta, X) = U(p) - EU(E^{G'}\chi_{\{\theta\}}) \tag{3.10}$$

and

$$I_U(\theta, T(X)) = U(p) - EU(E^{G''}\chi_{\{\theta\}}). \tag{3.11}$$

Thus the theorem is equivalent to Lemma 4.5. □

Corollary 4.2.

Let U be an uncertainty function. Then

$$I_U(\theta, X) \geq 0. \tag{3.12}$$

Proof:

Let $T(X)$ be a constant. Then

$$E^{G''}\chi_{\{\theta\}} = E\chi_{\{\theta\}} = p, \tag{3.13}$$

and therefore

$$I_U(\theta, T(X)) = 0. \tag{3.14}$$

The result then follows from (3.9). □

This corollary generalizes the "if" portion of Theorem 2.1 by DeGroot (1962) to a countable state space.

The result of Rényi (1967) that data reduction via a sufficient statistic loses no Shannon information will now be generalized. First, we will define a statistic $T(X)$ to be sufficient for θ if it yields the same posterior distribution as X does (technically, almost surely).

Definition 4.1.

Suppose T and X induce sub-σ-fields, $\mathcal{G}''$ and $\mathcal{G}'$, respectively. Then T is sufficient for θ if

$$E^{\mathcal{G}'} \chi_{\{\theta\}} = E^{\mathcal{G}''} \chi_{\{\theta\}} \quad \text{a.s.}$$

Theorem 4.3.

Let U be an uncertainty function. If T is a sufficient statistic, then

$$I_U(\theta, T(X)) = I_U(\theta, X). \tag{3.15}$$

Proof:

Sufficiency of T is equivalent to

$$E^{\mathcal{G}'} \chi_{\{\theta\}} = E^{\mathcal{G}''} \chi_{\{\theta\}} \quad \text{a.s.} \tag{3.16}$$

Equations (3.10) and (3.11) then give the theorem. □

The concluding remarks of Halmos and Savage (1949) are interesting in this context:

"We think that confusion has from time to time been thrown on the subject by (a) the unfortunate use of the term "sufficient estimate," (b) the undue emphasis on the factorability of sufficient statistics, and (c) the assumption that a sufficient statistic contains all the information in only a technical sense of 'information' as measured by variance."

Remark (a) seems to have been heeded, while remark (b) was met by Bahadur (1954). The results of this section are intended to be conformable with the spirit of remark (c).

4. Questionnaire Information and Shannon Information.

As was suggested in Section 1 of this chapter, a decision maker who ultimately is going to make use of the sure device of an arborescence questionnaire for determination of the true state, might first perform an experiment X. Naturally, the questioner would attempt to choose that experiment which would tend to lower his average charge in using a questionnaire. In this context, it is reasonable to define the questionnaire information in X about $\theta \in \Theta$, a finite or countably infinite state space, by

$$(4.1) \qquad I_K(\theta, X) = K(p) - EK(p(X))$$

where K was defined in Chapter III by

$$(4.2) \qquad K(p) = \inf_Q E_p C(Q).$$

The infimum is taken over all valid, arborescence questionnaires.

This quantity might then be interpreted as the value of the experiment X in terms of questionnaire charges.

Now since K is a continuous, concave function, it follows from Corollary 4.2 that

(4.3) $\quad I_K(\theta, X) \geq 0.$

The usefulness of this information content, as is suggested above, lies in the design of experiments. The experimenter has available an optimality criterion which leads to calling an experiment X^* *optimal in a class* $\mathcal{G}$ if

(4.4) $\quad I_K(\theta, X^*) \geq I_K(\theta, X)$

for all $X \in \mathcal{G}$. Naturally, an optimal experiment may not exist in a particular situation.

The Shannon information has been proposed for use in experimental design in this same manner by Lindley (1956, 1957). DeGroot (1962, 1970) discusses sequential optimal experimental design for general concave uncertainty functions. In a regression framework, Draper and Hunter (1967) show an equivalence between an information optimal experimental design and a Bayes optimal experimental design. Normality is assumed (for both errors and prior) and the Wiener entropy is used as the uncertainty function since the state space is uncountable. The Bayes criterion used is minimization of the determinant of the expected posterior mean square error matrix.

Questionnaire information has a very close relationship to Shannon information. Shannon information might be denoted by $I_H(\theta, X)$ and is defined by using the Shannon entropy as an uncertainty function in the general definition. Thus,

(4.5) $\quad I_H(\theta, X) = H(p) - EH(p(X)).$

As one step in demonstrating this relationship, consider the following:

Theorem 4.4.

If Θ is finite,

$$I_H(\theta, X) - 1 < I_K(\theta, X) < I_H(\theta, X) + 1; \tag{4.6}$$

if Θ is countably infinite,

$$I_H(\theta, X) - 1 \leq I_K(\theta, X) \leq I_H(\theta, X) + 1. \tag{4.7}$$

Proof:

By the results of Chapter II, in the finite case,

$$H(p) \leq K(p) < H(p) + 1. \tag{4.8}$$

Further,

$$H(p(x)) \leq K(p(x)) < H(p(x)) + 1 \tag{4.9}$$

which implies

$$EH(p(X)) \leq EK(p(X)) < EH(p(X)) + 1. \tag{4.10}$$

Therefore,

$$H(p) - EH(p(X)) - 1 < K(p) - EK(p(X)) < H(p) - EH(p(X)) + 1. \tag{4.11}$$

Then (4.6) follows by definition. The countably infinite case is similar with possible equality in (4.8) (see Theorem 2.4). □

If the task set is to determine a sequence of realizations of Θ, as might be the case in a quality control application, a

"block coding" scheme indicates a further tightening of the relationship between questionnaire information and Shannon information. (The "block coding" idea is discussed in Ash (p. 39; 1965).)

The questionnaire information per θ-determination may be made arbitrarily close to the Shannon information by simultaneously determining a sufficiently large number of θ-values. This is established by the following theorem:

Theorem 4.5.

Let Θ^n denote the n-fold Cartesian product of copies of Θ (finite or countably infinite). Let $\mu = (\tilde{\theta}^{(1)}, \ldots, \tilde{\theta}^{(n)})$ be a finite sequence of n random variables with values in Θ^n. They may be arbitrarily jointly distributed with probability vector, p^{μ}, over the elements of Θ^n (m^n of them in the finite case). Then given $\epsilon > 0$ there exists n such that

$$(4.12) \qquad \frac{1}{n} I_H(\mu, X) - \epsilon < \frac{1}{n} I_K(\mu, X) < \frac{1}{n} I_H(\mu, X) + \epsilon.$$

Proof:

Under the stated conditions, Theorem 4.4 shows that there exists a questionnaire Q which resolves Θ^n satisfying

$$(4.13) \qquad I_H(\mu, X) - 1 \leq I_K(\mu, X) \leq I_H(\mu, X) + 1.$$

Hence,

$$(4.14) \qquad \frac{1}{n} I_H(\mu, X) - \frac{1}{n} \leq \frac{1}{n} I_K(\mu, X) \leq \frac{1}{n} I_H(\mu, X) + \frac{1}{n}.$$

So choose n so that $\frac{1}{n} < \epsilon$ and the theorem follows. □

Corollary 4.3.

Suppose $(\tilde{\theta}^{(1)}, X_1), \ldots, (\tilde{\theta}^{(n)}, X_n)$ is a finite sequence of n jointly distributed random variables, independent and identically distributed. Suppose $\tilde{\theta}^{(i)}$ $(i = 1, \ldots, n)$ takes values in Θ while if $\theta_j \in \Theta$, $P\{\theta_j\} = p_j$ $(j = 1, \ldots, m)$ and $p = (p_1, \ldots, p_m)$. Then given $\epsilon > 0$ there exists n such that

$$(4.15) \qquad I_H(\theta, X) - \epsilon < \frac{1}{n} I_K(\mu, X) < I_H(\theta, X) + \epsilon.$$

Proof:

Note that under the conditions of the corollary,

$$(4.16) \qquad H(p^{\mu}) = nH(p). \ \square$$

Therefore in the particular sense of these results, it might be said that the "asymptotic" value of the experiment X in terms of questions is $I_H(\theta, X)$.

5. Forecasters and Questionnaires.

5.1. Payment Schedules for Forecasters.

The decision maker who finds that his probability measure over the state space is unsatisfactory may have another alternative to the statistical approach. This is the use of a forecaster, or the employment of expert opinion to produce a probability measure which is more acceptable.

Naturally, it is in the interest of the decision maker to devise a payment schedule for the forecaster's services which will motivate the forecaster to be most efficient. Efficiency for a forecaster has two aspects:

First, an efficient forecaster should be most diligent in his activities, which might perhaps involve collecting and analyzing data, or even consulting others, so that he may, a priori, be close to certainty as to the true state.

Second, an efficient forecaster should report his findings to his client accurately.

The first aspect deserves to be examined further according to the nature of the process by which the true state is determined. One possible model is that the true state is determined by some actual, physical random process which is inviolable in the sense that its probability structure is not changed by "conditioning." The true states are, to use the standard terminology, elementary outcomes. In this case, the best the forecaster might hope for is to determine the probability measure, P^T, which governs the process. One would imagine that in most cases he would actually succeed in establishing a probability measure, P^F, which would only approximate P^T. Note that in this context, while P^T may have a frequency interpretation, P^F is particular to the forecaster and hence should be interpreted in a personal or subjective sense.

Very often a decision maker is faced with "one of a kind" decisions,making it highly artificial to imagine the true state of nature as being generated by some random process. In this situation, the forecaster might seek a probability measure, P^F, which approximated a distribution placing probability one on the

true state. This might then be considered a degenerate case of the previous situation. Naturally, prudence would dictate not going beyond ones state of knowledge in this endeavor. In either case, the forecaster will actually report to his client a third probability measure, P^R.

The goal of the payment schedule is to provide encouragement to the forecaster to act efficiently. Thus the forecaster should be encouraged to present a P^R which is a close approximation to P^T. In attempting to devise such a payment schedule, a very important distinction between the two situations that the forecaster might face must be taken into consideration. This distinction is that, in the second situation, P^T may well become revealed at some time in the future while this is necessarily not the case in the first situation. Thus a more definitive assessment of the forecaster's effectiveness is possible in the second situation and hence stronger statements can be made as to what constitutes a desirable payment schedule.

Concentrating then on the second situation, suppose θ is observed to occur and the state space Θ is discrete. Then it is known that $P^T\{\theta\} = 1$. Therefore a reasonable payment schedule encourages the forecaster to develop P^T so that $P^F\{\theta\}$ is close to one. This would mean, in particular, that the payment schedule should depend on θ. Symbolically, therefore, the payment schedule might be represented as a function h from the Cartesian product space of the parameter space and the space of probability

measures on the parameter space to the real numbers; the actual payment would be $h(\theta, P^F)$. It is desirable in this situation to have

(5.1.1) $\quad h(\theta, P^F) > h(\theta, P^{F'})$ if $P^F\{\theta\} > P^{F'}\{\theta\}$.

Such a payment schedule is said to encourage diligence. If possible equality were allowed in (5.1.1), the payment schedule would be said to not discourage diligence.

This requirement (5.1.1) is satisfied by many of the commonly proposed payment schedules. For example, any linear function of the following:

(5.1.2) $\quad$ logarithmic: $h(\theta_i, P^F) = \log p_i$

(5.1.3) $\quad$ quadratic: $h(\theta_i, P^F) = 2p_i - \sum_j p_j^2$

(5.1.4) $\quad$ spherical: $h(\theta_i, P^F) = p_i / \{\sum_j p_j^2\}^{1/2}$,

provided that in (5.1.3) and (5.1.4), $\sum_j p_j^2$ remains fixed. This last, in a sense, keeps the variability of the probability estimates constant.

The logarithmic payment schedule is discussed by Good (1952), the quadratic payment schedule is suggested by de Finetti (1962), and the spherical payment schedule is examined by Toda (1963). A discussion of much of the literature is contained in Winkler (1967).

However, the client, who must make the payment, does not know P^F. Instead what is reported to him is P^R. Now it is in

the client's interest that P^R accurately reflect P^F. Indeed, he would want the two measures to identically coincide. McCarthy (1956) has said that a payment schedule encourages honesty if the forecaster believes his average earnings will be maximized if and only if $P^R = P^F$, i.e.,

$$(5.1.5) \qquad E_{P^F} h(\theta, P^F) \geq E_{P^F} h(\theta, P^R) \quad \text{for all } P^R,$$

with equality if and only if $P^R = P^F$. All of the above payment schedules encourage honesty. Dropping the "only if" portion of the above definition gives a payment schedule in which there is no profit in dishonesty.

Aczel and Pfanzagl (1966) consider payoff functions h with the property that

$$(5.1.6) \qquad h(\theta_i, P^F) = g(p_i).$$

Thus the payoff is required to be a function only of the forecasted p_i when θ_i is realized. Clearly, in such a situation, diligence is encouraged when g is strictly increasing. Aczel and Pfanzagl (1966) show that the logarithmic payoff is the only differentiable payoff in the class determined by (5.1.6) for which there is no profit in dishonesty.

5.2. Payments Based on the Value of Information.

Marschak (1959) draws a distinction between "amount of information" and "value of information." He suggests the Shannon entropy as a possible measure of the amount of information in a

forecast, while suggesting that the value of information must be specific to the needs of the buyer of the information. This idea is also discussed by Hurley (1964).

We will discuss the value of information in the context of statistical decision theory. In a statistical decision problem we are given an action space A and a loss function L, assumed to be nonnegative and bounded, on $\Theta \times A$. The decision maker will buy from the forecaster a probability measure on Θ, say p. He may then carry out further experimentation of his own--producing a random variable X with a known sampling distribution depending on θ. He will have chosen a decision function δ^* with smallest Bayes risk (provided one exists), i.e., if

$$(5.2.1) \qquad r(\delta, p) = E_{(\theta,X)}L(\theta, \delta(X)),$$

then

$$(5.2.2) \qquad r(\delta^*, p) \leq r(\delta, p) \quad \text{for all} \quad \delta.$$

A particular $\theta_0 \in \Theta$ will then be realized and the decision maker will sustain a loss of $L(\theta_0, \delta(x))$ where x is the sample value actually obtained from his experiment.

We now assert that a reasonable payoff to the forecaster, fully in accord with the value of the information given to the decision maker, would be a linear function of the risk associated with θ_0. Thus if

$$(5.2.3) \qquad R(\theta, \delta) = E_{X|\theta}L(\theta, \delta(X)),$$

we propose a payoff of

$$a - bR(\theta_0, \delta^*)$$

where a is a flat fee and b is a nonnegative constant reflecting extent of penalty.

This payoff function does not depend on observations obtained by the decision maker in his experimentation; this would appear to be fair to the forecaster. Further the following theorem demonstrates that under this payment schedule the forecaster will find no profit in dishonesty.

Theorem 4.6.

Assume that Bayes decision functions exist for each probability distribution on Θ. Let δ_F^* and δ_R^* be Bayes decision functions with respect to probability distributions, P^F and P^R, respectively on Θ. Then a payment schedule which is a linear function of the risk evaluated at δ^* and θ_0 allows no profit in dishonesty.

Proof:

We want to show that the forecaster believes his average payment will be maximized when $P^R = P^F$, i.e.,

$$(5.2.4) \qquad \int (a - bR(\theta_0, \delta_F^*))dP^F \geq \int (a - bR(\theta_0, \delta_R^*))dP^R.$$

But (5.2.4) is equivalent to

$$(5.2.5) \qquad \int R(\theta, \delta_F^*)dP^F \leq \int R(\theta, \delta_R^*)dP^F$$

which is true by definition of δ_F^*. □

One aspect of the payment schedule discussed in this section which merits attention is the fact that the forecaster is necessarily drawn more closely into the decision maker's problem. Presumably in the negotiation of the constants a and b in (5.2.3), the forecaster would have to be informed of the decision maker's loss function and the nature of the experimentation which the decision maker plans to carry out. This "identification of interests" might conceivably be desirable in some circumstances but in others it might well be better to develop a payment schedule which would avoid such disclosures (or even advance planning of experimentation). As Marschak (1959) has pointed out, the logarithmic payment schedule might be used in this latter situation. It enjoys the double advantage of encouraging both diligence and honesty. Another payment schedule which might be considered for use in this context is discussed in the next section.

5.3. The Client is a Questioner.

Suppose now that the client is a questioner concerned with a finite state space. He will be making use of the probability vector that the forecaster reports, P^R, in order to effectively design a questionnaire. Then it seems reasonable to base the payment to the forecaster on the amount it will cost the questioner to determine the true state using the best questionnaire, Q^*, he can construct with P^R. Thus Q^* satisfies

$$(5.3.1) \qquad E_{P^R} C(Q^*) = \inf_{Q} E_{P^R} C(Q).$$

A simple form of payment schedule would be of the following form:

(5.3.2) $h(\theta_i, P^R) = a - b\varphi_i(Q^*)$

where a is a positive constant meant to reflect a basic flat fee, b is a positive constant giving extent of penalty, and $\varphi_i(Q^*)$ is the charge to determine the true state θ_i using Q^*.

It might first be noted that the payment schedule given by (5.3.2) fails to encourage diligence. This is evident from the $m = 4$ case discussed in Chapter III where for $p_1 = .35$ and $p_2 = .23$, $Q_3 \succ Q_2$; whereas the domination reverses for $p_1 = .40$ and $p_2 = .30$. But $\varphi_1(Q_3) = 1$ and $\varphi_1(Q_2) = \log 3$.

On the other hand it is an immediate corollary of Theorem 4.6 that the payment schedule 5.3.1 does not discourage honesty.

Corollary 4.4.

Under payment schedule (5.3.2) there is no profit in dishonesty.

Proof:

Identify $\varphi_i(Q^*)$ with $R(\theta_0, \delta^*)$ in Theorem 4.6. □

6. Subjective Probability.

Suppose it is desired to know something of a person's subjective probability in a particular context. The most definitive statements about this probability can be made by proceeding normatively with a Rational Man, well-schooled in the construction of questionnaires. The basic approach is to observe how he goes about this construction. From these observations, one would hope to infer the nature of his subjective probability. The Rational Man will be assumed to make decisions according to an ordering

of possible decisions by their Bayes risk with weights given by his subjective probabilities. He will be neither a risk taker nor a risk avoider. It is intended to confront the Rational Man with a finite state space made up of m states, precisely one of which--not initially specified to the Rational Man--is designated as "true." He will be entitled to know the location of the true state among the various sets of a partition of the state space. This will be done according to the rules of a questionnaire upon payment of the appropriate charge.

Since the questionnaire consisting of a single resolution m question--the minimax questionnaire--is always available, with charge $\log m$, the Rational Man will be rewarded with a payment of $\log m$ upon discovery of the true state, thereby achieving a partial balance between payments and charges.

Suppose the event A is under consideration. Identify this event with state θ_1. Partition A^c into two arbitrary sets and identify them with states θ_2 and θ_3. Then confront the Rational Man with a choice of four questionnaires:

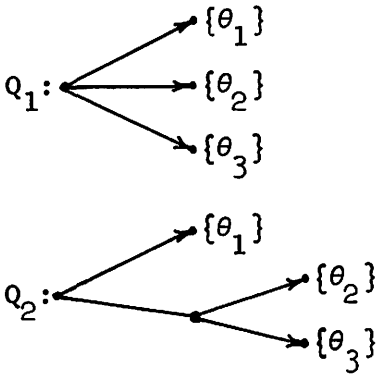


Figure 4.1a

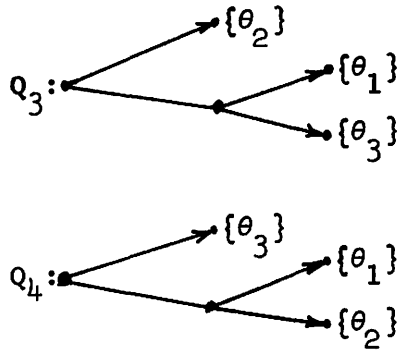
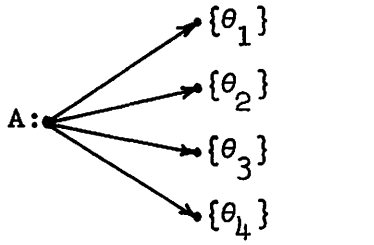


Figure 4.1b

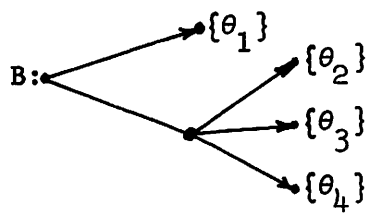
Some statement can be made about his subjective probability for A from his choice. Namely,

Choice	P(A)
Q_1	$1/3 \leq P(A) \leq 2 - \log_2 3$
Q_2	$P(A) \geq 2 - \log_2 3$
Q_3	$P(A) \leq 1/3$
Q_4	$P(A) \leq 1/3$

Suppose there are two disjoint events A_1 and A_2; he can be confronted with an essentially complete class of questionnaires of order 4:



(one of this type)



(four of these)

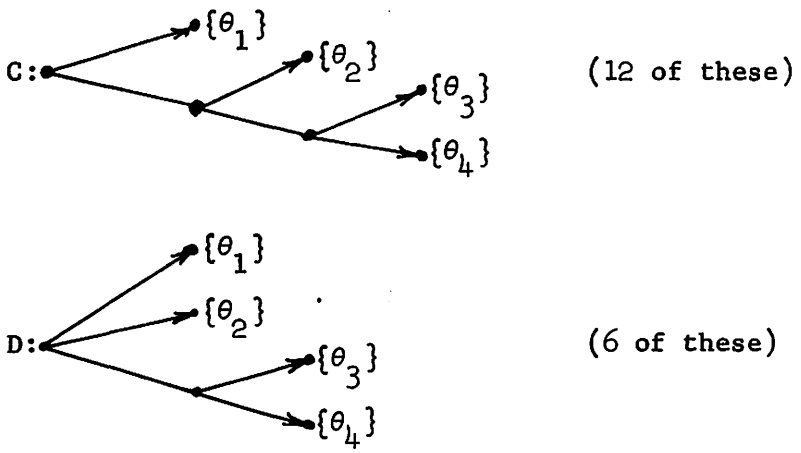


Figure 4.2

A selection of one of these will provide bounds on the probabilities, $P(A_1)$ and $P(A_2)$.

A special problem exists when $P(A) > 1/2$ since no distinction between questionnaires will require $P(A) > 1/2$. This can be handled by independently tossing a fair coin and looking at $P(A \cap H)$.

REFERENCES

1. Aczel, J. and Pfanzagl, J. (1966). Remarks on the measurement of subjective probability and information. Metrika 11 91-105.
2. Ash, Robert (1965). Information Theory. Interscience Tracts in Pure and Applied Mathematics No. 19, John Wiley, New York.
3. Bahadur, R. R. (1954). Sufficiency and Statistical Decision Functions. Annals of Mathematical Statistics 25 423-462.
4. Barnard, G. A. (1951). The theory of information. Journal of the Royal Statistical Society Series B 13 46-64.
5. Bartlett, M. S. (1951). Discussion of Professor Barnard's Paper. Journal of the Royal Statistical Society Series B 13 59-60.
6. Bellman, Richard (1957). Dynamic Programming. Princeton Univ. Press, Princeton, New Jersey.
7. Billingsley, Patrick (1961). On the coding theorem for the noiseless channel. Annals of Mathematical Statistics 32 594-601.
8. Burge, W. H. (1958). Sorting, Trees and Measures of Order. Information and Control 1 181.
9. Campbell, L. L. (1968). Note on the connection between search theory and coding theory. Proceedings of the Colloquium on Information Theory (Budapest), A. Renyi, ed. 85-88.
10. Cox, Richard T. (1961). The Algebra of Probable Inference. Johns Hopkins Press, Baltimore, Maryland.

11. de Finetti, Bruno (1962). Does It Make Sense to Speak of 'Good Probability Appraisers'? in I. J. Good, ed. The Scientist Speculates--An Anthology of Partly-Baked Ideas. Basic Books, New York, 357-363.

12. De Groot, M. H. (1962). Uncertainty, Information and Sequential Experiments. Annals of Mathematical Statistics 33 404-419.

13. De Groot, M. H. (1970). Optimal Statistical Decisions. McGraw-Hill, New York.

14. Draper, Norman R. and Hunter, William G. (1967). The use of prior distributions in the design of experiments for parameter estimation in nonlinear situations: multi-response case. Biometrika 54 662-665.

15. Dubail, Francoise (1967). Algorithms de questionnaires réalisables, optimaux au sens de différents critères. Thèse présentee a la faculté des sciences de l'Université de Lyon pour obtenir le titre de Docteur de Spécialité (Mathématique Appliqués).

16. Fano, Robert M. (1961). Transmission of Information. Massachusetts Institute of Technology Press, Cambridge.

17. Ferguson, Thomas S. (1967). Mathematical Statistics: A Decision Theoretic Approach. Academic Press, New York.

18. Flament, Claude (1963). Applications of Graph Theory to Group Structure. Prentice-Hall, Englewood Cliffs, New Jersey.

19. Ford, L. R., Jr. and Johnson, S. M. (1959). A tournament problem. American Mathematical Monthly 66 387-389.

20. Good, I. J. (1952). Rational decisions. Journal of the Royal Statistical Society Series B 14 107-114.

21. Hadian, Abdollah (1969). Optimality properties of various procedures for ranking n different numbers using only binary comparisons. Technical Report No. 117, Department of Statistics, University of Minnesota, Minneapolis, Minnesota.

22. Halmos, P. R. and Savage, L. J. (1949). Application of the Radon-Nikodym theorem to the theory of sufficient statistics. Annals of Mathematical Statistics 20 225-241.

23. Hall, Marshall, Jr. (1967). Combinatorial Theory. Blaisdell, Waltham, Massachusetts.

24. Huffman, David A. (1952). A method for the construction of minimum-redundancy codes. Proceedings of the Institute of Radio Engineers 9 1098-1101.

25. Hurley, W. V. (1964). A mathematical theory of the value of information. International Journal of Computational Mathematics 1 97-146.

26. Karlin, Samuel (1959). Mathematical Methods and Theory in Games, Programming and Economics, Volume I. Addison-Wesley, Reading, Massachusetts.

27. Kerridge, D. F. (1961). Inaccuracy and inference. Journal of the Royal Statistical Society Series B 23 184-194.

28. Kolmogorov, Andrei N. (1956). The Shannon theory of information transmission in the case of continuous signals. Institute of Radio Engineers Transactions on Information Theory IT-2 102-108.

29. Kolmogorov, Andrei N. (1965). Three approaches to the quantitative definition of information. Problemy Peredacii Informacii 1 3-11 (Russian)(translation: Problems of Information Transmission. Faraday Press, New York.)

30. Kraft, L. G. (1949). A device for quantizing, grouping and coding amplitude modulated pulses. M.S. Thesis, Electrical Engineering Department, Massachusetts Institute of Technology.

31. Lehmer, Derrick H. (1964). The Machine Tools of Combinatorics. In Applied Combinatorial Analysis, E. F. Beckenbach, ed. John Wiley, New York.

32. Lindley, D. V. (1956). On a measure of the information provided by an experiment. Annals of Mathematical Statistics 27 986-1005.

33. Lindley, D. V. (1957). Binomial sampling and the concept of information. Biometrika 44 179-186.

34. Loève, M. (1963). Probability Theory (Third Edition). Van Nostrand, New York.

35. MacDonald, D. K. C. (1952). Information theory and its applications to taxonomy. Journal of Applied Physics 23 529.

36. Marschak, J. (1959). Remarks on the economics of information. Contributions to Scientific Research in Management, 79-100. Western Data Processing Center, University of California, Los Angeles.

37. McCarthy, John (1956). Measures of the Value of Information. Proceedings of the National Academy of Sciences 42 654-655.

38. Moran, P. A. (1951). Discussion of Professor Barnard's Paper. Journal of the Royal Statistical Society Series B 13 60-61.

39. Ore, O. (1962). Theory of Graphs. American Mathematical Society, Providence, Rhode Island.

40. Osborne, Donald V. (1963). Some aspects of the theory of dichotomous keys. New Phytologist 62 No. 2.

41. Parkhomenko, P. P. (1969). Optimal questionnaires with unequal question prices. Doklady Akademii Nauk SSR 184 51.

42. Petolla, Sylvette (1966). Problems de minimisation de couts dans les questionnaires. Exposé au séminaire sur des questionnaires du 31 janu. 1966 l'Université de Lyon.

43. Petolla, Sylvette (1969). Extension de l'algorithme d'Huffman a une classe de questionnaires avec couts. Thèse, présentee a l'U.E.R. de Mathématiques de l'Université de Lyon. No. d'ordre 442.

44. Picard, Claude F. (1965). Théorie des Questionnaires. Gauthiers-Villars, Paris.

45. Picard, Claude F. (1968). Sur la longueur de cheminement d'un questionnaire latticiel. Publication No. AMT/2.10.8/AF, Institute Blaise Pascal, Paris.

46. Picard, Claude F. (1969). Quasi-questionnaires, codage et longueur de Huffman. Manuscript, April, 1969.

47. Rényi, Alfred (1962). Statistical laws of accumulation of information. Bulletin de l'Institut International de Statistique (The Hague) 39 311-316.

48. Rényi, Alfred (1965). On some basic problems of statistics from the point of view of information theory. Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, University of California Press.

49. Rényi, Alfred (1967). Statistics and Information Theory. Studia Scientarium Mathematicarum Hungarica 2 249-256.

50. Shannon, C. E. (1948). A Mathematical Theory of Communication. Bell System Technical Journal 27 379-423, 623-656.

51. Shannon, C. E. and Weaver, W. (1949). The Mathematical Theory of Communication. University of Illinois Press, Urbana, Illinois.

52. Sobel, Milton (1960). Group testing to classify efficiently all units in a binomial sample. Information and Decision Processes, R. E. Machol ed. McGraw-Hill, New York.

53. Toda, Masanao (1963). Measurement of Subjective Probability Distribution. State College of Pennsylvania: Report No. 3, Division of Mathematical Psychology, Institute for Research.

54. Tribus, Myron, Shannon, Paul T. and Evans, Robert B. (1966). Why Thermodynamics is a logical consequence of Information Theory. American Institute of Chemical Engineering Journal, March, 1966, 244-248.

55. Wiener, Norbert (1948). Cybernetics. John Wiley, New York.

56. Wilde, Douglass J. (1964). Optimum Seeking Methods. Prentice-Hall, Inc., Englewood Cliffs, New Jersey.

57. Winkler, Robert L. (1967). The Quantification of Judgement: Some Methodological Suggestions. Journal of the American Statistical Association 62 1105-1120.