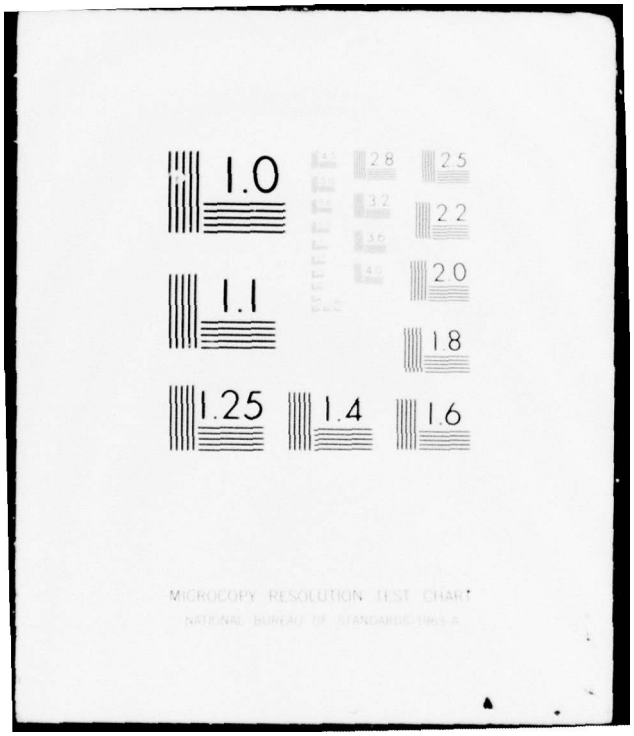
AD-A042 852 CALIFORNIA UNIV LOS ANGELES SCHOOL OF ENGINEERING A--ETC F/G 5/10
GROUP DECISION THEORY.(U)
JUL 77 N C DALKEY
UNCLASSIFIED UCLA-ENG-7749 N00014-69-A-0200-4056
NL

1 OF 4
AD A042852


1.0
1.1
1.25
1.4
1.6
2.8
2.5
3.2
2.2
3.6
2.0
4.0
1.8
MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS 1963-A

This research was supported by the
Advanced Research Projects Agency
of the
Department of Defense
and was monitored by the
Office of Naval Research
Under Contract No. N00014-69-A-0200-4056/452.

404 637

UCLA-ENG-7749
JULY 1977

GROUP DECISION THEORY

N.C. DALKEY

UCLA • SCHOOL OF ENGINEERING AND APPLIED SCIENCE

DISCLAIMER

The views and conclusions contained in this document are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of the Advanced Research Projects Agency or the U.S. Government.

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

1. REPORT NUMBER: UCLA-ENG-7749
2. GOVT ACCESSION NO.
3. RECIPIENT'S CATALOG NUMBER

4. TITLE (and Subtitle): GROUP DECISION THEORY,
5. TYPE OF REPORT & PERIOD COVERED: Final Report 9/1/74 - 6/30/75
6. PERFORMING ORG. REPORT NUMBER

7. AUTHOR(s): NORMAN C. DALKEY
8. CONTRACT OR GRANT NUMBER(s): N00014-69-A-0200-4056

9. PERFORMING ORGANIZATION NAME AND ADDRESS
University of California, Los Angeles
School of Engineering and Applied Science
Los Angeles, California 90024

10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS: 5D20

11. CONTROLLING OFFICE NAME AND ADDRESS
Defense Advanced Research Projects Agency
Human Resources Research
1400 Wilson Blvd.
Arlington, VA. 22209

12. REPORT DATE: July 1977
13. NUMBER OF PAGES: 314

14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office)
Office of Naval Research
Pasadena Branch Office
1030 East Green Street
Pasadena, CA. 91105

15. SECURITY CLASS. (of this report): Unclassified
15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Reproduction in whole or in part is permitted for any purpose of the United States Government

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

Final rept. 1 Sep 74 - 30 Jun 75;

DDC RECEIVED AUG 11 1977 C

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Decision theory, group judgment, Delphi, group decision

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

A general foundation for group decisions is formulated involving:
(a) Resolution of disagreement on factual questions by demonstrating greater accuracy of group judgment over average accuracy of individual judgments.
(b) Resolution of inconsistencies between individual and group preferences (Arrow paradox) by means of ordinal scales with fixes reference points.
(c) Demonstration that, if the individual members of the group have cardinal utility scales, then the conditions of dominance, acyclicity; continuity

(over)

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE S/N 0102 LF 014 6601

404 637

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

and equivalence (if the group is indifferent between actions A and B, it is indifferent between A and any probability combination of A and B), then there is a group utility function which is a weighted sum of the individual utilities.

UCLA-ENG-7749
JULY 1977

GROUP DECISION THEORY

Norman C. Dalkey

This research was supported by the
Advanced Research Projects Agency
of the
Department of Defense
and was monitored by the
Office of Naval Research
Under Contract No. N00014-69-A-0200-4056/452.

School of Engineering and Applied Science
University of California
Los Angeles, California

TABLE OF CONTENTS

List of Figures

CHAPTER I. INTRODUCTION

1. Disagreement

Group decisions are basic components of society. They occur at every level of social interaction—from the selection of an evening's entertainment by a casually dating couple to the election of a president by a large nation. But despite their ubiquitous presence, group decisions are not well understood. That comment holds both for descriptive theory—the theory of how groups make decisions—and for prescriptive theory—rules for making rational group decisions.

This situation contrasts with the state of decision theory for individuals. There is a fairly rich literature dealing with empirical studies of individual choice; and the theory of rational individual choice—often called decision analysis—has made rapid progress in the past quarter of a century. Stemming primarily from the theory of games, a coherent set of rules has evolved which has proved highly fruitful in identifying the major elements of individual choice, and in establishing a framework within which rational individual decisions can be defined. This framework involves the notions of numerical value scales (utilities), estimated probabilities (sometimes called subjective probabilities), and the rule, select the action which maximizes expected utility.

The stumbling block in attempting to extend these notions to group decisions is the existence of disagreement. If all the members of a group agree on the salient features of a decision problem, no special difficulties arise. But in almost all interesting decisions in practice, members of the group can differ widely on any relevant aspect of the decision problem.

There are two generic bases for disagreement: uncertainty (incomplete information) and conflicting interests. If crucial aspects of the decision are poorly understood, then differences of opinion are pretty much inevitable.

In addition, if various courses of action open to the group lead to differential rewards to members of the group, then the members are likely to evaluate the courses of action differently.

These two generic causes of disagreement spawn a wide variety of discords. The following short list is not intended to be comprehensive, but to pinpoint some of the more critical types. Appended to each type is a touchstone question which highlights the bone of contention.

1. Point of view. What is the problem?
2. Factual. What is the case?
3. Value. What is worthwhile?
4. Interest. Who gets what?

Point of view differences are the hardest to characterize and the most difficult to deal with in preactice. In any given decision, individuals can and do have quite different "models" of the situation. In an environmental dispute, one individual can take an ecological point of view, another an economic, still another a humanistic or aesthetic, and so on. These differences cannot be summed up entirely by phrases such as different emphasis, or different beliefs. The different points of view are different ways of representing the world. A cognate notion is problem formulation. Different individuals formulate the decision problem (as they see it) in categories which taken together, do not form a coherent structure.

Some of the issues involved in point of view disagreement are discussed in Chapter II, Section 2 under the topic universe of discourse. Different individuals can, in effect, be talking in different languages if they bound the problem differently. For example, one individual can maintain that a given possibility is irrelevant to some central feature of the decision, and another individual maintain that on the contrary it is highly relevant, and both be correct within their respective universes of discourse.

Methods of dealing with point of view differences are not well developed at the present time. The ancient rule, "First define your terms," is relatively feeble. It essentially assumes a common universe of discourse. Some technology exists which is helpful if individuals can formulate their individual models explicitly. Among these are cluster analysis to generate common sets of categories, and relevance trees to allow for several levels of aggregation.[1] However, the requisite theory to apply these techniques to point of view disagreement has not been generated, and more to the point, figures of merit for evaluating the effectiveness of the techniques have not been defined.

In this report, point of view disagreement will usually be sidestepped by the assumption that the group has already agreed on a common model of the basic factors in the decision. The formal resolution procedures then deal with the other types of disagreement which can arise within the common model.

Factual disagreement is the clearest of the four, and the type for which methods of resolution are most advanced. A major fraction of this report will be devoted to the topic.

Value and interest differences are easier discussed together, since they are often confused. In the terminology of decision theory, values relate to criteria, objectives, or payoffs, whereas interests relate to the allocation or distribution of rewards. It is possible for two individuals to agree completely on what is worth having, and disagree completely on who should have it. In fact, there is a significant inverse relationship between value and interest disagreement. The greater the disagreement on values, the smaller the disagreement on interests. The relationship can be illustrated by the old nursery rhyme about Jack Sprat and his wife. Jack Sprat could eat no fat, and his wife could eat no lean—complete disagreement about values. As a result, they could eat the platter clean—no conflict of interest.

Many discussions of value conflict, especially in the economic literature, obscure the distinction between these two types of disagreement by simplifying the motivational component of decisions to a single notion, namely preference. An object A is considered more valuable to individual I than object B if I prefers A to B (i.e., I would select A over B given a free choice). However, preference has two "dimensions", value per se and amount. An individual can prefer item A to item B for either or both of two reasons, item A is a more valuable kind of item than B, or they are both the same kind of item and B includes more.

The most intense form of interest disagreement occurs when values are identical, but there is a scarcity of rewards. In the theory of games, the sharpest conflict occurs with the zero-sum two-person game where the payoff is equivalent for each player, but the rules of the game determine that whatever one player gains the other loses. On the other hand, if individual I does not want what individual J wants, and vice versa, it is hard to start a quarrel.

The status of value judgments is somewhat up in the air at the present time. Individuals do disagree on the relative worth of different kinds of rewards as well as on allocations. However, there is no generally accepted criterion of correctness of value judgments. In the prevalent view, value judgments are "ad lib".

2. Resolution of Disagreement

In practical affairs, there are a number of reasons for avoiding or resolving disagreement. Above all, of course, disagreement can be an impediment to action, providing action requires consent on the part of members of the group. But in addition, disagreement usually entails costs in delayed action, and in abrasive interaction. It can lead to conflict, ranging from "verbal battles" to more violent forms of confrontation. A more insidious kind of cost can

occur in the form of degraded decisions. The resolution process, especially if it involves so-called compromises, can lead to large biases in the final choice.

Historically, a number of procedures have evolved to deal with disagreement. Some of the more widely practiced are:*

1. Dictatorial. One individual makes the decision.
2. Objective. The decision is made according to preestablished rules.
3. Darwinian. The decision is the outcome of competition.
4. Collective. The decision results from amalgamation of the individual judgments.

The dictatorial solution is by far the most common way of resolving disagreement. It occurs not only in tyrannies, but in all walks of life. There is nothing necessarily despotic (i.e., arbitrary) in the notion. The industrial manager, the government agency head, the head of a household, any one who can claim "final authority", is a device to resolve disagreement. In practice, the "one man" nature of the procedure is obscured by complexity--e.g., the hierarchical structure of large management staffs--and by the constraints which the "system" places on the abuse of power.

The dictatorial solution has more than historical usage to justify it. It is effective, it is efficient, and it has the advantage that it is free of some of the more vexing conceptual issues of multi-person procedures. Because of these strong advantages, it is likely to be the most widely used method of resolving disagreement for some time to come. However, the dictatorial solution has a number of weaknesses beyond the potential abuse of power. Above all, it

*I have omitted from this list the more violent procedures such as physical coercion, and the more blatantly totalitarian procedures such as information control, not because they are rare, but because they are outside the scope of the present treatment, which is limited to cooperative decisions.

is subject to the biases, limited point of view, and other pathologies of individual judgment.

Objective resolution of disagreement takes a number of forms. Perhaps the most relevant to the present discussion is the form exemplified by institutionalized science. The scientific community has developed a set of criteria to settle factual issues. An essential element of these criteria is objectivity.* The rules can be expressed in terms that do not refer to the individual researcher, i.e., in terms of data, and inferences from data. There may be some controversy concerning the precision of the criteria, especially with regard to inferences. But the contrast between the relatively objective, rule-prescribed procedures of the natural sciences and the fuzzier procedures in other social domains is clear.

Perhaps the most salient feature of science from the present perspective is its extraordinary success. By relegating pure debate and personal influence to the background in settling factual disputes it has exhibited a power to solidify knowledge in a way that is well beyond reasonable doubt. For this reason, there appears to be little question that scientific knowledge is the most excellent kind of information that can be input into a decision--when it is available.**

There is only one serious weakness of the scientific method for most decisions; namely, it is incomplete. If a firm scientific basis can be found for an assertion, it is a valuable input to any decision for which it is

---

*Some scientists would claim that intersubjectivity is all that is required. Whether intersubjectivity can be achieved without objective reference points is a moot subject.

**The only dubiety here is a matter of relative solidity. Much of the "know-how" of technology also has a high-order of credence, even though it does not have the overt validation structure of systematic science. The technologists test "Does it work?" appears to be about as powerful as the scientists' "Is it substantiated by experiment?" in weeding out groundless beliefs.

relevant. But if a firm scientific basis is not available then the statement remains in scientific limbo; it is simply unproved. For most interesting decisions, a large proportion of factual issues are in the scientific limbo.

Scientists have imposed another incompleteness on their method, namely the contention that science can say nothing about values. There appears to be an active debate beginning on this subject within some scientific communities. For the time being, however, there are no value judgments that can claim the "official" sanction of scientific method.

The term Darwinian refers to a wide variety of types of disagreement resolution that involve competition. Perhaps the purest example is the debate, where two individuals present as powerful array of arguments for and against a given statement as they can, and "the best man wins". The typical formal debate requires a judge (or judges) who is, in effect, the agency of resolution. In more general settings, expressed by phrases such as the marketplace of ideas, the intellectual forum, and the like, the role of judge is presumably taken by a loosely defined interested community.

For more narrowly defined group decisions, the group itself may be the judge, and may also include the contenders. Resolution is by "consensus" a somewhat vaguely defined process including, usually, face-to-face discussion, various forms of mutual persuasion, and other influences which may lead to agreement.

I have labeled this fuzzy class of procedures Darwinian because of the implied assumption that the competition leads to "survival of the fittest"—i.e., that the most excellent judgment is the one that wins out. This assumption appears to be more an article of faith than the result of careful evaluation. There are serious problems in designing experiments to evaluate the effectiveness of competitive processes for selecting the best (e.g., the most accurate)

judgment out of a list of contenders. Nevertheless the question whether competitive processes are effective is an empirical one. My own attitude, based on a few experiments of my own[2] and after surveying the rather sparse literature on the subject is that competitive procedures are probably better than dictatorial ones, at least on the average. However, if competition is viewed as a filtering process, then my impression is that the efficiency of filtering per stage is rather low.

The first three methods of resolution are roughly ways of selecting one judgment out of a group of judgments. One somewhat vague rationale that can be forwarded on their behalf is that, given disagreement, there is one judgment that is correct, and the others wrong; or somewhat weaker, there is one judgment which is better than the others, and the goal is to find that judgment. The fourth method has a different rationale. It starts from the assumption that if there is major disagreement, especially within a knowledgeable group, then in all likelihood, none of the members of the group knows the answer to the question. In such a case, rather than selecting a single answer, more can be gained by amalgamating all the answers—hence the term collective.

Methods of amalgamating individual judgments are in an early stage of development. Procedures which are implementable in practice come down to some form of measure of central tendency (mean, median, geometric mean, etc.) with a measure of dispersion (standard deviation, interquartile range, etc.) to indicate the degree of disagreement. However, in theory at least, more sophisticated methods of pooling individual judgments are possible, and are discussed in Chapter V.

Most of the results which form the body of this report are presented within the framework of the collective approach to disagreement resolution.

3. The Emerson Principle

The aggregation problem can be expressed formally as follows: there is a group of individuals who, on a given subject, have a set of judgments J_i where the index refers to individual i. To obtain a group judgment on the same topic, there is a function F(J), $J = (J_1,\ldots,J_n)$, which aggregates the set of n individual judgments into a single group judgment. The function F should fulfill some straightforward conditions:

1. Substantive Conditions. F(J) should be the same sort of judgment as the J_i. Example: If the J_i are probabilities for a given event, then F(J) should be a probability.
2. Consistency Conditions. Consistency here refers to coherence between the individual judgments and the group judgment. Consistency at the individual or group level is part of the substantive conditions. Example: If all the members of the group are in agreement, $J_i = J_k$ for all i and k, then $F(J) = J_i$ (the unanimity principle).
3. Performance Conditions. If there is a figure of merit for the individual judgments, then F(J) should not perform poorly with respect to this figure of merit. Example: If J_i is individual i's answers on a test, and each individual gets a high score on the test, then F(J) should not get a low score.

Conditions of type 3 have not received a great deal of attention in the literature on group decisions, primarily because those of type 2 already appear to pose insurmountable difficulties. Probably the best known of these difficulties is the result derived by Kenneth Arrow that, if the J_i are individual preference relations, there is no F which fulfills a few, highly plausible, consistency conditions. This result is discussed in some detail in Chapter VI. A similar difficulty is exemplified by a result I demonstrated some time ago

to the effect that if the J_i are probabilities, then no F exists which fulfills the usual axioms of probability for both the individual judgments and the group judgment. This result is expounded in Chapter V, Section 5.

A basic theme of the collective resolution of disagreement is that conditions of type three can compensate for difficulties with conditions of type two. The idea is straightforward; if a group judgment can be shown to perform well on a given figure of merit, then a certain amount of non-conformity between individual and group judgments is tolerable. I have called this the Emerson Principle--performance is at least as important a criterion for aggregation as consistency.*

To invoke the Emerson Principle, it is necessary to have a well-defined figure of merit that applies both to individual judgments and to group judgments. For factual judgments, there is a large family of figures of merit, or scores, which enable comparing the performance of individual and group estimates. This is the topic of Chapter III. Using these scores it is possible to derive a corresponding family of n-heads rules, i.e., statements to the effect that the group score is better, in some well-specified sense, than the corresponding individual scores. This is the topic of Chapter V. The n-heads rules appear to be a satisfactory justification for using group estimates in decisions where the individual members disagree on factual issues. This "resolution" of disagreement is a good deal stronger than simply finding a "compatible" group

*Historically, there has been a wide range of reactions to the discovery of inconsistencies, from panic to stubborn unconcern. The story has it that the logician Frege died of a heart attack when Bertrand Russell informed him of the paradox of the class of all classes which do not contain themselves. But mathematicians continued to use the notion of a differential despite Bishop Berkeley's slashing attack. Zero gradually achieved the status of a full fledged number even though contradictions can be derived if it is "misused". In the case of the differential and zero, the concepts were judged by the mathematical community to be more useful than dangerous.

judgment. In most cases, the group judgment is better than the typical individual judgment; and in theory at least, the group judgment can be better than the judgment of any member of the group.

4. Value and Interest Disagreement

When we turn from factual estimates to value judgments or conflicts of interest, as has been noted previously, there are no agreed on figures of merit which apply equally to individuals and to the group. Thus, the Emerson Principle cannot be used to sidestep the consistency conditions. As it turns out, there is a fairly straighttorward resolution of inconsistencies of the Arrow type which does not depend on performance criteria. If individual preferences are expressed as ordinal scales--i.e., some set of objects is selected as a reference set and preferences for other objects expressed by their location in the scale formed by this reference set--then it is feasible to construct a group preference scale that is compatible with the individual scales. Demonstration of this possibility is a principal topic of Chapter VI.

Since reference objects have a number of desirable features in themselves--they assure the stability of individual preferences, and form the bases for extending preferences to more numerical kinds of measurement--introducing them into the formal apparatus of decisionmaking appears to have multiple advantages beyond simply allowing consistent group preference scales.

In the absence of a figure of merit for group value judgments, it is not possible to assert that the group will be "better off" if it uses collective value judgments. There is a weaker form of n-heads rule that can be derived for collective value judgments, but an additional notion is requried, namely the notion of cooperative decisions.

It is useful at this point to have some additional terminology. An individual decision can be analyzed with the help of a decision matrix,

illustrated in Figure 1. There is a list of potential actions $A = (A_1,\ldots,A_n)$ (strategies, plans, policies, etc.) among which the individual can choose.* There are two properties required of this list: (a) each action must be feasible, i.e., the individual must be able to carry out any action which he selects, and (b) the individual must be able to select one action out of the list—the "free will" condition. The result of taking a given action is dependent upon a set of contingencies $\{E_j\} = (E_1,\ldots,E_m)$ (states of the world, uncontrolled events, etc.) The outcome of selecting action A_k and the occurrence of contingency E_j is designated O_{kj}. The set of contingencies is taken to be an event space in the probability sense, i.e., there is a probability distribution $P(E_j)$ that any given contingency E_j will occur, where these probabilities do not depend on the action taken.**

To complete the analysis, it is assumed that there is a value function (utility or payoff function) $V(O_{kj})$ which defines the value of the outcome O_{kj} to the individual.

The decision rule for a decision expressed by a decision matrix is select the action A_k which maximizes the expected value $\sum_j P_j V(O_{kj})$.

In the individual case, the value function V is interpreted as the value to the individual of a given outcome, and the probabilities $P(E_j)$ as the probabilities as seen by the individual.

In the group situation, each individual has his own matrix--a set of actions that he can take, and a set of contingencies which he perceives to be relevant.

*In some forms of decision analysis, the set of actions may be extended to a tree, i.e., a branching process in which options at a later stage are dependent on what has occurred before. Although this more extensive model has a number of valuable features, the critical issues for group decisions can be discussed using the simpler matrix description.

**In the tree version, the probabilities of events can depend on previous actions. Again, this more general possibility is not needed for most of the following discussion.

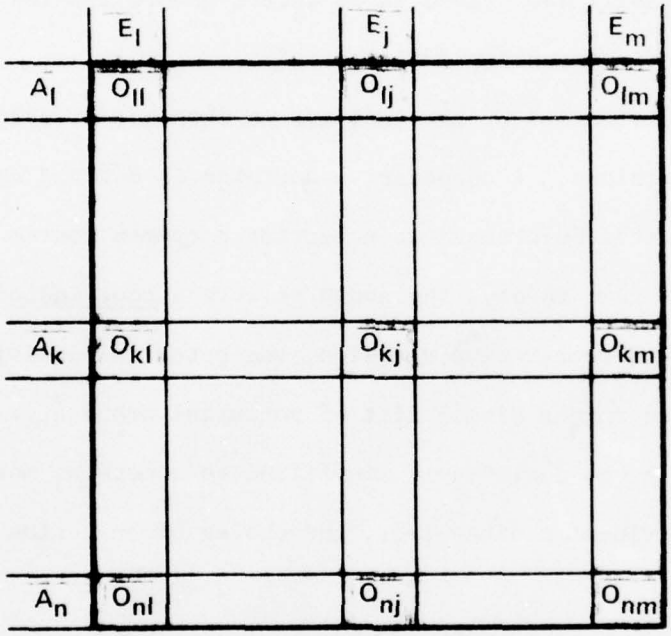


Figure 1. Decision Matrix

But now, in addition to the contingencies, the outcomes are determined by the actions of the other members of the group. The situation resembles a game in the sense of von Neumann and Morgenstern, but is a little more general.[3] In a von Neumann and Morgenstern game, the set of contingencies and the outcome matrix are common to all the players.

Since in this disaggregated case there is no common value function, and no common set of probabilities, there is no direct generalization of the maximization rule which defines a group decision rule.

A basic simplification of the analysis is obtained if attention is limited to cooperative decisions. A cooperative decision is defined as one in which the group is committed beforehand to selecting a common course of action. In the terminology of game theory, the group selects a coordinated strategy. In other words, in a cooperative decision, the potential individual courses of action are compiled into a single list of potential group actions. To this extent, then, the group decision is simplified to something more closely resembling an individual choice--i.e., the choice of an action out of a single list of actions.

Limiting attention to cooperative decisions omits a number of group processes that are relevant to group decision analysis. In particular, it slides over the question how the group "decides" to take a common action. However a broad area of important types of decisions remains. Typical decisions encountered in business firms and government agencies still remain, as well as those of most voluntary organizations.

The notion of cooperation defined above is quite narrow. Note that any of the four resolution techniques described in Section 2 can operate within cooperative groups as defined. The resolution procedures relate to the way in which a given course of action is selected; the cooperative assumption merely

determines that the selection will be from a single, common list. Thus, a group can commit itself to a common action, and still "allow" one individual to make the selection.

In the dictatorial "solution" to the cooperative decision problem, the set of contingencies, their probabilities, and the outcomes, are those perceived as relevant by the single decisionmaker. Similarly, the value function is one that the decisionmaker finds appropriate. But there is nothing in the dictatorial solution which says that the value function reflects the "selfish" interests of the decisionmaker; for most organizational managers, presumably the welfare of the organization, as well as their own welfare, would count in their value functions.

An even more drastic simplification is commonly made in formal treatments of group decisions, namely, that the entire decision matrix, except for probabilities on contingencies and the value function, is common to all members of the group. The assumption sweeps most of the problems associated with point of view disagreement under the rug. There is no good justification for the assumption, other than the fact that it bypasses many thorny problems. With that unadorned excuse, the assumption of a common decision matrix will be adopted for most of the formal models of group decision in this report.

The two assumptions of cooperative actions and a common decision matrix lead to a greatly simplified arena of disagreement. Disagreement is limited to probabilities for contingencies and to the value function. The set of actions $\{A_K\}$, the set of contingencies $\{E_j\}$ and the outcome matrix $||O_{ij}||$ are identical for all participants. Each individual, however, may have his own set of estimates $P_i(E_j)$ for the probabilities of contingencies, and his own value function $V_i(O_{kj})$ on outcomes.

The n-heads rules mentioned earlier furnish a basis for the resolution of disagreement on probabilities. Ordinal scales, as previously noted, allow the formulation of a consistent group preference scale. If in addition, each individual can express his value function in a numerical form which is linear in probabilities—technically known as a utility function—then the step to a numerical group value function is rather small. For example, if it is assumed that the group, whenever it is indifferent between two outcomes A and B, is also indifferent between A and any probability combination of A and B, then the group value function is just a weighted sum of the individual value functions. By a probability combination is meant, e.g., a lottery in which A will result with some probability p and B will result with probability 1-p. The assumption that the group is indifferent between an outcome A and a probability combination of A and any equivalent outcome B can be called the equivalence condition.

The assumption of individual utility functions is one that has been rather generally accepted by decision theorists. The equivalence condition for group preferences is more controversial. However, it can be bolstered by a form of n-heads rule. The weighted sum of the individual value functions minimizes the weighted total regret of the members of the group. An individual's regret is the difference between what he expects the group can achieve (in terms of his value function) and his expectation of the value of the action the group selects. The total group regret is just the sum of the individual regrets.

The min total regret result is rather weak since there is no separate criterion to indicate that the group is "better off" if it adopts a weighted average of the individual utilities as a group value function. It does give

a stronger justification for the weighted average than the equivalence condition alone.*

5. Some Limitations

Two formal limitations on group decisions were proposed above to simplify procedures for resolution of disagreement, namely, the assumption of a common point of view, and the assumption of cooperative, i.e., coordinated actions. In addition, there are some caveats concerning group judgment that are difficult to characterize in a completely formal fashion. Strictly speaking, these caveats are not part of formal group decision analysis; however, they are relevant to applying group decision procedures in practice.

N-heads rules are appropriate where there is no cost-effective way to obtain a more objective answer to a decision-relevant question. If the answer to a question can be obtained from well-validated sources, or by relatively inexpensive data collection, then the objective answer is clearly to be preferred to group judgment. This consideration does not create any conceptual problems for group decisions; in theory, at least, each individual should prefer the more solid objective information to his own judgment.

There is, however, another circumstance in which collective judgment may not be appropriate which does raise conceptual problems, namely, the case where the group knows so little about a question that a group judgment may be "misleading". This case is closely related to the situation variously labelled in the literature as "radical uncertainty", "unknown factors", or "incomplete

*In the demonstration of the possibility of a consistent group preference scale, and the demonstration of the min regret result, values and interests are not separated. Each individual is assumed to have either a single preference scale, or a single utility function, which expresses both values and interests. There is some reason for believing that separating values and interests can simplify resolution of disagreement, at least on values, and possibly on interests as well. Exploiting this possibility requires developing a group theory for multidimensional criteria. This topic is too extensive to include in the present report and will be treated in a separate publication.

information". For reasons which are still obscure, if an individual is poorly informed on a given question he is likely to give a biassed answer. Bias, in this context, does not imply distortion by the interests of the individual, although that factor may play a role, but only implies a systematic deviation from the true answer. Another way to express the same phenomenon is that if the individual is sufficiently poorly informed, he can be a counterpredictor; if he asserts A, then not-A is more likely to be true than A. Under these circumstances, for a yes-no question, flipping a fair coin will give a more accurate answer (on the average) than the individual's best guess.

The elementary n-heads rules apply to any set of judgments. Thus, no matter how poor the individual judgments, the error of the mean will always be less than or equal to the average individual error. This result does not guarantee, however, that the mean is free from bias. The group can be a counter-predictor as well as the individual members. The question thus arises whether there are circumstances under which, crudely speaking, the group would be "better off" to use a random device to obtain an answer to a decision-relevant question.

This topic is the subject of Chapter IV. In theory, both for individual's and groups, there are questions for which a better answer can be obtained by some variation on random choice. I have labelled such choices nominal estimates, i.e., estimates obtained by formula rather than by judgment. The practical problem raised by the possibility that nominal estimates can perform better than judgment is to find an indicator, i.e., some well-defined method of identifying the circumstances under which nominal estimates are called for.

Group judgment is to some extent a guard against individual bias. In Chapter V, the data from a study of probability estimates by a professional group is analyzed to show that individual judgments can perform more poorly

than chance; whereas the group does better than chance. Thus, any indicator would have to take into account the fact that the group compensates in part for individual error.

There are two potential indices that might be used as indicators of questions for which the group is a counterpredictor. One is the dispersion, as measured, e.g., by the standard deviation of the individual responses. The other is a self-rating, i.e., an appraisal by each individual of the degree of confidence he has in his estimate. In experiments, both the standard deviation and the average of individual self-ratings have shown high correlations with the accuracy of group estimates.[5] However, there are difficult problems of calibrating these indices so that a given standard deviation or a given average self-rating will indicate the same degree of information deficit across a variety of types of questions.

Chapter IV investigates the phenomenon of counterprediction and explores some plausible rules for generating nominal estimates. For one body of data, a form of nominal judgment (uniform weights for linear estimation) generates more accurate estimates than either individual or group judgment. To this extent, the potential value of nominal estimates in cases of low information can be illustrated. But the material in Chapter IV does not resolve the issue of specifying an indicator of counterprediction, nor does it give a complete set of criteria for selecting a nominal rule in practice. Chapter IV is thus an initial excursion into an area that requires additional research.

6. Summary Comments

The advantages of physical cooperation among individuals have long been apparent. Tasks which are impossible for individuals to perform can be carried out by teams of individuals. When specialization and division of labor are added to collective effort an even wider range of beneficial activities becomes

feasible. Potential similar advantages for collective behavior at the decisionmaking level have been obscured by a number of factors. Although large decisionmaking teams are commonplace—the managerial staff of any large organization is a case in point—their task has been viewed more as one of coordination than as one of directing the organization. By and large, the kind of thinking required to make non-routine decisions has been viewed as an activity unsuited to groups. Disagreement has been seen as a divisive force, requiring a "firm hand on the reins" to prevent disruption of organized activities. Emphasis has been placed on obtaining concurrence with policies—i.e., with obtaining the minimum agreement needed to maintain activity—rather than on an active participation in decisionmaking.

Unless I am mistaken, these attitudes are changing. In particular, the notion that decisionmaking can be diffused more widely within an organization is gaining ground. In part, this shift arises from a desire to improve the sense of participation, and hence the self-image of the members of the organization.

A number of the results presented in this report suggest that a stronger justification exists for expanding the role of group decisionmaking. In particular, group judgments can be more accurate than individual judgments (on the average) and group values can be more stable, and more comprehensive than individual value structures. Presently identified criteria for selecting a specific group value function are not wholly demonstrative. But they are a good deal stronger than simply finding some kind of accommodation among the disagreeing interests of the members of a group. For example, the fact that decisions based on the average of individual values minimizes the group regret may not appear particularly impressive; however, for those cases where a group

has decided to adopt a collective decision procedure, the minimum regret rule is a non-trivial incentive for choosing the weighted average value function.

The approach to group decision elaborated in the chapters which follow is necessarily elementary in character and scope. It appears likely that more powerful methods of aggregating individual judgments, and more intuitively plausible bases for generating group value functions will be uncovered in the near future. The collective approach to group decisions appears to offer a framework in which such improvements can be meaningfully pursued.

CHAPTER II. INDIVIDUAL ESTIMATION

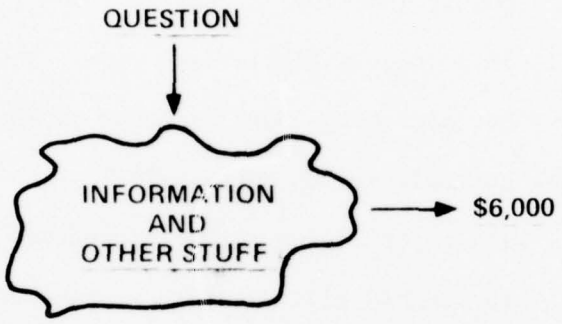
1. The Estimation Process

Although this book is primarily concerned with group judgment, some attention must be given to the role of the individual. Individual judgments are the basic ingredients of the group process. For decisions involving uncertainty and disagreement, the quality of the individual judgments is a fundamental limiting factor on the quality of the group decision. A basic theme of the book is various possibilities for using the group process to improve individual judgments. But roughly speaking, if the individual judgments are poor, the group judgment will be at best a little less poor. The old adage "You can't make a silk purse out of a sow's ear" is about as applicable to improving judgments as adages normally are to anything.

In this Chapter I will be examining rather elementary kinds of individual judgments, those which are roughly equivalent to simple declarative statements. In addition, the discussion will be restricted to factual judgments. There exists a fair amount of theory and some relevant experimental data concerning such judgments. In the following exposition I have been highly selective, dealing only with those approaches which I have found valuable for assessing group decisions. There is a much richer body of theory and experiment dealing with cognitive psychology that is in some sense relevant. Hopefully someday all of that will help illuminate the stubborn obscurities that plague the topic.

It is helpful to have a crude diagram of what is involved in making an estimate. A common type of judgment is that in which the individual is asked a specific, numerical question, where he doesn't know the answer, but can make an "educated guess." That last feature implies that the individual has some relevant information "in his head," and that the information is sufficient

to generate a properly formatted answer to the question. Consider the following example which I cooked up to provide a vehicle for introspecting about what goes on in making an estimate. "What is the cost of a young, well-trained elephant, FOB Thailand?" If your background is anything like mine, you won't know the answer to that question; and yet with no great effort you can come up with a number. You may not be happy with the number, but that's another matter. In my case, the number that came into my head was $6,000. A crude diagram is suggestive.



The question triggers the recall of related information. The material in the amorphous box has been labelled "information and other stuff" for the obvious reason that the question "brings to mind" (at least to mine!) a quite amazing variety of relevant and not so relevant material. When I thought up the elephant question, images of steaming jungles, turbanned mahouts, a scene from the movie "Around the World in Eighty Days," teak logs being trunkled into piles, and a great deal more of highly colorful mental imagery flooded in. Somehow, all of that added up to $6,000.*

The example suggests some important considerations with respect to the idea of using individual judgment as a surrogate for data. Even for uncertain

*A phone call to the Thai consulate in Los Angeles elicited an estimate of $5,000 as a "round figure."

questions, a great deal of miscellaneous and possibly low grade material exists in the minds of suitable individuals. But none of it answers the question directly. The individual is needed to recall that material, and more importantly, to fashion it into a reply to the question. It is clear that this process must be at least as complex as diagrammed in Figure 2.

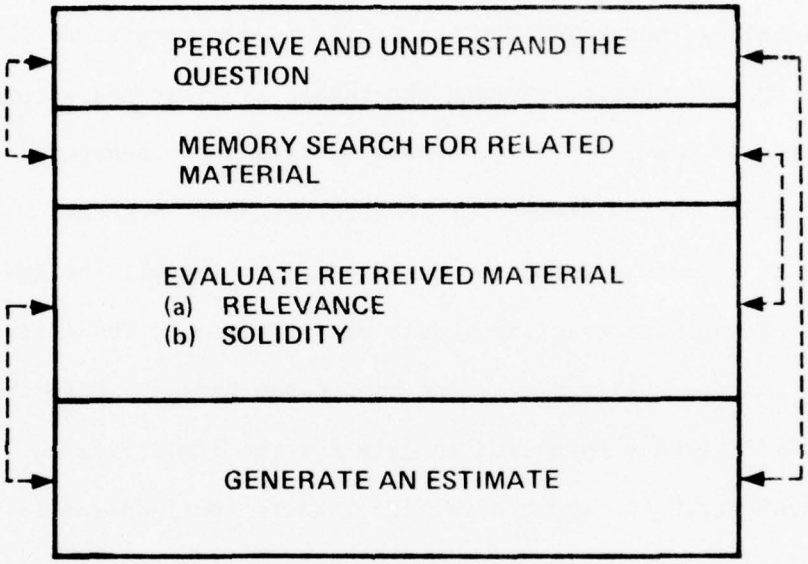


Figure 2. Basic Processes in Estimation.

The first step, perceive and understand the question, is probably a complex operation by itself. There is some reason to believe that in order to understand a question, the individual must have some relevant information.* Thus understanding probably interacts with the second step, retrieving related material. One of the gaping holes in the theories of estimation that will be elaborated below is the lack of any substantial treatment of these first two steps. Tversky's concept of anchoring suggests that the process is crucially affected by the way it gets started.[1]

As in the elephant example, the products of recall need evaluation and screening. There appear to be at least two basic criteria: (a) relevance — is the material useful in answering the question? (b) solidity — is the

*Cf. the discussion of universe of discourse in Section 2 below.

material substantial, or is it "flimsy?" The solidity rating has been dealt with in the literature, for example, under the topic "grounds for belief." A representive list of pertinent factors might be: the individual's direct experiences, congruence with other attitudes, consistency with the beliefs of acquaintances, perceived authority. In my own experience all of this is modulated by the question whether the recall is clear and strong or fuzzy and weak. There is a host of Ph.D. theses waiting to be generated focussing on the identification and measurement of the factors affecting the solidity rating.

Several indices related to the solidity rating will be examined in the sections dealing with specific models of estimation. These include probability estimates, self-ratings, and confidence ranges. Unfortunately, most of these indices have been studied only for the final judgment, and not for the "ingredients." In the general case, where the judgment is not part of the individual's reportory, but must be made up for the occasion, some of the material that is retrieved may be from old wives tales, from fiction, or from one's own imagination. (I had to reluctantly discard the scene from "Around the World in Eighty Days" as fiction—reluctantly because it was the only thing that came to mind that had a relevant number in it.) Some of the rest may be dubious, and hopefully some is fairly firm.

I sometimes refer to the final step, the generation of the estimate, as a minor miracle. Out of the culled heap of mostly qualitative matter, an estimate appears—in the elephant case a precise, though pretty shaky number. Equally amazing is the swiftness with which all of this occurs—a little less than 30 seconds for the price of the elephant. Figure 3 shows the results of a sequence of experiments on timed estimates. Several groups of upper class

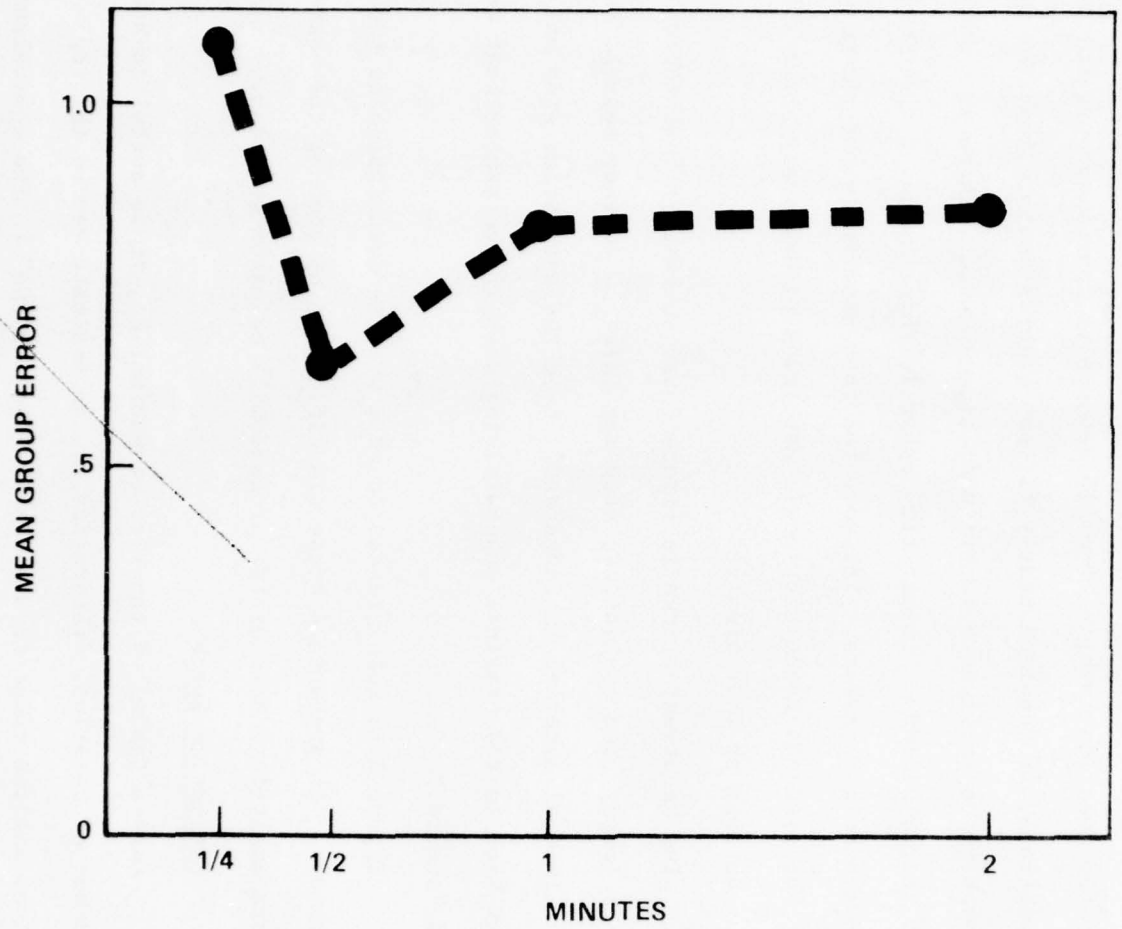


Figure 3. Effect of Time to Respond

and graduate students were asked a series of general information questions not unlike the elephant question (typical: "How many girls in the United States under the age of nineteen gave their status as divorced in the 1970 census?") They were given various time intervals in which to read and respond to the question, ranging from 15 seconds to four minutes. For many of the questions, it required nearly fifteen seconds just to read the question. The graph shows error as a function of time allowed. There is a clear minimum at thirty seconds. Longer time spent in thinking about the question led to less accurate answers. The data for four minutes is not shown because most of the students found they could not think productively about the questions for as long as four minutes.

The reentrant arrows in Figure 2 are to suggest that there may be feedback loops. The evaluations step may initiate further search, especially if the initial material is discarded. Even the estimation step may go all the way back to the original understanding step if the number that comes to mind is "absurd."

Figure 2 is not intended to be a precise description of the estimation process. It presents a crude classification of some of the basic features. More analytic models of the process will be taken up below.

2. Estimation Space

Before taking up theories of estimation, it is useful to have a certain amount of notation. Theoretically, an estimate can be the reply to any question, and can range from a simple "yes" or "no" to the equivalent of a book — "What do you think the world will be like in the year 2050?" Of necessity, the present discussion must be limited to questions less global than a world forecast. To the extent possible, overt replies to questions will be

designated by the letter R (R for "response" or "reply"), with various subscripts to indicate who is responding and about what. The letter Q will designate the belief of an individual which may or may not be the same as R. The letter I will designate information, usually the information on which a belief (and where appropriate, a reply) is based. The notation $(Q|I)$ and $(R|I)$ will designate the relationship that Q or R is based on I. This relationship is not well defined, a fact that will be occasionally embarrassing. At times the relationship will be treated as a relative probability, indicated by $P(Q|I)$—the probability of Q given I—but usually the relationship is not that of a probability.

In addition to the individual's estimate there is a true answer to the question. Loosely, we say the individual response is correct if it corresponds to the true answer. In general, the true answer will be designated by variants of the letter T. Thus, if the response is a magnitude estimate like the price of the elephant, then T is the actual price. The notation becomes more complex (and even controversial) for some types of estimates, especially probability judgments. This is further compounded by the fact that for the interesting cases the true answer is unknown. This topic will be elaborated in Chapter III.

A question implicitly determines a set of possibilities concerning the world and a set of possible responses. These can sometimes have identical structures. The question "will it rain tomorrow?" identifies two possible states of the world—rain or no rain—and two possible replies, equivalent to "rain" and "no rain." However, the question "What is the probability of rain tomorrow?" can be interpreted in two ways. On one interpretation, there are still the same two possibilities, rain or no rain, but an infinite number

of possible replies namely, any number between 0 and 1. On the other interpretation, the infinite set of potential probabilities are possible states of the world. The set of possibilities concerning the world will be referred to by the slightly fancy term "event space," and the set of possible replies by the term "response space." The event space will be designated by E, again with subscripts to indicate specific events, or sometimes by U (universe of discourse). The response space will be designated by R.

A basic property of either event or response spaces is the amount of structure that is imposed on the space (usually by definition) prior to asking a question. The simplest structure for either is a list of miscellaneous possibilities. For example, the question "Who will be the next president of the United States" may refer to a list of several names. Clearly, the order of the list doesn't matter. However, even such an elementary set of possibilities will be expected to have a minimum of structure; namely, the items on the list will be considered separate or exclusive. This is not logically necessary, of course, but for most practical situations, it would be awkward to have two different items (two different labels referring to the same alternative).

The structure of an event space can range all the way from the simple list to highly complex mathematical frameworks, e.g., the input-output coefficients of the world economy in the year 2025.

There is usually a strong coupling between the event space and the response space. They may have the same structure, or if not, the former sharply delimits the latter. One of the thorny issues in estimation theory is how to deal with the real life situation where the coupling breaks down. To the theorist a reply to the question "What is the probability of rain

tomorrow?" like "The probability of rain is .7 and the probability of no rain is .6," won't do. It is logically unacceptable that the probabilities of two exclusive events add to more than 1. Nevertheless, such "inconsistent" responses are encountered frequently in the psychological laboratory, and in councils of industry and government—usually not in such a bare form, but often only thinly veiled. Generally I will assume that the logical properties of the response space are consistent with those of the event space; but for a general theory it is necessary to allow the possibility that the estimator does not know (or "slips up on") the structure of the event space.

Another relevant aspect of event spaces relates to the numerical properties of the structure. It is often convenient, and at times essential, to quantify the event space, i.e., to describe the set of possibilities by one or more scales. Usually this is done as a matter of course where the question is inherently numerical. But in many instances, quantification has additional value within the context of group decisions. It is much easier to aggregate numerical judgments than purely verbal statements.

The topic of quantification is a whole field of investigation of its own. For the purposes of this book, the central issue is the degree to which available procedures justify the mathemetical operations performed on the numbers. A common classification of the possibilities is:

1. Nominal scales: A nominal scale consists of the assignment of numbers to items on a (mathematically) arbitrary basis, to be used as tags or names. For example, the list of presidential candidates could be numbered "in order" and thereafter the candidates referred to as "number one" etc. The numbers assigned in this fashion are traditionally assumed to have no mathematically interesting properties, other than being distinct. I am

inclined to think this attitude overlooks some of the advantages of nominal scales. E.g., a nominal scale can furnish counts ("have we voted on all the candidates?"). But from the standpoint of a scale of measurement, the nominal assignment of numbers is not equivalent to the imputation of some quantity to the items.

2. Ordinal scales. An ordinal scale consists of a relation which puts the items in a well-defined sequence. Typical relations are: greater than, better than, later than, more costly than, etc. Numbers may be attached to the items, to form an ordinal scale. If $N(x)$ is the number attached to item x and $N(y)$ is the number attached to y, and x has the given relation to y, then $N(x) > N(y)$. In general this is the only restriction on the numbers $N(x)$, so that any other set of numbers fulfilling the condition are an "equivalent" scale. In technical terms, the number assignment is determined only up to a monotonic transformation.

3. Interval scales. With interval scales, the differences between any two numbers are ordered. In effect the ratio

$$\frac{N(x) - N(Y)}{N(z) - N(w)}$$

for any two pairs of items (x,y) and (z,w) is fixed. In technical terms the scale is fixed up to a linear transformation; i.e., if $N(x)$ is an interval scale, than $AN(x) + B$, where A is a positive constant and B is any constant is an equivalent scale. A typical example is the ordinary scale of temperature which is fixed up to two reference points. Various scales of temperature are possible depending on the choice of reference points. The freezing point and boiling point of water (at sea level) is the common choice for everyday scales. Even with the selection of reference points, there is still the freedom of assigning numbers to these. We have two common scales, the

Fahrenheit and Celsius scales. The former assigns -32° and 212° to the two points. The latter assigns 0° and 100°.

4. Ratio scales. A ratio scale is one which is fixed except for one arbitrary (multiplicative) constant. Most common physical quantities are of this sort—length, weight, time duration, etc. Various alternative sets of scales—English, metric, etc.—are interchangeable by multiplication by suitable constants. In the case of ratio scales, only one reference object is necessary to fix the scale—the 0 comes for free. In technical terms, there is a fixed, absolute zero.

5. Absolute scales. An absolute scale is one for which there is no freedom whatsoever; the scale is completely fixed. The best known scale of this sort is the scale of cardinal numbers—those used for counting, tallying and so on. Another absolute scale is probability. The reference points, 0 and 1, are fixed. This feature of probability plays an important role in trying to tie the theory of subjective probability to the theory of objective probability measures.

In addition to these 5 typical kinds of scales, there are many variants. In the discussion of group utilities, the kind of scale obtained by introducing reference points for ordinal scales will play an important role in resolving inconsistencies between individual utility judgments. Psychologists and social scientists frequently use category scales, i.e., a sequence of "soft" reference points specified by verbal descriptions such as very desirable, desirable, neutral, undesirable, very undesirable. Numbers may be attached to these categories, very desirable = 5, very undesirable = 1, etc. Whether manipulating these numbers in usual arithmetic fashion—e.g., taking averages or standard deviations—is justifiable depends on properties of the numbers that are rarely tested in practice.

One of the guiding principles of the application of group judgment to uncertain problem areas is the availability of a remarkably rich assortment of judgmental scales. Humans have a rather astonishing ability to quantify practically any aspect of a problem, at least in a rough "intuitive" way. However, whether the numbers generated by human judgment have the necessary properties to justify treating them as mathematical entities requires demonstration. This is just as true where the individual is trying to estimate a well known physical quantity, such as length, or time duration, as it is when the quantity is "subjective," such as desirability. The psychological magnitude defined by the judgments may not have the same mathematical properties as the physical magnitude being estimated. We will encounter a situation of this sort in the psychonumeric phenomenon discussed below.

One final topic needs to be examined before leaving the discussion of estimation spaces, namely, the identification of a universe of discourse. This topic is full of obscurities and quasi-paradoxes, and I wouldn't bring it up at all if it were not one of the crucial aspects of increasing the solidity of judgment. The question is, how are the boundaries of the estimation space delimited—how can one specify what is to be included and what is to be left out? One appealing point of view (at first glance) is the straightforward proposal: Why leave anything out? Why not specify the elements of the problem you are interested in, and then sweep everything else into a "throw away" category "everything not included in the above"? In this way you don't clutter up your event space with every possible state of every possible universe, but at the same time, you have at least a weak guard against omitting a crucial feature that is not initially apparent.

Unfortunately, this ingenuous suggestion runs into a host of difficulties when the event space is uncertain. One type of trouble is illustrated by the

well-known paradox of confirmation, as formulated by Hemple.[2] A general statement such as "All ravens are black" is logically equivalent to the contra-positive "All non-black things are non-ravens." This creates no difficulties as long as we are dealing with a tidy universe of true-false assertions. However, if we are concerned with the messier world of incomplete information, and examine the equivalence of these two with respect to confirming evidence, an embarrassing situation arises. According to well-established custom, anything that is both a raven and black confirms the direct formulation of the general statement. But by the same custom, anything that is both a non-raven and non-black confirms the second, and hence by logic, confirms the first. Thus, if we go down to the beach, each brown grain of sand is a confirming instance for the assertion "All ravens are black." We have a ready made store of billions of confirming instances!

There doesn't seem to be much doubt that the pathology here is related to the universe of discourse for the sentence in the doubtful mood. Extending the universe of discourse to include a beachful of sand results in allowing non-relevant cases. In fact, at first blush, something which is neither a raven, nor black seems to be beside the point, and you might want to contract the universe of discourse to include only those things which are either ravens or black. You are forced by logic to do something drastic, because the negation sweeps in the whole remainder of the universe. However, that won't do either.

To see this, we have to generalize the problem slightly. Suppose you are a psychology graduate student, doing research for your Ph.D., and you want to do experiments to establish (or, heaven forbid, reject) an hypothesis you have conjectured to the effect that a given procedure A will influence positively

the performance of a task, B. You divide people into two sorts, those that have received treatment A and those who haven't, and redivide them into those who exceed a given criterion on task B (call them B+), and those who don't, call them B-. You recruit a group of students from the department subject pool (students who earn departmental brownie points by volunteering for duty as subjects in experiments) and proceed to expose them to treatment A and count the number of B+'s and B-'s so produced. You write it up, with χ^2 for one degree of freedom duly noted, and your lab advisor hits the ceiling. Why? You didn't have a control group.

The situation can be diagrammed by the matrix

	B+	B-
A	x	y
not A	z	w

You obtained results for the first row, x and y. But nothing in that experiment guarantees that the proportion x/y is not precisely the proportion you would have observed if the subjects had not been exposed to your treatment; perhaps x/y = z/w, in which case, the treatment has no effect. But notice, if you now dutifully go back to the lab and run a control group (no treatment) and count z and w, you are doing exactly what seemed so strange for the ravens and blackness. Your hypothesis is that A has a positive effect on B+; if there are no non-A's that are B-, your hypothesis is in trouble.

The apparent difference between the two cases stems from an illusion induced by the universal form of the statement "All ravens are black." If you restrict your universe to ravens, then your assertion is equivalent to "Everything is black." If you allow some non-ravens, then there had better

be some non-ravens which are non-black, or the assertion is still "Everything is black."

The embarrassment I mentioned is just that there are no rules that anyone has been able to think of that do not outrage logic and at the same time exclude the silly consequence of accepting a grain of sand as a confirming instance for "All ravens are black." Actually, the Ph.D. student is saved from thinking this through by the fact that the psych pool is a well defined universe. That is one reason that many "hard-headed" people in the "real world" are dubious of the "transfer of laboratory findings."

The paradox of confirmation is perhaps the most dramatic pathology connected with selecting a universe of discourse, but there are others. If you look into logic texts, the requirements for a specified universe of discourse is usually stated.[3] In some, the possibility of getting into trouble if the universe is interpreted too generously is given lip service, and is then promptly dropped. You will see illustrations of the Venn diagram for a logic problem that look like Figure 4 where the square box is the universe, and the ovals are the classes of things of interest. Illustrated is the statement, "All ravens are black"—the oval representing ravens is included in the oval representing black things. Also illustrated is the statement "Some swans are black."

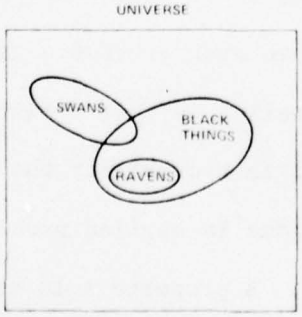


Figure 4. Normal Universe of Discourse.

Interestingly, you never see a diagram that looks like Figure 5 which shows what is going on outside the neat world you have enclosed. You never meet statements like "All ravens in my world are black (but some outside are not)."

What the logic texts never deal with is what happens if you change the boundaries, if you expand or contract your universe. In pure logic that is not an exciting question, especially if you are careful to keep all the classes you are interested in well inside the boundaries. Relations definable in pure logic, such as class inclusion, overlap, exclusion, and the like, remain invariant under changes of the boundary. This is quite different if the universe is uncertain, and you are concerned with things like confirmation or probability.

Ordinarily, the probability of the universe is defined as 1. The Venn diagram is a useful device to illustrate that the probability of a class can be represented by the ratio of its area to the area of the universe. Suppose the universe is the class of people living in the United States. The probability that someone living in the United States has a Ph.D. is rather small, which can be represented by a tiny area. The probability that someone in the United States is female can be boldly represented by a line cutting the universe in half. Obviously, if you expand the universe to the population of the world the probabilities of some classes are going to change, e.g., the probability of making an income over $10,000 a year.

What is much more interesting is that relations between classes which are thought to be fundamental in probability theory can change quite drastically. One of the basic notions in applied probability theory is <u>independence</u> and its side-kick <u>dependence</u>. A property A is said to be independent of a

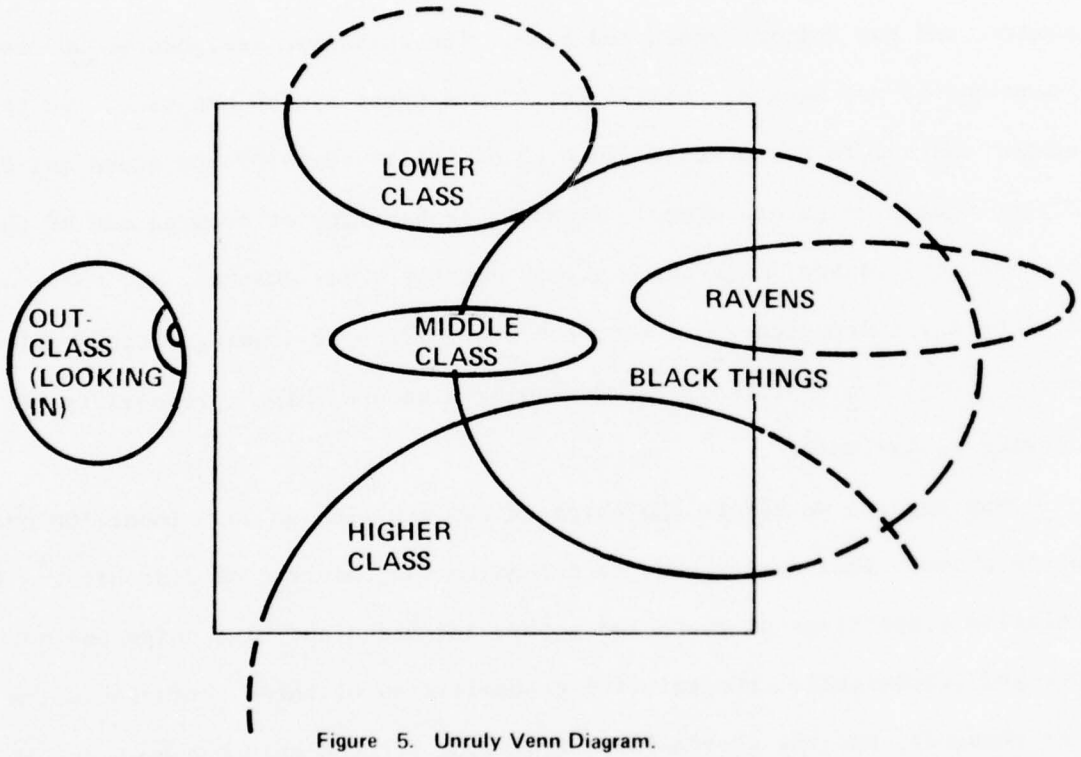


Figure 5. Unruly Venn Diagram.

property B if the probability of A, given that B obtains is the same as the probability of A in general. Or, equivalently, the two are independent if the probability of the conjunction A and B is equal to the product of the probabilities of the two separately.

Consider a classic type of example, an urn in which there is a certain number, say 100, of poker chips. Suppose there are two shapes, round and square, and two colors, green and blue. The chips are designed so 50 are round and 50 are square. Similarly, 50 are green and 50 are blue, and the shapes and colors are mixed so that 25 of the round chips are green and 25 of the square chips are green. Thus the probability of drawing any of the combinations—square and blue, e.g.,—is precisely one quarter, and the conditions for independence are met; the probability of drawing a square blue chip = $\frac{1}{4} = \frac{1}{2} \times \frac{1}{2}$ = probability of drawing a square chip × probability of drawing a blue chip.

Now suppose we dilute the chips in the urn with an additional 100 oval white chips. This corresponds to expanding the universe of discourse.* The relative proportions of round and square and green and blue chips has not changed (technically, the relative probabilities of these characteristics has not changed), but the probability of drawing a round chip has been cut in half, and the probability of drawing a green chip has also been halved. In addition the probability of the combination, round and green has also been cut in half. Thus the probability of drawing a square blue chip = $\frac{1}{8} \neq$ probability of a square chip × probability of a blue chip = $\frac{1}{4} \times \frac{1}{4} = \frac{1}{16}$. The characteristics are no longer independent.

*The same effect could be achieved by leaving the original urn unchanged, and introducing a second urn containing 100 oval white chips, then flipping a coin to determine which urn would be drawn from.

This example may feel artificial to someone who is accustomed to urn drawing illustrations of probability relations. But a little reflection on Figure 5 should make it clear that the seeming artificiality stems from posing a question which is outside standard logic and not from some insight stemming from the principles of logic.

There are some rules which have been established by logicians. These are rules which attempt to exclude the most serious form of logical pathology, namely antinomies or logical contradictions. Thus, potential contradictions associated with the words "true" and "false"* have led logicians to partition the universe of discourse into different levels of language, each of which can refer only to languages below it. Similarly, contradictions associated with a too-liberal usage of the notion class, have been excluded by a variety of restrictions such as partitioning classes into a hierarchy (theory of types) where a class can include only classes immediately below it in the hierarchy, or restricting classes to those that can be defined in a specific way starting from a fixed set of initial classes, and the like.

These grand logical restrictions are well to keep in mind, but usually they are not serious bugaboos for the practitioner. Not many practical

---

*If you allow statements of the form "This sentence is false," where the "this" refers to the sentence in quotes, then, if you assume the sentence is true, since it says it is false, it must be false. Conversely, if you assume it is false, it says it is false, and therefore must be true. The contradiction involving classes has the same sort of self reference. Suppose, following Bertrand Russell, you define the number three as the class of all classes that have three members. Now, there are certainly more than three classes which have three members, so the number three does not belong to the class of things with three members. Hence, the number three is a class which does not contain itself. Now contemplate the class of all such classes, namely the class of all classes which do not contain themselves, and call it A. Does A contain itself? If it does, by definition it doesn't; but if it doesn't, then by definition it does.

decisions involve investigating the consequences for policy of the world belonging to a class that doesn't include itself. But the type of puzzle represented by the paradox of induction, or the relativity of the notion of independence to the selected universe of discourse, are precisely the sort of thing that can bedevil practical decisions.

The notion of universe of discourse has affinities to the notion of "closed system" in physical science. Nowadays it is not too difficult to define a closed system; e.g., it can be defined as a region of space such that (during the time interval of interest) no energy flows either way across the boundary. Once upon a time, when the notion of energy was not as clear as today, it might have been much more difficult to say sharply what a closed system is. At the present time, there is no similar summative notion for decisions. The term "information" is beginning to assume some such role, and perhaps we could define a decisional universe of discourse as one for which no information flows across the boundary during the time period of interest. This is not proposed as a definition. Neither the term "boundary" nor the term "information" is sufficiently well defined to make a technical definition appropriate.

There is one attempt in the literature to pin down the problem under discussion more than I have indicated; this is the treatment of grand and small worlds by L. J. Savage.[4] His notion of small world – one that is decisionally manageable—is not too far from the notion of universe of discourse as I have loosely introduced it here. Savage assumes there is one grand world within which any small world can be identified by aggregating states of the grand world. For example, the small world "Rain tomorrow" or "No rain tomorrow" with potential acts "Start for a drive in the country

tomorrow morning" or "Stay home," can be defined by lumping under Rain all the possible conditions of the world compatible with rain, and similarly lumping with no rain all possible conditions compatible with no rain. The two acts are conceived as each having an extension which defines what will happen (outcomes) given any of the possible states of the grand world lumped under rain or no rain. Presumably two different small worlds are compatible if they have the same probability functions and the same utility functions in the grand world.

Essentially what Savage is suggesting is that all decision problems have the same universe of discourse (for a given decision maker) and universes of discourse for specific decisions be formulated by aggregating in some appropriate fashion the elements of the grand universe. Difficulties with this program will be discussed more fully in the section on personal probabilities. In essence, the grand world is simply too grand. Difficulties arise which are analogous to trying to measure the diameter of the physical universe with a yardstick.

To sum up this quite unsatisfactory discussion of universes of discourse: Before anything interesting in a formal sense can be done with individual or group estimates, an event space E and a response space R must be specified. This implies that a meaningful question cannot be asked unless the individual making the response knows a fair amount about the topic "a priori." For example, the question "How high is that tree?" cannot be understood without knowing quite a bit about measurement, about the heights of everyday objects, and something about tress—e.g., that they don't change their heights within the space of a few seconds. This prior knowledge is part of the universe of discourse implied by the question. At the moment, there does not appear to

be a well-defined technique for specifying this implicit knowledge, or determining the extent to which it limits the potential responses to the question.

Most of the discussion in this book will center around three types of event spaces: (a) a simple list, (b) a set of classes (types) of events, and (c) a simple Euclidean space; i.e., real number continua of one or more dimensions. However, it may be worth warning the reader that these rather tidy event spaces are drastic simplifications of the intricate conceptual contexts in which individual judgments are normally formulated.

3. Models of Estimation

We turn now to some specific models of the estimation process. As I commented at the beginning of this chapter, at the present time there does not appear to be a unified model of the entire process as outlined in Fig. 2. Rather, fragments of the process have been modeled. These fragments are of critical value in establishing some of the important properties of group judgment, but—as fragments—leave some major gaps when it comes to formulating a well-rounded set of guidelines for group judgment.

Instrospection gives a rather bewildering impression of the estimation process, especially of the generation step. At times the number "just comes." At other times, the mind appears to engage in a miniature reasoning process, frequently of a "narrowing down sort." The following is an outline of the way my daughter (who is left handed) arrived at the answer to the question "What is the proportion of people in the U.S. who are left handed?"

> "I know that left handers are not in the majority in the U.S., therefore it is less than 50%. Maybe 30%. But if the proportion were as large as 30% manufacturers would make a lot of things for

left handed people—like scissors. But these are rare. So maybe it's like half of 30%—say between 10% and 15%."

According to the Encyclopedia Brittanica, she should have reduced the figure by another factor of two. But in any event, her line of reasoning was relatively clear.

The view that "thinking" is a quasi-logical reasoning process has been adopted by a number of researchers in the field of artificial intelligence.[5] One of the basic tools in this approach is the elicitation of "protocols" — introspectively generated descriptions of how the individual deals with a problem. On the basis of these protocols, "heuristics," incomplete algorithms, are generated as approximations to the thought processes of the individual. The heuristics are incomplete in the sense that they do not guarantee the solution to a problem, but usually they do guarantee that if a solution is encountered during the process, the heuristic will recognize it as such.

The heuristic approach has achieved some success in dealing with well-structured problems like computer game playing routines for checkers and chess, and theorem generators for elementary logic. The heuristic model fits a great deal that is observed in introspection, and the artificial intelligence program has generated a valuable stock of algorithms for attacking some kinds of problems, in particular, powerful routines for searching large spaces of possibilities.

Nevertheless, the heuristic model does not appear to be particularly useful at this stage of the game for dealing with the problem of group decisions. There appear to be two reasons for this. In the first place, the kinds of estimates required for decisions have a remarkably miscellaneous quality. Each question, viewed as a reasoning process, requires a separate logical

pattern which is surprisingly ad-hoc. There does not appear to be a formal way to deal with the wide variety of miniature models encountered in practice.

The second reason is more fundamental. The heuristic approach requires that the estimate be posed as a problem, i.e., as a set of conditions for which a well-defined solution can be described. Estimation can rarely be couched in this form. Perhaps the most frustrating aspect of research with estimation is the lack of simple criteria to determine when the "right" answer has been obtained. It is even very difficult to formulate rules specifying when a given step is in the right direction. To the extent that such rules can be formulated, they are likely to be of a probabilistic nature -- "go this way and you will probably get a better estimate."

I'm not sure to what extent these comments represent limited imagination on my part, and to what extent they express objective features of human judgment. One major field of artificial intelligence research, pattern recognition, appears close to some of the approaches to estimation that will be elaborated below. Some of the most interesting experiments with communication in group processes have used tasks that can be expressed sharply as problems for the group to solve, with a clearly defined criterion of solution. Among these are the experiments by Bavelas with communication networks, where the task can easily be solved by a single individual having all the fragmentary information initially spread among the group.[6]

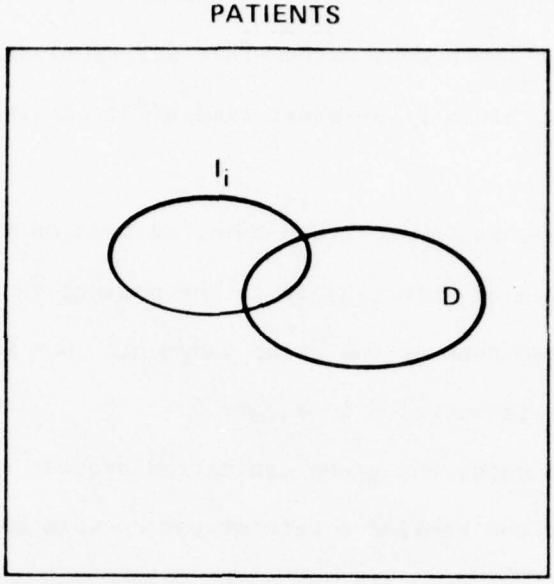
One elementary way to represent the information available to an individual is a subset in an event space. Suppose the question is (for a doctor) "Will this patient die?" The doctor has a certain amount of information about the patient -- symptoms, life stage, life style, etc. This can be represented in the space of patients as one type, or class, or set of patients, I_i, the

subscript i indicating that this information is available to doctor i. The question of interest is to what extent this set overlaps the set D of those patients who die (within a given short time after examination), as illustrated in Figure 6a.

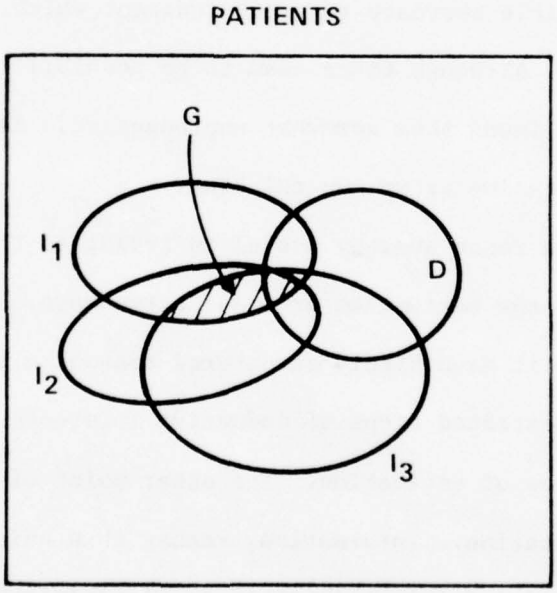
To extend Figure 6a to the group case, we need only assume that there are several doctors, each of whom classifies the patient in a set I_i and take the intersection of these sets as the group judgment. $G = \prod_j I_j$, where $\prod$ is the logical product, as illustrated in Figure 6b.

In this simple case, the group estimation process consists in determining the intersection of the knowledge sets of the doctors and relating this common set to the set of those who will die. The common set will be smaller than any individual set, and thus is more likely to be either totally within or totally outside the set of interest, D. The example is extremely elementary, but it illustrates a possible approach to group judgment which invokes only notions from formal logic. Although there seem to be possibilities inherent in such approaches, I have found them somewhat unproductive. As I remarked earlier, that may be a limitation of my own thinking.

I have found a rough analogy useful in trying to think about estimation. We can conceive of the estimation process in two ways. One is the traditional way of thinking of it as a highly structured reasoning process analogous to the precisely orchestrated steps of deductive inference. We could call this the algorithmic view of estimation. The other point of view might be called the chemical orientation. Information, rather than being put together in an intricate and stylized fashion, is mixed or blended like ingredients in a brew. The analogy is similar to the contrast between mechanics and thermodynamics. In mechanics, the detailed configuration of the elements and forces



(a)



(b)

Figure 6. Individual and Group Information Sets.

is used to predict the behavior of the system; in thermodynamics, on the other hand, gross averages of the properties of the particles are used to predict gross averages at a later time. Knowing the pressure and temperature and heat inflow of a system, it is not necessary to know the locations of individual molecules to preduct the pressure and temperature at a later time.

The analogy is perhaps only suggestive; but it allows using notions such as mixing, diffusion, adding or subtracting amounts of information and the like with a freedom of conscience that is hard to attain for one steeped in the rigors of formal logic.

4. Factor Models

One fruitful approach to a theory of estimation is the family of factor models. On this approach, the output of the memory search activity is a set $\{f_k\}$ of relevant factors (cues, components, items of information, etc.) The estimate R is assumed to be a function of this set of factors,

$$R = F(f_1,\ldots,f_k,\ldots,f_m) \tag{1}$$

Although in theory F could be about anything, only a small range of the possibilities has been explored. At one extreme are the single variable psychophysical laws, $R = F(x)$, where x is a physical magnitude (the stimulus) such as weight, sound intensity, and the like, and F is a power law (S. S. Stevens) or an exponential law (the classic Weber-Fechner law).[7] A relatively sophisticated formulation is the algebraic model approach of Anderson.[8]

The most widely exploited form of F is one of the simplest, namely the linear form. On this approach, the output of the evaluation step is a set of weights $\{w_k\}$ which perform the triple function of expressing the relevance of each factor, of discounting the factor for solidity, and scaling the numerical value of the factor to match the size of the required estimate. The estimate is then given by

$$R = \sum_k w_k f_k + c \qquad (2)$$

i.e., the estimate is a weighted sum of the factors, where c is an additive constant.

The linear model has been utilized to describe multi-cue perception,[9] complex value functions[10] and more general kinds of estimation.[11]

The notion of "solidity" has not received a large amount of attention in the psychological literature. Probably, the reason is that most experimental investigations have dealt with the case where the factors are "given" either as environmental cues in the case of perceptual tasks, or as experimenter-furnished values in other tasks. In these experiments, the factors are all "completely solid;" the only problem for the subject is to assess how significant they are for the given estimate (relevance). In the more general case we are examining, the factors are self furnished and the additional consideration of how well-established the factors are is a significant part of the task.

Unfortunately, on the linear model, there is no direct way to separate these three functions. It would be feasible in theory to assume that $w_k = g(s_k, r_k, m_k)$ where s_k is the individual's assessment of the solidity of the given information, r_k is his assessment of its relevance, and m_k is some assessment of the relative size of the factor and the desired estimate. As an example for m_k, consider the question "How many telephones are there in Africa?" One relevant factor is income. The estimator might reason, "The average income in Africa is very low—no more than a few hundred dollars per year—thus the number of telephones is probably small." The average income

is not the same "size" as the number of telephones. Thus a scaling factor is needed to bring the two in line.

Whether at this stage of the game it is worth introducing all of this complication into the model is difficult to determine. Individuals can differentiate between relevance and solidity, and the size feature is an obvious consideration. However, for purposes of applying the factor model to group decisions, it is difficult to see how much more intricacy than formula (2) can be used.

Some obvious variants of (2) appear worthy of notice. As we shall see, there is reason to believe that for many kinds of estimates individuals scale their responses on the logarithm of the quantity being estimated. In those cases the formula

$$r = \sum_k w_k \log f_k + c \qquad (3)$$

is more appropriate, where r is the logarithm of the individual's response, $R = e^r$.*

To compensate for different sized factors, and also to compensate for possible large differences in range or variability of the factors, a commonly used transformation is the z score

$$z_f = \frac{f - \bar{f}}{s_f}$$

where $\bar{f}$ is the mean of the values of a given factor f, and s_f is the observed standard deviation of the factor. The estimate then becomes

$$R = \sum_k w_k z_{f_k} + c \qquad (4)$$

*(3) is equivalent to the statement $R = c \prod_k f_k^{w_k}$, where $\prod$ denotes the product. Note that in this form, the weights appear as exponents.

For estimation problems where the f_k are reasonably well defined and objective, F can be identified using multiple correlation (linear estimation) techniques. Given a sufficiently large set of estimates by an individual of a specific type of quantity, an optimal (for the given data) linear model of the individual's estimates can be computed. For example, a number of experimenters have investigated the ability of college students and faculty members to predict the first year grade-point average of entering students, based on three factors; a score on a college entrance examination, the high-school grade-point average, and a rating of the excellence of the high school. The multiple regression of the individual's estimate against these three factors furnishes a set of weights which, with a little care, can be interpreted as the relative importance that the individual attaches to each factor.

The commonly employed *figure of merit (score)* for factor models is correlation with the true answer. Given a computed model, there are two potential scores, the correlation of the "raw" estimates of the individual with the true, and the correlation of the estimates computed from the individual's model with the true. On the estimation of grade-point averages, both students and faculty make a relatively poor showing. Average correlations range around .3, even for faculty with experience in college admissions.

A rather surprising result of these investigations has been that the individual's model uniformly outperforms the individual. This result has been labelled bootstrapping by investigators.[12] A common interpretation of this result is that individuals are more variable than their model.

In addition to the model of the individual's estimate, there is a corresponding model of the relationship between the true answer and the factors; thus we can write

$$T = G(f_1, \ldots, f_m) \tag{5}$$

Since (5) is the mirror image of (1), the combination of the two was dubbed the "lens model" by Brunswick. If (5) is also treated as a linear relationship, then it is clear that the correlation of R and T furnished by (2) can be no better than the multiple correlation of T and $\{f_k\}$.

A lively debate has been going on for a decade or more concerting the significance of research into factor models for professional decisions. It is a truism that the objective model (5) will outperform intuitive judgments, providing the model is correct. If individual judgments can be approximated very well by linear models, then an optimal linear model of the sort (2) will outperform the individuals. In many types of clinical judgments, linear models have proved to be good approximations, and, in fact, as the bootstrapping phenomenon indicates, better than the individual. This raises the question whether certain kinds of professional judgment can be replaced by models—by optimal objective models where sufficient data exists to compute the models, otherwise by models of the individual's judgment. A rather radical suggestion along these lines has been proposed by Robin Dawes that will be discussed in Chapter IV on nominal judgment.

These proposals appear to have made little headway in professional circles. The reason may be that professional people resist giving up certain roles, or just cultural lag, or possibly the fact that professional judgment usually cannot be reduced to a few well-specified types of estimates. The doctor must not only decide how sick a given patient is on the basis of a pre-specified set of symptoms, but also which set of symptoms to examine, what course of treatment to undertake, when to terminate treatment, and the like. There are active investigations underway studying whether these broader

contexts can also be reduced to well-defined models, either based on objective data, or on professional judgment, or both.

The set of issues raised by efforts to model professional judgment are all relevant to the "one-head rule." In the most general case, where the judgments of interest cannot be embedded in a well-defined family of judgments, and where the information is miscellaneous and non-objective, attempts to model the process would probably not be productive; each estimation task would require its own special model. However, this is not definitive. It is possible that even in these extreme cases, attempts to identify at least the major factors which influence the judgments, and to formulate a rough, linear model, may produce results which are more accurate, and more reliable (in the sense of including less random variation) than less systematic methods of arriving at estimates.

5. Probabilistic Models

One feature that appears to be lacking in the factor theory of estimation is an explicit statement of the degree of certainty of the judgment. An individual can take account of his own uncertainty (in the ingredients) via the weights he attaches to factors, and formulate his estimate accordingly, but the overall degree of certainty is not transferred to the final response. Since it is clear that the judgments of greatest interest are those which are plagued to some extent by uncertainty, many decision analysts have emphasized probabilistic judgments which contain an overt expression of the estimator's certainty.

Probabilistic estimation is a different species from magnitude estimation. Most of the theory and experimentation associated with probability estimation arise from a different context, namely, theories of rational

choice. The latter have grown out of the theory of rational economic behavior. By and large there has been less concern with the generation of estimates, and more concern with the explicit representation of uncertainty, the consistency of separate but related estimates, and revision (updating) estimates based on additional information.

These emphases have resulted from the close association of the theory of a decision with the calculus of probabilities. The calculus is not a theory of specific probabilities, but a statement of the relationships between probability assertions. Like formal logic, the calculus of probabilities is empty. It does not deal with the correctness of specific probability assertions, but rather is concerned with questions such as: given the probabilities of some events, how do you compute the probabilities of other, related events?

The approach to probabilistic estimation most closely related to decision analysis is the theory of personalistic or subjective probability associated with the names of Ramsy, de Finetti, and Savage.[13] There has been a welter of discussion about the signification of the probability estimates defined by this theory—do they denote degrees of belief, propensities to wager, shadow prices (trade-off weights) on events, and the like.

The question of signification has been complicated by an additional issue, namely the so-called problem of the probability of a single event. For most objective theories of probability, events which can be assigned a probability are repeatable. Thus a coin can be flipped (theoretically) an indefinite number of times. Subjective theories have been applied to events which—at first glance—are not repeatable. In fact, the theories were developed in part to meet an apparent requirement for dealing with uncertain

but non-repeatable events. For example, if we ask, "Will there be a major nuclear war between the United States and the Soviet Union during the next twenty years?" there is no question but what the reply is uncertain. Yet a major nuclear war within the next twenty years is not a repeatable event in the same sense in which a toss of a coin showing heads is repeatable.

The point of view I take on this topic is that so-called non-repeatable events are theoretically repeatable. Thus, one can imagine a super entity (cosmic scientist) conducting an experiment in which a set of earths is configured to resemble the earth at present, including its human population and political structure, and the entire set is allowed to run on for twenty years, while the super being carefully tabulates the number of earths on which nuclear wars occur. This grisly gedanke experiment doesn't appear to violate any logical laws, and possibly no physical laws. Some events are technologically repeatable (for present day humans) like the flip of a coin; for others, there are continuing systems in which the events in fact repeat, like various kinds of telephone calls; still others are repetitious, like tides or seasons; some are repeatable, but very rare, like Richter magnitude 10 earthquakes — none has occurred in recent history. The interaction of history and possibility provides a rich and fuzzy conglomeration of event types. To try to organize all of these into one neat concept like the collective of von Mises[14] is straining too hard.

Of particular interest are the nonobserved events which have very low probabilities, but are not necessarily impossible, like the Ricter magnitude 10 earthquake, or the delightful example of Frederick Mosteller in a reference that now escapes me of the likelihood that a human being (under present circumstances) will live to be 1000 years old. If the present distribution

of ages is taken seriously, then the probability of a millinarian is not zero--though very, very small. In order to deal with these types of events as "repeatable" it is necessary to envisage hypothetical sequences of events, in short to perform gedanke experiments. The gedanke experiment has a long and fruitful history in the physical sciences. Reluctance to use the device in the social sciences stems, I suppose, from the fragmentary condition of theory, and the fear that the imagination can run wild with no firm theoretical constraints. But the theory of probability is relatively well advanced, and gedanke experiments can be formulated in a fairly well-disciplined manner.

The contention that so-called non-repeatable events are repeatable in theory does not have the implication that probability is to be defined as a relative frequency. Relative frequencies are, on this point of view, one way to measure probabilities, in much the same sense that the position of a column of mercury in a thermometer is one way to measure a temperature. The temperature is not the height of the column of mercury. Of course, the statement that the probability of an event is p has the consequence (derived from the calculus of probability) that if the antecedent of the event (e.g., the flip of the coin for the event heads) is repeated, and the repetitions are independent, then the event will occur with the relative frequency p in the long run.

A probability estimate, on this point of view, is an individual's judgment of an objective property of a system. The estimate is no more subjective than an estimate by someone of the height of a visible, but unmeasured tree, or an estimate of the width of a river encountered by an explorer without a transit and chain.

The basic datum for decision analysis is that individuals can and do estimate the probabilities of relevant events in numerical terms. One

approach to a theory would be to start with that fact and ask "How good are such estimates, and how useful are they for making decisions?" An obvious problem, until recently, with this approach has been the difficulty of specifying what is meant by a good probability estimate of a non-repeating event. This issue will be examined more fully in the next Chapter on scoring methods.

The subjectivist theories start a little farther back, namely, with the fact that individuals make choices and these choices are influenced by their perception of the likelihood of events relevant to the choices. The theory lays down certain criteria for the choices to be "good" and investigates the consequences of these criteria for the properties of probability estimates.

Although the subjectivist point of view is somewhat at variance with the general perspective of this book, it is useful to have an exposition of the theory. It is a well thought out formulation of some of the criteria for good decisions, and it furnishes some useful conceptual apparatus for later investigations.*

The theory begins with the notion of choice, or alternatively with the notion of preference. The two are tied together by the assumption that if the individual is presented with a choice out of a set of alternatives, he will select the one he most prefers. The nature of the alternatives used as a starting point by various theorists have differed somewhat--Savage prefers starting with preferences among acts, for Ramsey it is goods, for others it is the outcomes of acts, and for some it is reward value, or

*The exposition which follows is basically that of L. J. Savage for the ordinal theory of subjective probability.

utility of states of the world. Most of these starting points lead to about the same conclusions.

For the present exposition, I will use a rather bland concept which will be designated by the term situation. A situation is a state of the world viewed by an individual from the standpoint of his interests. To give an illustration, a physicist might describe a chair as a complex assemblage of atoms. For the average man, most of that description would be irrelevant for everyday decisions such as whether to sit on the chair, or buy a new one and the like. This distinction is sometimes expressed by distinguishing between state variables and criteria variables, where the latter are the descriptors that are relevant to preferences. This formalization is pertinent, but more intricate than is needed for the exposition of the theory of subjective probability.* I use the term situation mainly to emphasize that it is the interests of the individual that define the relevant objects.

We envisage a set $X = \{x,y,z,....\}$ of situations. In general these will be potential situations--they are possibilities that may or may not be realized. The basic assumption concerning X is that a given individual has feelings about the relative desirability of different situations. This can be expressed by saying that there is a preference relation on X. For technical reasons it is convenient to start with the notion of "prefers or is indifferent" rather than strict preference. Thus $x \geq y$ means the individual either prefers x to y or is indifferent between them. (The

*For many situations of practical concern, the features which determine desirability are not known. In the most elementary cases, it is necessary to have a treatment which does not explicitly depend on criteria.

analogy with "greater than or equal to" is close, and is one reason for borrowing the notation $\geq$.)

The two basic conditions determining $\geq$ are:

P1a. Connexity. For every pair x,y in X, either $x \geq y$ or $y \geq x$.

P1b. Transitivity. For every triple x,y,z in X, if $x \geq y$, and $y \geq z$, then $x \geq z$.

P1a asserts that the preference relation is complete; for every pair of situations, the individual knows whether he prefers one to the other, or is indifferent. P1b is usually considered the property which makes preference relations rational. Thus, for example, with P1a, it asserts that preferences will not go around in a circle; it rules out x preferred to y, y preferred to z and z preferred to x.

Corresponding strict motions can be defined.

D1a. Strict Preference. $x > y$ means $x \geq y$ and not $y \geq x$.

D1b. Equivalence. $x \sim y$ means $x \geq y$ and $y \geq x$.

It is easy to prove that for any pair x,y, one of three things hold, either $x > y$ or $y > x$ or $x \sim y$.

Included among situations are a type that will be called contingencies. A contingency is a complex situation where the outcome depends upon the occurrence of some event. For example, the condition of Los Angeles in the year 1990 will depend upon the occurrence of a major earthquake between now and then. The state of the pocketbook of a citizen rolling dice in a casino in Las Vegas will depend upon the appearance of a seven in his first roll, etc. This type of dependence will be expressed by the notation $(x|E)$ read "the situation x will obtain if the event E occurs." There is nothing probabilistic about the dependence expressed by this notation. It could

be physical "Given a major earthquake, a majority of the brick buildings erected before 1932 will be heavily damaged." It could be the result of a social contract: "Given that the ball falls into a slot with the same number as the one on which you put a $1,000 chip, I will give you chips worth $36,000." Or it can be logical, "Given that Cleopatra was born a peasant and her nose was 3 1/2 inches long, her nose was 3 1/2 inches long."*

The expression $(x|E)$ is incomplete in that it doesn't say what will obtain if E does not occur. The notation $(x,y|E)$ will be used to express the more complete contingency $(x|E)$ and $(y|\bar{E})$ where $\bar{E}$ means "E does not occur." For example, with the social contract on the roulette wheel, x is "you get chips worth $36,000," y is "I take your chip and you get nothing." E is "the ball falls in the slot with the same number as the one on which you put a $1,000 chip," $\bar{E}$ is "the ball falls in any other slot." $(x,y|E)$ will also be called a contingency.

More generally, if $\{E_i\}$ is any partition of the universe of discourse (complete and exhaustive division) and $\{x_i\}$ a set of situations such that each x_i is contingent on the corresponding E_i, the expression $(x_1|E_1, x_2|E_2,\ldots,x_n|E_n)$, abbreviated $(x_i|E_i)$, represents an n-fold contingency. Note that $(x,y|E)$ is just $(x|E, y|\bar{E})$. Complex contingencies can be formulated where the situations are themselves contingencies. The prize in a lottery can be another lottery ticket. Unfortunately for the neatness of the theory, the distinction between situations which are not contingencies and those which are needs to be maintained for the early stages. Situations

*There has been a massive (and not completely benign) neglect of this relationship in the literature on the foundations of probability. It is not the same as implication. Some of the problems will be discussed below.

which are not contingencies will be called elementary. The idea expressed by a contingency has analogies in all the subjective probability theories. Thus, Ramsey uses the term wager, Savage uses act and consequence, von Neumann and Morgenstern, probability combinations. Other cognate terms are probability mixtures, lotteries, de Finetti's random quantities. I am sorry to add to all of this. There are some technical differences. For example, $(x,y|E)$ is not identical to $(x,y|F)$ providing $E \neq F$, even if the probability of E is equal to the probability of F.

To complete the building blocks, we need a set of events, $U = \{E,F,G,...\}$. U is the universe of discourse of events, and contains all the events worth considering for a given problem. The symbol U will also be used to designate the universal set, i.e., the set that includes all events. There should be no problem with ambiguity here, since most of the references to U in the following will be in the second sense. U is the same as the estimation (event) space discussed earlier. All of the problems associated with specifying estimation spaces apply to U.

Technically, U will be assumed to be an algebra of sets. This means U contains all the sums and differences of members of U, and it contains the null (empty) set 0. For every set E, U also contains $\bar{E}$ (the complement of E or not-E). In addition, we need the notion of joint occurrence of events, E.F (both E and F) and the disjunction E v F (either E or F or both). Since I will not be concerned with the fine structure of U, its properties will not be spelled out in axiomatic form. The interested reader can get details from any text on the theory of sets or any book on measure theory.

The distinction between U and X is not as sharp as one might wish. Normally, the distinction is made in terms of control; the events in U are

those which are not under the control of a given individual or group, those in X can at least be influenced, hence they are often called consequences or outcomes. However, this distinction breaks down in many types of analysis. The basic distinction on the present approach is one of evaluation. The items in X are those for which the individual has a clear preference—they "make a difference." The items in U make a difference through their effects on items in X.

To summarize, the elements of the theory are: the set of situations X, the preference relation $\succeq$, the set of events U, and the operation $(x_i|E_i)$ which generates contingencies. P1a and P1b specify the properties of $\succeq$. P2 extends P1 to contingencies.

P2. Closure for contingencies. Given any set of situations $\{x_i\}$ and a corresponding (equi-numerous) set of events $\{E_i\}$, which is a partition of U, the contingency (x_i,E_i) is in X.

P2 asserts that the individual has preferences for contingencies as well as for non-contingent situations, and in light of P1, any contingency can be compared with any situation. The significance of the wholesale independence of elementary (non-contingent) situations and events will be discussed below.

P3. Properties of $(x,y|E)$

(a) $(x,y|E) \sim (y,x|\bar{E})$

(b) $(x,y|0) \sim (y,x|U) \sim y$

(c) $(x,x|E) \sim x$

(d) $((x,y|E),y|F) \sim (x,y|E.F)$

(e) $(x,(x,y|E)|F) \sim (x,y|E \vee F)$

P3 expresses a number of properties which are immediate consequences of the "meaning" of $(x,y|E)$. It is redundant, in the sense that some of the properties can be derived from the others. However, to do this with entire rigor, it would be necessary to axiomatize the pertinent parts of set theory, which I promised not to do above. (a) simply emphasizes that in $(x,y|E)$, the situation x is contingent on the occurrence of E, and y is contingent on the occurrence of $\bar{E}$. (b) states the obvious, that any situation contingent on the null set is never realized, and conversely, any situation contingent on the universal set is always realized. (c) is equally obvious. A situation contingent on either an event or its negation is always realized. (d) and (e) express the appropriate representation of complex contingencies, as can be seen from the diagrams.

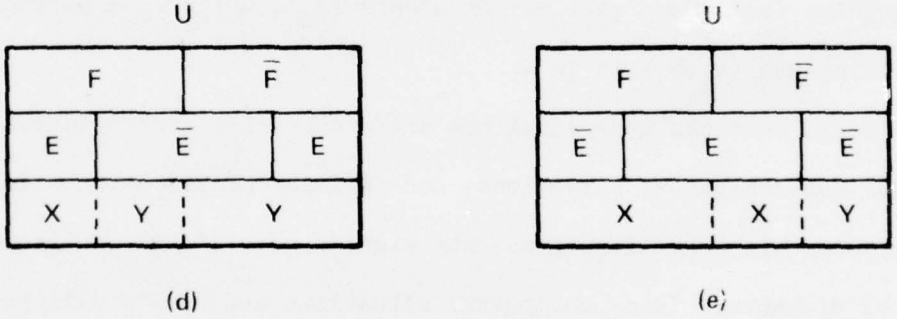


Figure 7. Compound Contingencies

P3b suggests a more general notion, namely that of a null event. Roughly, the idea is an event whose probability is zero, even though it may not be empty.

D2. E is null means $(x,y|E) \sim y$ for every x and y.

P4. Dominance. If $\{x_i\}$ and $\{y_i\}$ are sets of elementary situations, and $x_i \geq y_i$ for every i, then $(x_i|E_i) \geq (y_i|E_i)$. If, in addition, $x_j \succ y_j$ for some j, and E_j is not null, then $(x_i|E_i) \succ (y_i|E_i)$.

P4 is a familiar axiom in decision theory. It asserts that given two contingencies if the situation in the first contingency is at least as good as the situation in the second, no matter which event occurs, then the first is at least as desirable as the second. If for at least one of the events, the situation is strictly better in the first contingency than in the second, then the first contingency is strictly preferred to the second, providing, of course, that that event is not null.

The postulate is formulated for elementary situations rather than for all members of X for the simple reason that it does not hold if some of the situations are themselves contingencies. I don't believe this fact has been noted in most previous formulations of subjective probability theory; but it is difficult to be sure because many of the manipulations which depend on the logical properties of sets are left implicit. The difficulty will be illustrated by a simple example. Suppose one contingency is that the individual receives a dollar if it doesn't rain, otherwise nothing, so $C_1 = (\$1,0|\bar{R})$. The other contingency $C_2 = (x,0|\bar{R})$ where $x = (\$10,0|R)$ i.e., if it doesn't rain the individual will receive \$10 if it rains. This complex contingency can be evaluated using P4, but the reader is probably well ahead of analysis. The second contingency is worth precisely 0 despite the fact that $x > 0$. In general, P4 holds for both elementary situations and contingencies in the case that the events in the primary partition are independent of the events involved in the sub-contingencies; however the notion of independence cannot be defined with the conceptual structure developed up to this point. The notion of independence can be defined for arbitrary events only within the context of numerical probabilities, which we won't get to for several pages. The fact that independence cannot be

defined for ordinal probabilities, even with a fixed U, appears to be a deep property of the subjectivist approach which supplements the comments in Section 2 concerning the relativity of independence to a specific universe of discourse.

The next assumption is intended to give substantiality to the notion of one event being more probable than another. Suppose you are offered a choice between two contingencies, $C_1 = (x,y|E)$ and $C_2 = (x,y|F)$, where $x > y$. Since x is preferred to y, you would prefer that it be contingent on the more likely event. Thus, if you feel that C_1 is preferable to C_2, this is prima-facie evidence that you think E is more likely than F. As an obvious example, if C_1 is "you get \$10 if a head shows on a flip of a coin, otherwise nothing" and C_2 is "you get \$10 if a five shows on a roll of a die, otherwise nothing," you would in all likelihood select C_1.

This approach to perceived relative likelihood wouldn't be worth much if you changed your feelings depending on the kind of reward. P5 is intended to assure the requisite stability. As in P4, we have to restrict the postulate to elementary situations. It is patently false if asserted for situations which are themselves contingencies.

P5. Stability. If x,y,z,w are elementary situations and if $x > y$ and $(x,y|E) \geq (x,y|F)$, then if $z > w$ $(z,w|E) \geq (z,w|F)$.

The postulate looks more complicated than it is; it merely asserts that if you prefer the more valuable of two particular situations to be contingent on the event E rather than the event F, then for any other pair of situations you would prefer the more valuable to be contingent on E.

Savage defends P5 on the grounds that preferences for contingencies should not be dependent on the size of the prizes, no matter how small, as

long as one is definitely preferred to the other. I am inclined to think that the question of absolute size of prizes is a bit of a red herring, comparable to the question whether individual's make estimates which are "really continuous." In essence P5 assures that the following definition won't create trouble when situations are shuffled in contingencies.

D3. $E \geq F$ (read "E is at least as probable as F") means that, given x,y are elementary events and $x > y$, $(x,y|E) \geq (x,y|F)$.

The use of the same symbol $\geq$ to indicate the preference relation between situations and the relation *more probable* between events should not be too bothersome, since the two uses will be distinguished by lower case letters for situations and upper case letters for events.

In order to assure that D3 is not empty, it is necessary to assert the trivial assumption that there is at least one pair of situations x, y such that $x > y$. I'm willing to make that assumption without dignifying it with a P number.

P1-5 and D1-3 are sufficient to establish what could be called the pure ordinal theory of subjective probability. As we shall see in a moment, they determine for any two events E and F that the individual has a consistent judgment as to which is the more probable. This judgment lays out all events (in U) in a serial order, with the null event 0 at the low end, and the universal event U at the upper end. As should be the case, the disjunction $E \vee F$ of any two events is at least as probable as either, and either is at least as probable as the conjunction E.F.

Theorem 1. $\geq$ is a complete ordering for events, that is, it is connected and transitive.

Proof: Let E and F be any two events. For connexity, consider any pair of situations $x > y$. By P1, either $(x,y|E) > (x,y|F)$ or $(x,y|E) \sim (x,y|F)$ or $(x,y|E) < x,y|F)$. Using D3 the corresponding relationship is transferred to E and F. For transitivity, $E \geq F$ and $F \geq G$ means there is a pair x,y $x > y$ and $(x,y|E) \geq (x,y|F)$ and there is a pair z,w $z > w$, and $(z,w|F) \geq (z,w|G)$. From P5 we get $(x,y|F) \geq (x,y|G)$. Hence from P1 (transitivity) we conclude $(x,y|E) \geq (x,y|G)$, and D3 implies $E \geq F$.

Theorem 2. $U > 0$.

Proof: Assume $x > y$. $(x,y|U) \sim x$ and $(x,y|0) \sim y$ by P3b

Hence $U > 0$ by D3.

Theorem 3. $0 \leq E \leq U$.

Proof: Assuming $x > y$, $(x,y|0) \sim y \sim (y,y|E) \leq (x,y|E) \leq (x,x|E) \sim x \sim (x,y|U)$. The equalities are from P3, the inequalities from P4.

Theorem 4. $E \vee F \geq \frac{E}{F} \geq E.F$.

Proof: Consider the table

	$E.F$	$E.\bar{F}$	$\bar{E}.F$	$\bar{E}.\bar{F}$
C_1	x	x	x	y
C_2	x	x	y	y
C_3	x	y	x	y
C_4	x	y	y	y

If $x > y$, then C_1 dominates C_2 and C_3, which in turn dominate C_4, all by P4. The table entries, e.g., $C_2 = (x,y|E) = (x|E.F, x|E.\bar{F}, y|\bar{E}.F, y|\bar{E}.\bar{F})$ follow from the logical rule $E = E.F \vee E.\bar{F}$ and P3.

There is a tendency in subjectivist theories of probability to use the theory of ordinal probability simply as a stepping stone to the more familiar

numerical probability.* This may be too hasty a leap. As we have seen, numerical probability runs into difficulties as soon as we try to move from one universe of discourse to another. In addition, as will crop up later, numerical probabilities do not seem to rationalize some kinds of choice behavior when the amount of information concerning uncertain events becomes too sparse. We can ask, does the set of postulates and definitions we have just gone through remain valid for these "non-normal" conditions? We will return to this theme in Chapter IV on nominal judgments.

The final postulate which bridges the gap between purely ordinal probability and numerical probability is the somewhat more controversial "sure-thing" principle. This postulate introduces a form of strong independence between events and situations that furnishes the basis for the additivity of probabilities for exclusive events. To formulate the postulate in our notation we need an auxiliary idea.

D4. C and D agree on E if there is a partition $\{E_i\}$ of E, such that $C = ((x_i|E_i), R|\bar{E})$ and $D = ((y_i|E_i), S|\bar{E})$ and $x_i = y_i$ for every i. $R|\bar{E}$ and $S|\bar{E}$ are shorthand for "C and D can be anything on $\bar{E}$."

P6. (Sure-thing) If C_1 agrees with C_3 on E and C_2 agrees with C_4 on E and C_1 agrees with C_2 on $\bar{E}$ and C_3 agrees with C_4 on $\bar{E}$, then, if $C_1 \geq C_2$, $C_3 \geq C_4$.

The intention of P6 is probably clearer displayed in a diagram.

	E	$\bar{E}$
C_1	r	s
C_2	t	s
C_3	r	u
C_4	t	u

If $C_1 \geq C_2$, then $C_3 \geq C_4$

*"....for here I have little interest in qualitative probabilities, except as a foundation for quantitative probability." L. J. Savage,[13] p. 45.

Here r, s, t and u are not situations, but abbreviations for "whatever pattern of situations obtains for subsets of E or subsets of $\bar{E}$ as the case might be."*

The basic intent of the postulate is expressed by the diagram. Whatever makes C_1 preferable to C_2 must involve only E, because the two are identical on $\bar{E}$. But then C_3 and C_4 are also identical on $\bar{E}$, and therefore should be affected by exactly the same considerations that made C_1 preferable to C_2. The reason for the name "sure-thing" should be clear.

It is unfortunate that the postulate assumes such an intricate form, since the basic notion is quite simple. "The only features of two contingencies that make a difference are those parts where they are different."

P6 furnishes the theorem

Theorem 4. $E \geq F$ if and only if $E \vee G \geq F \vee G$, providing $E.G = F.G = 0$.

Proof: Consider the diagram, where as usual $x > y$,

	$E.F$	$F.\bar{F}$	$\bar{E}.F$	$\bar{E}.\bar{F}$	G
C_1	x	x	y	y	y
C_2	x	y	x	y	y
C_3	x	x	y	y	x
C_4	x	y	x	y	x

G is included in $\bar{E}.\bar{F}$ by assumption. $E \geq F$ if and only if $C_1 \geq C_2$ and $E \vee G \geq F \vee G$ if and only if $C_3 \geq C_4$. Set $E.\bar{F} \vee \bar{E}.F = H$ and $E.F \vee \bar{E}.\bar{F} = \bar{H}$, C_1 agrees with C_3 on H, C_2 agrees with C_4 on H, C_1 agrees with C_2 on $\bar{H}$, and C_3 agrees with C_4 on $\bar{H}$. Thus the conditions of P6 are fulfilled, and

*If r,s,t,u are construed as elementary situations, then P6 is just a special case of P4, dominance, since $C_1 \geq C_2$ implies $r \geq t$, whence $C_3 \geq C_4$.

if $C_1 \geq C_2$, then $C_3 \geq C_4$. Since the conditions are symmetrical, the reverse also holds.

Hence by D3 the theorem follows.

Theorem 4 is the analogue for ordinal probabilities of the additivity of numerical probabilities for exclusive events. It permits "cancellation" of G in the inequality E v G $\geq$ F v G. An important corrolary of Theorem 4 is

Corollary 1. E $\geq$ F implies $\bar{F} \geq \bar{E}$

Proof: Consider the table, with x > y

	E.F	E.$\bar{F}$	$\bar{E}$.F	$\bar{E}$.$\bar{F}$
C_1	x	x	y	y
C_2	x	y	x	y
C_3	y	y	x	x
C_4	y	x	y	x

Define H and $\bar{H}$ as in the proof for Theorem 4.

E $\geq$ F implies $C_1 \geq C_2$. C_1 agrees with C_4 on H and agrees with C_2 on $\bar{H}$. C_2 agrees with C_3 on H and C_3 agrees with C_4 on $\bar{H}$. Thus, by P6, $\bar{F} \geq \bar{E}$.

Turning to numerical probabilities, the elementary calculus of probabilities is remarkably simple. You can get by with the following three assumptions:

A1. $0 \leq P(E)$

A2. $P(U) = 1$

A3. $P(E \text{ v } F) = P(E) + P(F)$, providing $E.F = 0$

There are a number of routes to take to A1-3. I will outline what appears to be the simplest of the procedures. More complete treatments are found in Savage and de Finneti. D3 suggests a natural definition for the notion probability 1/2.

D5. $P(E) = 1/2$ means, given $x > y$, $(x,y|E) \sim (x,y|\bar{E})$ i.e., the individual is indifferent whether the more valuable alternative is contingent on E or $\bar{E}$.

Corollary 2. If $P(E) = 1/2$ and $P(F) = 1/2$, then $E \sim F$.

Proof: By corollary 1, if $E > F$, then $\bar{F} > \bar{E}$. By definition of $P(E) = 1/2$, $E \sim \bar{E}$, and similarly $F \sim \bar{F}$, whence $F > E$, contrary to assumption. The same reasoning rejects $F > E$.

Although it does not seem possible to define the notion of independence for pairs of arbitrary events within the ordinal theory of probability, it is possible to define the notion for the special case that one of the events has probability 1/2.

D6.* Given $P(E) = 1/2$, E is independent of F means $E.F \sim \bar{E}.F$.

D6 can be extended to express the notion E is independent on repetition, providing the probability of E is 1/2.

D7. Given $P(E) = 1/2$, E is independent on repetition means: Let X_n designate a conjunctive sequence of n terms consisting of E's and $\bar{E}$'s in any proportion and any order; e.g., an X_3 might be $E.\bar{E}.E$. For any n, and any X_n, $X_n.E \sim X_n.\bar{E}$.

Theorem 5. If there exists an event E such that $P(E) = 1/2$ and E is independent on repetition, then there exists a unique mapping of U onto the real interval, such that A1-A3 hold.

Proof: It is elementary, but tedious, to prove that the hypothesis of the theorem implies there is a 2^n-fold equipartition of U for every integer n. Denote a member of the 2^n-fold equipartition by X_n, and the logical sum of any m of these by $X_{n,m}$. For any F, either $F \sim X_{n,m}$ for some n and m, or

*This definition can be related to the usual definition of independence, namely $P(E.F) = P(E)P(F)$, by noting that independence implies $P(\bar{E}) = P(\bar{E})P(F)$, and if $P(E) = P(\bar{E})$, we arrive at D6.

there is an infinite sequence of intervals $X_{n,m+1} > F > X_{n,m}$. Define P(F) to be $m/2^n$ in the first instance, otherwise the limit of the sequence of intervals $(m+1/2^n, m/2^n)$ as $n \to \infty$. This definition maps U onto the real interval (0,1), with P(U) = 1. The mapping is unique in the sense that for any two events that have probability 1/2 and are independent on repetition the identical mapping is generated. A3 follows from Theorem 4 and the additivity of the reals defined as limits of sequences of intervals.

Theorem 5 motivates the assumption

P7. There is an event E, P(E) = 1/2, and E is independent on repetition.*

Although Theorem 5 is in some sense an adequate basis for numerical probabilities, it does not assure a total fit between ordinal probabilities and the numerical mapping. It implies that if P(E) > P(F) then E > F, but not the reverse. In particular, it does not exclude E > F and P(E) = P(F). To rule out this possibility, an additional assumption is needed. This assumption, ruling out infinitesimal differences in probabilities, is common in measurement theory.

P8. (Archimedean) If E > F, then, for G such that P(G) = 1/2, independent on repetition, there is an $X_{n,m}$, such that $E > X_{n,m} > F$.

---

*Most investigators in the foundations of probability would probably find P7 overly specialized, "weak," and possibly old-fashioned. The same may be true of P8, below. I have preferred these two to more powerful assumptions on the grounds that, given Theorem 4 (existence of an ordinal probability scale), the basic issue appears to be attaching a numerical scale to probabilities with some psychological content. P7 and P8 seem to express widely accepted attitudes about events like observing a head on the flip of a coin.

P8 assures that the strict inequality E > F means there is a finite difference between E and F. It also assures that E and F will be mapped onto different numbers.

This is probably a good place to list several definitions and formulae that can be derived from A1-A3 which will be used in later sections.

D7. Relative probability. $P(E|F) = P(E.F)/P(E)$

Read "The probability of E given that F occurs."

D8. Independence. E and F are independent means $P(E.F) = P(E)P(F)$, or equivalently, $P(E|F) = P(E)$.

F1. Extended rule of addition. $P(E \vee F) = P(E) + P(F) - P(E.F)$

F2. Rule of the Product. $P(E.F) = P(E)P(F|E) = P(F)P(E|F)$

F3. Rule of elimination. If $\{F_i\}$ is an exclusive and exhaustive partition of U, $P(E) = \sum_i P(F_i)P(E|F_i)$

F4. Theorem of Bayes. If $\{H_i\}$ is an exclusive and exhaustive partition of U, $P(H_i|E) = \dfrac{P(H_i)P(E|H_i)}{\sum_j P(H_j)P(E|H_j)}$

There is a certain reluctance to accept idealizations like an event with probability 1/2 independent on repetition as a basis for probability measurements.* Idealizations in other areas of measurement are not so suspect — perfectly rigid and indefinitely divisible rods, isochronous clocks, and the like. I would suspect that the reason is not so much the idealization as the fact that in practice, probabilities are not measured by comparison with some set of equiprobable events, but rather are measured by relative frequencies.

*"It might fairly be objected that such a postulate would be flagrantly ad hoc." Savage, Foundations, p. 33.

There is no standard chance device, e.g., a platinum-iridium penny, at the International Bureau of Weights and Measures at Sèvres, France.

The world described by postulates P1-P8 is a very simple place. It is essentially the world of the gambler, where the interesting events are loosely coupled to the interesting rewards. In gambling, the coupling is effected by a social contract, not by physical interaction. To dramatize this point, P3 says that the contingency $(x,x|E)$ exists, whatever x and whatever E. But suppose E is "The sun goes nova tomorrow." What possible contingency could there be that makes the outcome of the sun going nova the same as the sun not going nova?

Although there have been attempts to model the "real" world, where events influence the relevant outcomes in a direct physical way, to my knowledge none of these have been successful. Savage begins his theory with something that looks very much like the real world, but he has to abandon it rather quickly. He needs the notion of "constant act" — that is, an act that produces the same consequences irrespective of the state of the world — in order to formulate the equivalents of P4 and P6. Thus, his definition of ordinal probability is formulated within the (unexpressed) restrictions of an assumption that is very much like P3.

The situation in probability theory is not too different from what it is in many other measurement theories. It is recognized by physicists, for example, that the elementary definition of length in terms of juxtaposing a sequence of equal-length rigid rods is feasible only in a limited geographical region. To measure lengths over more extended regions, e.g., to the planet Mars, complicated apparatus and complicated theories must be invoked. To

extend measures to intergalactic distances, rather shaky assumptions concerning the period and intrinsic brightness of variable stars must be made.

There is no reason for suspecting that the world is any more tractable when it comes to probability measurements. Defining the elementary notion of probability measure in terms of gambling-like situations does not imply that the same type of measurement extends to any situation where we would like to use the term probability.

The subjective theory of probability goes well beyond simple estimation of probabilities and includes a relatively complete theory of individual decisions. The extension of subjective probability theory to include numerical utilities is a relatively minor step, and in fact in the form developed by Ramsey and De Finneti, the theory of numerical utilities precedes and forms the basis for the finalization of a numerical theory of probabilities. The intimate tie between subjective probability theory and the complete theory of decisions is both a strength — it allows displaying the role of probabilities in decisions in a simple way — and a source of awkward consequences. If human decisions as observed, e.g., in the psychological laboratory, do not accord with the theory, it is not always clear whether the disparity involves the narrower concept of perceived probability, or the more general theory of decision in which it is embedded.

As presented by its founders, and most of those who want to apply it, the subjective theory has been given a kind of universality which is unnecessary for our purposes. Thus, for true-blue subjectivists, any individual (at least any one beyond the age of accountability) has a clear perception of the probability (from his perspective) of any event whatsoever. Furthermore, this probability is precisely the "correct" probability to guide any of his

decisions, to which the probability in question is relevant. Both of these appear to be unnecessarily strong assumptions. For our purposes, it looks sufficient to assume that for some universes of discourse, individuals have relatively clear perceptions of the probability distributions on those universes. Whether the perceived probabilities are "correct" is a quite different matter. We assume they can be incorrect in much the same sense in which guesstimates of any other physical quantity can be incorrect.

6. Calibration

The subjectivist theory of probability is essentially a theory of consistent probability estimates. The tie between the estimates postulated by the theory and reality is loosely drawn. By and large those who wholeheartedly embrace the theory become restive — if not downright surly — when the subject of correctness of probability judgments is raised. In part this appears to involve a feeling that probabilities are not part of the world, but are measures in some not fully specified sense of the amount of information which an individual has concerning the predicted event. Clearly, two different individuals with different information may announce quite different estimates of the probability of a given event. There is no pathology in this — it is analogous to the fact that $P(E|F)$ may be quite different from $P(E|G)$.

Nevertheless, there is a straightforward sense in which an individual can simply be mistaken in making a probability judgment. Almost everybody is agreed that the French mathematician D'Alembert was mistaken in asserting that the probability is one-third of obtaining a head and a tail in two tosses of a fair coin. Furthermore, if he had bet on that assumption, everyone is agreed that his shirt would have been in jeopardy. The problem is

to find a clear and general way to state what is meant by saying an individual is correct or incorrect in making a probability judgment. If the weather forecaster says that the probability of rain tomorrow is .2, and it pours, what crime can we accuse him of? He didn't say it wouldn't rain, just that it was unlikely. And on almost any concept of probability, unlikely things must happen once in a while. This topic is explored more thoroughly in the following chapter on scoring.

A different approach to this issue that has received a fair amount of attention lately is the notion of calibration. Given a set of probability estimates by an individual, and a set of data concerning the occurrence or nonoccurrence of the predicted events, it is possible to get a rough idea how good the individual's predictions are. Thus, if a subset of the predictions is selected, all predicting some event with the same probability R, then over many such predictions, the relative frequency F with which those events occur should settle down to about R. More, generally, if the individual generates many estimates with different probabilities, the relative frequency with which the events occur should approximate the dotted 45 degree line in Figure 8.

In actual experimental studies, the observed results are usually quite far from the theoretically "correct" 45 degree line. In Figure 9 the solid line is a plot of the data collected by Capen.[15] The subjects were 43 engineers. Each subject answered a set of 120 questions. Responses consisted of a true-false judgment and a probability estimate that the selected response was correct. Responses were restricted to the round-numbers .5,.6,...,.9,1.0; the restriction to responses greater than or equal to .5 resulted from the assumption that a subject would select the alternative

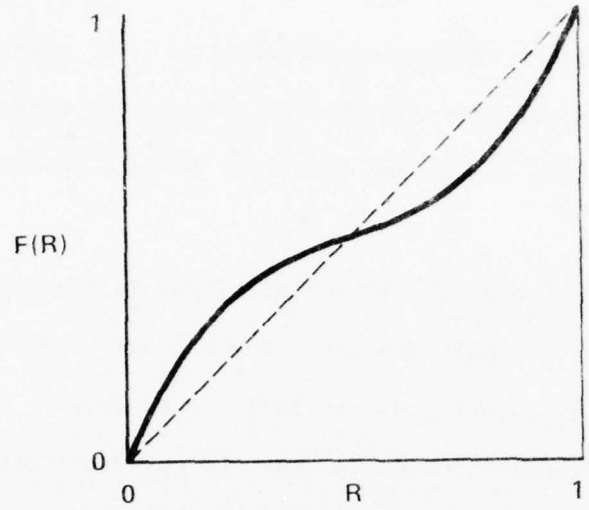


Figure 8. Typical Realism Curve.

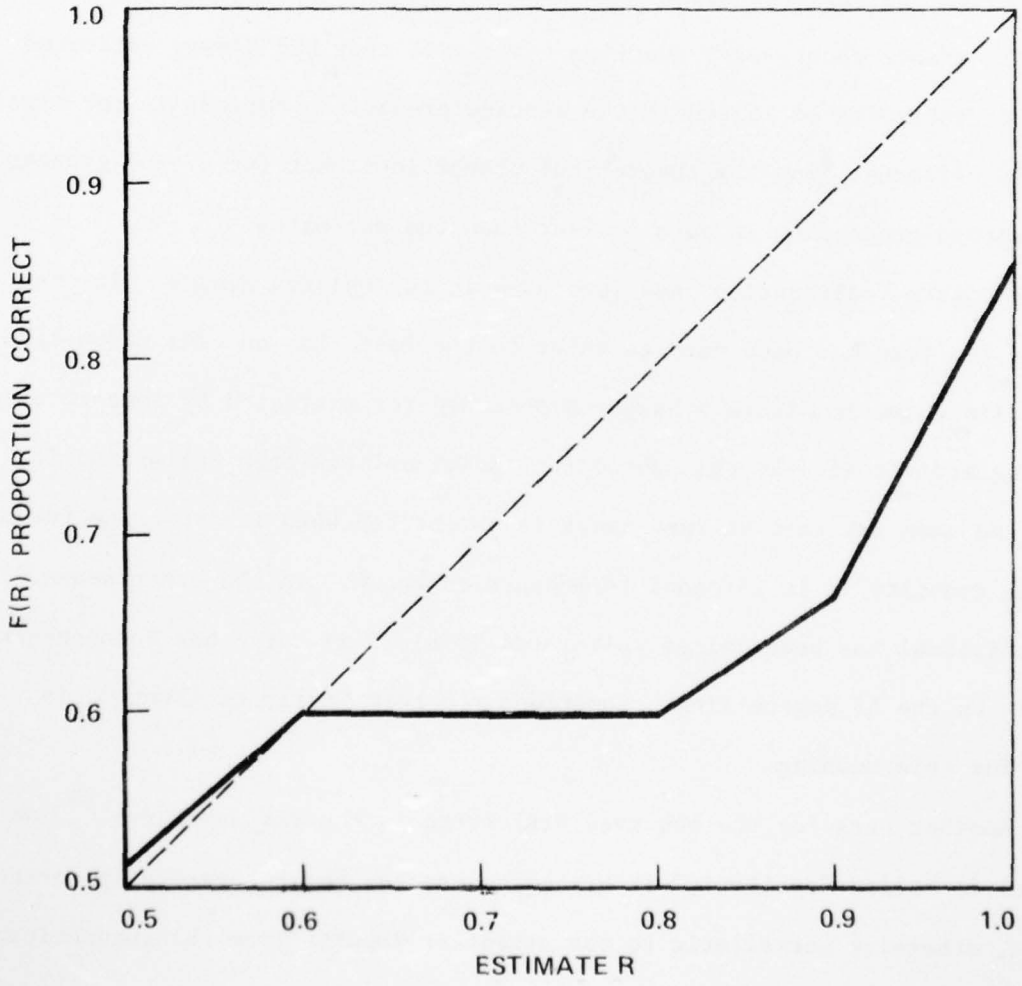


Figure 9. Calibration Curve (Data from Capen, *43 Subjects, 120 Questions*)

that had, for him, the higher subjective probability. The questions were a mixture of professionally relevant and general information types. The graph shows the average proportion of correct responses to total number of responses with a given probability.

Figure 9 is an average over the 43 subjects, and is somewhat smoother than what is observed for a single subject. Surprisingly erratic data can be observed with a single subject. Figure 10 is an example.

A quick glance at Figure 9 indicates that the engineers are not doing very well in making probability estimates. As we shall see later, the majority of the subjects would have done better if they had expressed complete ignorance about every question — i.e., if they had always estimated .5! For estimates of .5 and .6 the average proportion correct is not significantly different from the theoretical proportions; but for .7 and greater, the average proportion is much smaller than the estimates.

The term "calibration" has been used in two related senses. In one sense, the term has been used to refer to the fact that an F(R) curve like the solid curve in Figure 9 has been observed (or estimated by someone else) for the individual. In this sense the individual has been calibrated in much the same way that an instrument is calibrated when its response curve to the quantity it is intended to measure is known. In the other sense, an individual has been called calibrated if his F(R) curve has been observed to lie on the 45 degree line. Sometimes the term "fully calibrated" is used for this meaning.

Another term for the observed F(R) curve is the realism curve.[16] An individual is called "realistic" if his curve matches the theoretically correct curve, otherwise unrealistic to the extent it departs from the theoretical.

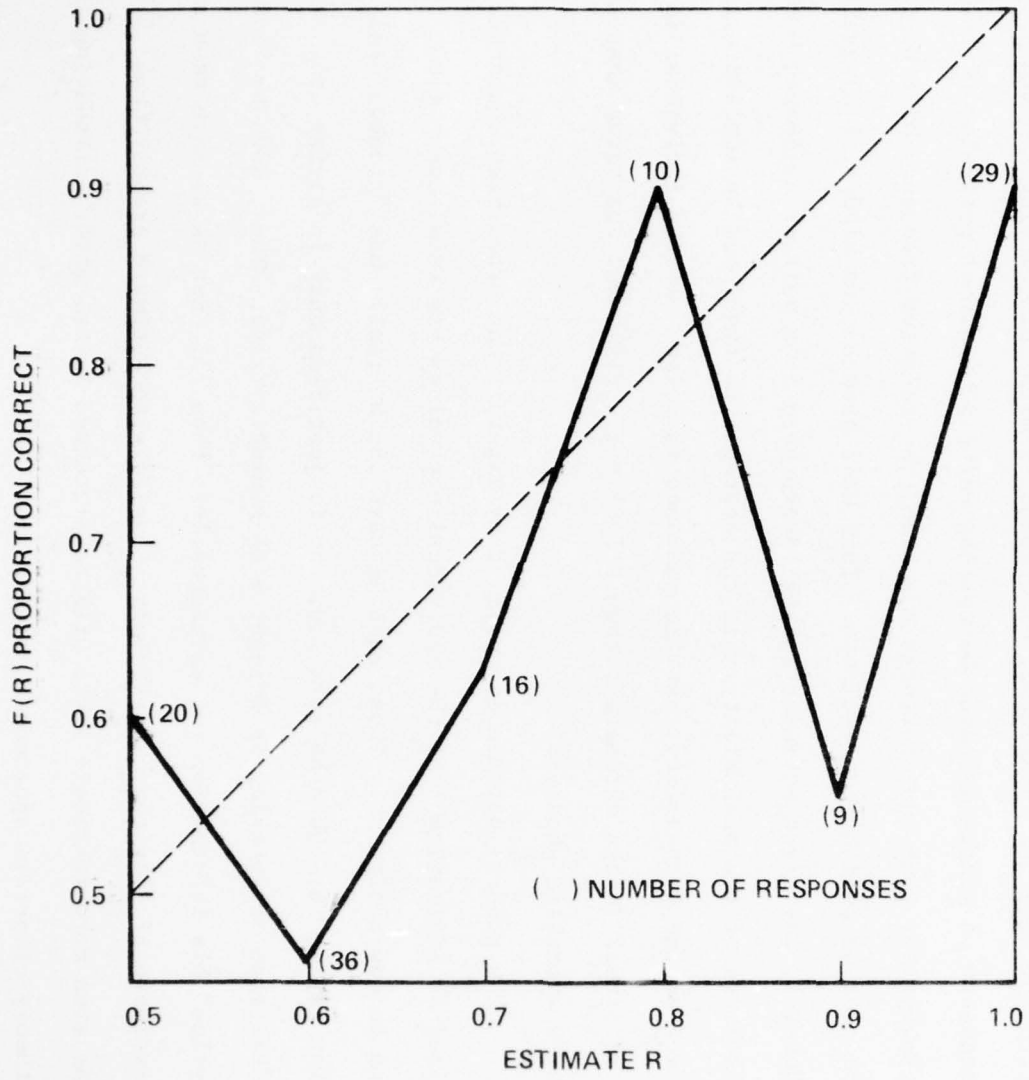


Figure 10. Individual Calibration (Data From Capen, Subject 29. 120 Questions)

Figure 9 is fairly typical of the results obtained by many investigators in this field.[17] In general, there is a tendency for subjects to fall below the 45 degree line for estimates exceeding .5, and to remain above it for estimates less than .5. Some difference in conventions of counting have led to apparent discrepancies with this finding. Given a sequence of sentences, and the corresponding sequence of estimates of the probability that the sentences are true, the relative frequency of true can be plotted against the estimate. This will produce one kind of F(R) graph. However, any given estimate can be interpreted as a pair of estimates — one estimate of the probability that the sentence is true, and an implied estimate of the probability that the sentence is false. If the individual is consistent in his estimates, then $F(1-R) = 1 - F(R)$, and the curve must be skew symmetric about .5.*

The general pattern exemplified by Figure 9 has often been characterized as indicating that the individual <u>overvalues</u> his knowledge — i.e., that he is overconfident. Thus, when he says .8, he really has "grounds" only for saying .6. When he says 1.0, he is justified only in saying .85, etc. This mode of speaking is fraught with semantic traps. Thus, does he "overvalue" his information for estimates less than .5? But as we have seen an estimate of less than .5 is always coupled with estimate greater than .5. An even more intricate snare will be discussed later, after introducing the theory of errors approach to estimation.

---

*This convention can lead to verbal puzzles at .5. In particular it implies that an individual is always completely realistic for the estimate .5. If an individual is presented with a set of sentences, all of which are true, and he responds, "The probability that this sentence is true is .5" to every sentence, he is right one-half the time.

The suggestion has been made that the realism curve be used as a method of generating objective probabilities from the individual's subjective reports. The F(R) curve, on this suggestion, can be used to correct the individual's estimates. At first glance, this looks like a fairly attractive idea. As mentioned above, the F(R) curve might be thought of as a kind of "theory of the instrument" of an individual making probability judgments. Various nonlinear relationships have been observed by psychologists between physical stimuli and perceived magnitudes. Offhand, there is no reason why there should not be a nonlinear scaling between objective probabilities and "perceived" probabilities. Presumably, such a relationship could be used to rescale the subjective probability estimates.

In order for this scheme to have any value, the observed F(R) curve must be a stable property or trait of the individual. That is, over a fairly wide variety of types of questions and circumstances, the observed relative frequencies must be roughly the same for the same reported probabilities. There is a fair amount of evidence that this is probably not the case; that the degree of realism is a function of the type of question being asked.*

A much more serious objection to using empirical F(R) curves for rescaling probability estimates is presented by the fact that probabilities are absolute scales, and allow no transformations. This statement appears to be significant enough to warrant being stated as a theorem.

Theorem 6. If P is a probability measure on the event space U, then there is no function $F(P) \neq P$, which is also a probability measure on U.

---

*Vide the discussion of "hard" questions in Chapter IV.

Proof: Let $\{E_j\}$ be an m-fold equipartition of U, i.e, $P(E_j) = P(E_k) = 1/m$ for all j and k. Let F be any function of P. We have $F(P(E_j)) = F(1/m)$. If F is a probability measure, since the partition is exhaustive, $\sum_j P(E_j) = 1 = mF(1/m)$. Whence, $F(1/m) = 1/m$. Consider any $n < m$ of the E_j. $P(\sum_j E_j) = n/m$. Since F(P) is a probability measure, and the E_j are exclusive, $F(P(\sum_j E_j)) = \sum_j F(P(E_j)) = n/m = F(n/m)$. Since an real number can be approximated by a sequence of rationals, if F is continuous, $F(x) = x$ for any real number x. The proof requires that U contain m-fold equipartitions for arbitrarily large m. This condition is assured by P7.

There are several ways Theorem 6 can be viewed with respect to calibration. Suppose P is the actual probability measure on U — i.e., P is the process which generates the occurrence or nonoccurrence of the events tabulated to give the relative frequency F(R), where R is the individual's subjective probability measure on U. If R differs from P, then there is no stable relationship between P and R. In other words, two sequences E_i and F_i can be selected out of U, each with the same estimated probability, and each with different probabilities of occurrence.

Another way to view the theorem is the following: If an individual is a consistent probability estimator, then no rescaling of his estimates is also consistent. On the other hand, if the individual is not consistent, then his estimates cannot be used with confidence, rescaled or not. To make the dilemma clear, suppose we are interested in P(E v F) where we know E and F are exclusive. There are two ways we can obtain this estimate; (1) Let the individual estimate P(E) and P(F), rescale these, and take the sum. (2) Let the individual estimate P(E v F) and rescale this estiamte. If the individual is not fully realistic, these two procedures will generally give two quite different numbers. Which is the best estimate?

If only one number is required — i.e., if there is no intention of using the individual's probability judgment in further computations — then possibly a rescaling procedure could be justified, where the individual estimates precisely the probability desired. However, in almost all interesting applications, various manipulations of estimates are needed to complete the analysis.*

In a way, this result is somewhat disappointing, since the notion of calibration appeared to be a way of tying probability estimates to reality. However, in another way the result is comforting. It says flatly that realism curves define objective probabilities if and only if the individual is fully realistic. The question of how to proceed if an individual is not "reasonably realistic" will be pursued in Chapter IV.

7. Theory of Errors Model

The theory of errors is perhaps the most widely used of the estimation models in experimental psychology. It is most often applied to simple magnitude estimates, but in theory applies to any quantifiable judgment. In elementary form the model assumes that an estimate has two components, a stable, non-variable, component, and a random error component. For estimates where a correct or true response is definable, it is usually assumed that the stable component is the true answer, and any given response of an individual is the sum of that true answer and a random perturbation, i.e.,

$$R = T + \epsilon \tag{6}$$

*In a previous publication[18] I was ambiguous concerning the notion of calibration as a foundation for a theory of group estimation. Theorem 6 pretty well clears up the ambiguity; calibration is an insubstantial foundation. The earlier formalism is still valid. The notation $P(E|R)$ must be interpreted as the probability that the event E will occur, given that the individual asserts R, and cannot be interpreted as $F(R)$ derived from some calibration curve.

The source of the random perturbation ϵ is usually not identified; it is assumed that variable factors both in the immediate environment and in the internal estimation process lead to variability in the individual's response. The theory consists primarily in characterizing the properties of this variation. It is usually assumed that the random error component has a mean of zero, and its value on any given response is independent of its value on any other response.

In addition, it is often assumed that the random error is normally distributed, i.e., it has the density function $\phi(\epsilon) = 1/\sqrt{2\pi}\sigma\, e^{-\epsilon^2/2\sigma^2}$ where σ is the standard deviation of the error. Thus the total response R is normally distributed with mean equal to T. To this extent, the theory is quite analogous to the theory of a fallible instrument in the physical sciences.

For certain kinds of estimates such as those involved in psychophysical measurements, it appears feasible to replicate the estimates so that direct observational verification of the assumptions can be made; the shape of the distribution can be determined, and parameters like the mean and standard deviation can be computed. However, for the kind of estimate we are interested in, direct observation of random errors is difficult. The individual is likely to remember his previous answer, and thus the basic assumption of independence on replication does not hold. Statements concerning random error have to be made indirectly, based on the consequences of assumptions about the form of the random variability. For this reason, some investigators prefer terms like "residual variability" or "unexplained variation."

To make the theory applicable to the kind of data that is obtained in experiments with uncertain questions, an addition feature is needed, namely a bias component. Thus, an individual response is considered to be the sum of three factors

$$R = T + B + \epsilon \tag{7}$$

The bias term, B, like T is a stable component — i.e., it is constant for a given individual and a given question. Equation (7) is equivalent to the assumption that the individual "selects" his response out of a distribution that is centered around some mean that is displaced by the bias from the true response, as illustrated in Figure 11.

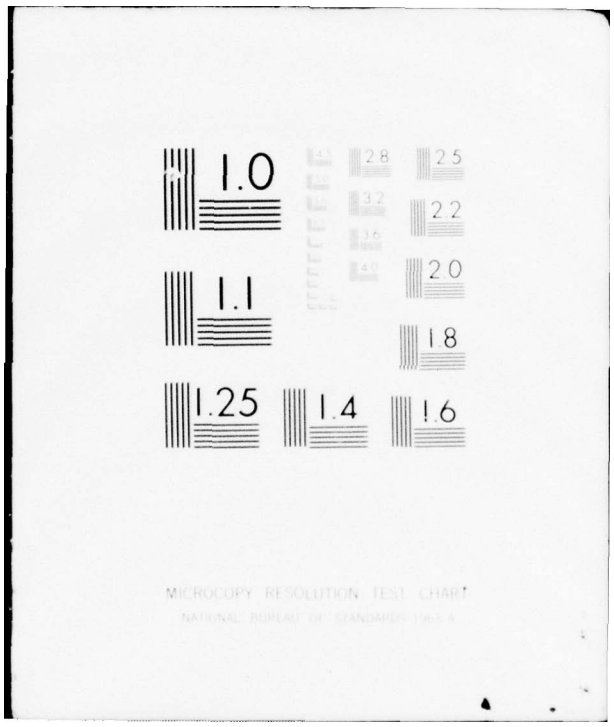
The notion of bias has not received as much attention in psychological literature as the notion of random error (they are often lumped together as "error"); a simple illustration from a physical situation may make the idea clearer. Suppose there is a marksman firing at a target who has not compensated adequately for windage or distance. His pattern of shots might look like the dots in Figure 12, which are clustered about a point displaced from the center of the target. The displacement illustrated by the solid line in the figure is the bias of the pattern; the offset from the center of the pattern, illustrated by the dashed line, is the random error of the specific shot labelled R.* It should be clear from this illustration that the notions of bias and random error are idealizations — the "biassing influences" such as wind and adjustment for distance are assumed to be constant throughout the trial.

---

*Figure 12 can be used to illustrate a common illusion in practical situations requiring judgment. Consider an individual who has a large bias, and an equally large random error. These two can compensate, and his answer be precisely correct. If this occurs in a significant decision context, the individual can be credited with remarkable acumen.

AD-A042 852 CALIFORNIA UNIV LOS ANGELES SCHOOL OF ENGINEERING A--ETC F/G 5/10
GROUP DECISION THEORY.(U)
JUL 77 N C DALKEY
UNCLASSIFIED UCLA-ENG-7749 N00014-69-A-0200-4056
NL

2 OF 4
AD A042852


1.0
1.1
1.25
1.4
1.6
2.8
2.5
3.2
2.2
3.6
2.0
4.0
1.8
MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS 1963-A

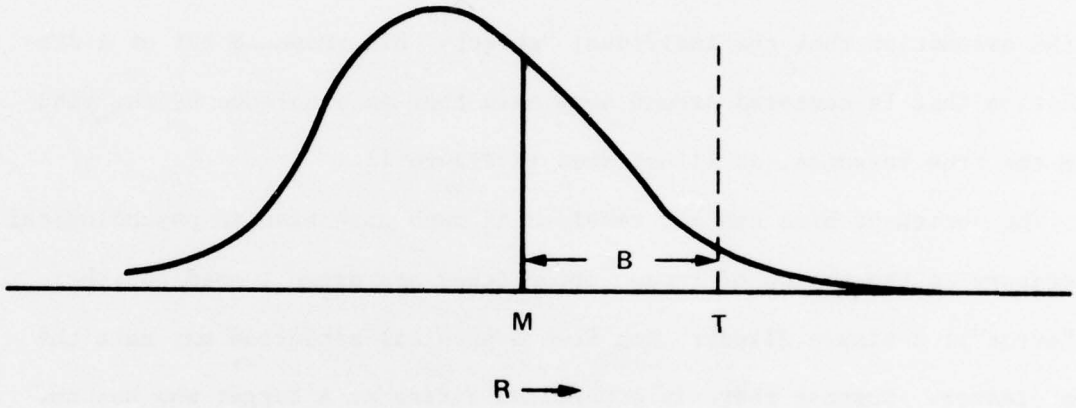


Figure 11. Illustrative Distribution of Response with Random Error and Bias

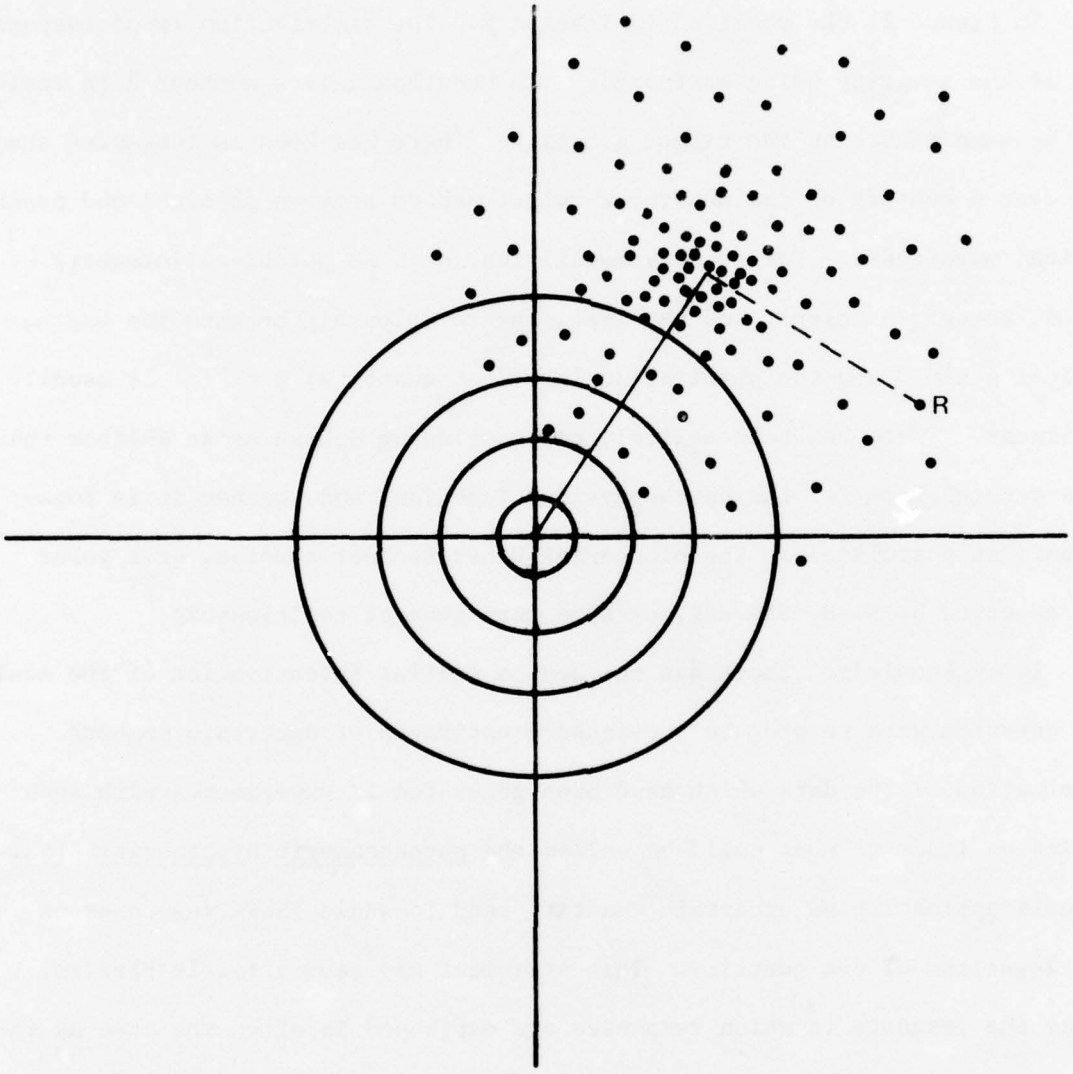


Figure 12. Illustration of Bias and Random Error

If the bias is unknown, the process appears to be a random selection of a response R out of a distribution with mean M where M = T + B. In Figure 11, B is negative.

In Figure 11 the abscissa is labeled R. The distribution is of responses, not of the quantity being estimated. The question arises whether R is scaled in the same manner as the target quantity. There has been an intensive study for over a century of the nonlinear relationships between physical and psychological magnitudes. For sensory modalities, such as perceived intensity of sound, perceived weight, and the like, the relationship between the psychological scale ψ and the physical scale q (for quantity) $\psi = f(q)$ is usually nonlinear. There has been a lively and continuing debate as to whether there is a general form for the psychophysical function, and whether it is logarithmic as postulated in the pioneering Weber-Fechner studies, or a power law as urged by S. S. Stevens, or some more general realtionship.

To my knowledge, there has not been a similar investigation of the scaling question with respect to non-sensory estimates of uncertain numbers. Examination of the data which have been generated in experiments with such estimates leads to what could be called the psychonumeric hypothesis: Individuals estimating an uncertain quantity tend to scale their responses on the logarithm of the quantity. This statement may seem a little bizarre, since the language in which responses are expressed is often the same as the language used to describe the physical quantity. In what sense can we say that 100 "seems" twice as large as 10 rather than, as arithmetic requires, ten times as large?

Rather than trying to resolve the semantic puzzles generated by this kind of talk, it is probably less mystifying to look at some of the data

which suggests the hypothesis. Figure 13 shows the distribution of several thousand responses of student subjects to numerical questions like "How many telephones are there in Africa?"[19] The size of the numbers being estimated covered a wide range — from "How many appointments have been made to the U.S. Supreme Court since 1930 (as of 1969), answer 20, to "How many gallons of beer were produced in the U.S. in 1964," answer 3 billion 193 million.

To make the responses comparable, they were transformed in the following fashion: z scores were computed for the logarithms of the responses, where the z score is computed as

$$z = (\log R - m)/s$$

m is the mean of the log responses (on a given question) and s is the standard deviation of the log responses (again for the same question), i.e., $s = \frac{1}{n} \sum_i (\log R_i - m)^2$, $m = \frac{1}{n} \sum_i \log R_i$. Figure 13 displays the distribution of e^z.

The smooth curve in Figure 13 is the log normal density function, i.e., it is the distribution that would be expected if the logarithms of the responses were normally distributed. As can be seen from the figure, the log normal distribution is a very good approximation to the data.

The skewness of the distribution of responses is to be expected on the grounds that all of the responses have a natural lower bound — all the questions involved answers greater than zero — but no natural upper bound. And the precise shape of the distribution can be explained by other assumptions than the psychonumeric hypothesis. However there is other evidence.

Figure 14 displays the standard deviations of the log responses graphed against the logarithm of the true answer. The true answer expresses the "size" of the number being estimated. The standard deviation has the property

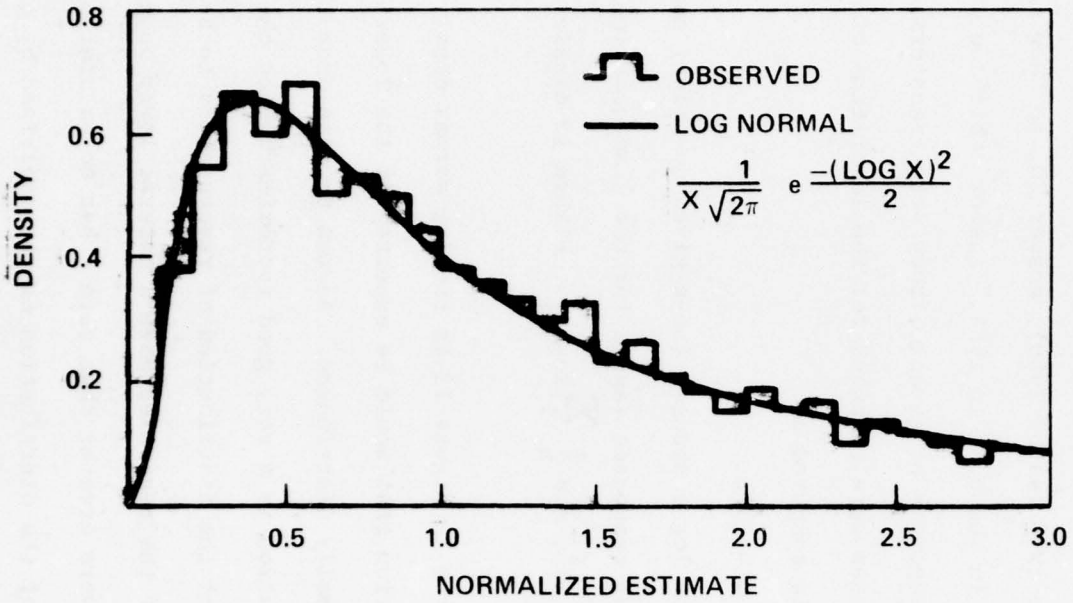


Figure 13. Distribution of Initial Answers

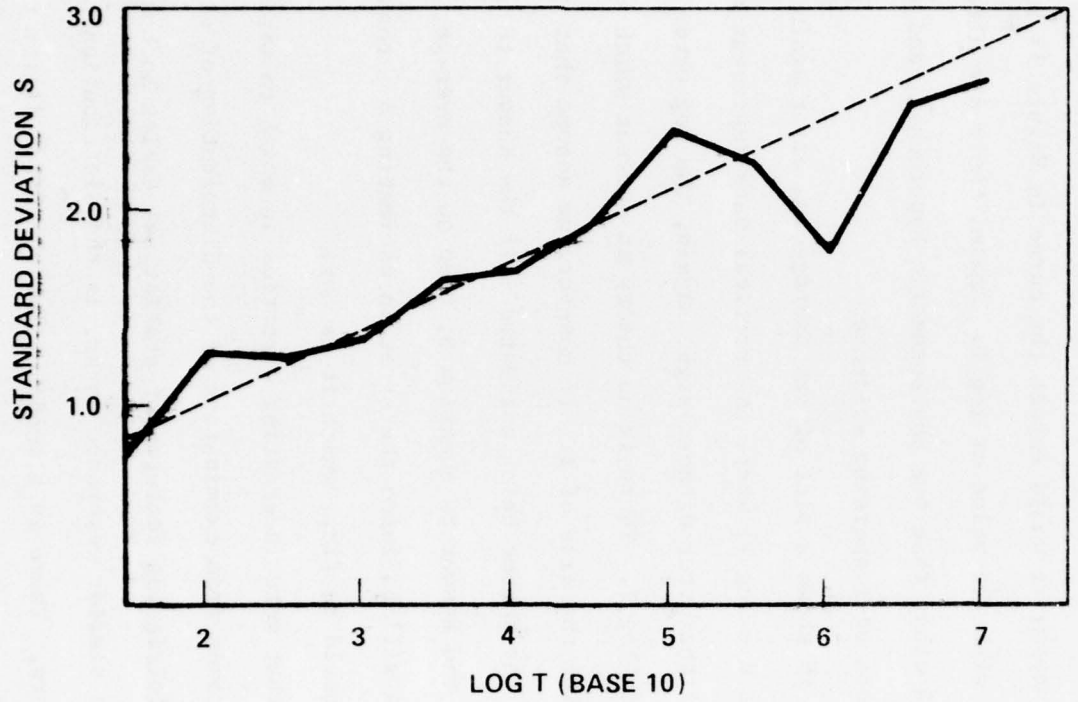


Figure 14. Average Standard Deviation as a Function of Log True

that it is invariant under a translation, i.e., $s(x + a) = s(x)$, where a is any constant. Similarly, the standard deviation of the logarithm of a set of responses is invariant under multiplication by a constant, i.e., $s(\log a\ x) = s(\log x)$. If the subjects were scaling their responses on the physical numbers, we would expect the curve in Figure 14 to be flat, rather than rising with the value of log T. Again, there are alternate explanations of Figure 14 other than the psychonumeric hypothesis, and it can be considered only one piece of supporting evidence.

Figure 15 shows a plot of the average log error against log T. log error = $|\log R - \log T|$ where the vertical bars represent absolute value, i.e., taking the error without sign. Again, the log error increases roughly linearly with log T. There is no theory at present which ties the size of the error with the size of T. If however, we assume that error scales with the size of the number being estimated — if the answer to question A is twice as large as the answer to question B, then on the average, the error in estimating A will be twice the error in estimating B — then the curve in Figure 15 should be flat, which it is not.

A somewhat more interesting direction in which to explore the question of scaling comes from examination of the distribution of digits in responses. One of the intriguing features of statistical tables such as are found in almanacs and similar reference works, is the distribution of the first digits of the numbers. There is a tendency for the first digits to be distributed in a logarithmic pattern; specifically the frequency of digit d ($d = 1,\ldots,9$) is roughly proportional to $\log(d + 1) - \log d$.[20] A somewhat more general hypothesis would be that the tabulated numbers x are themselves distributed as $1/x$. This would imply that not only the first digits but the second and

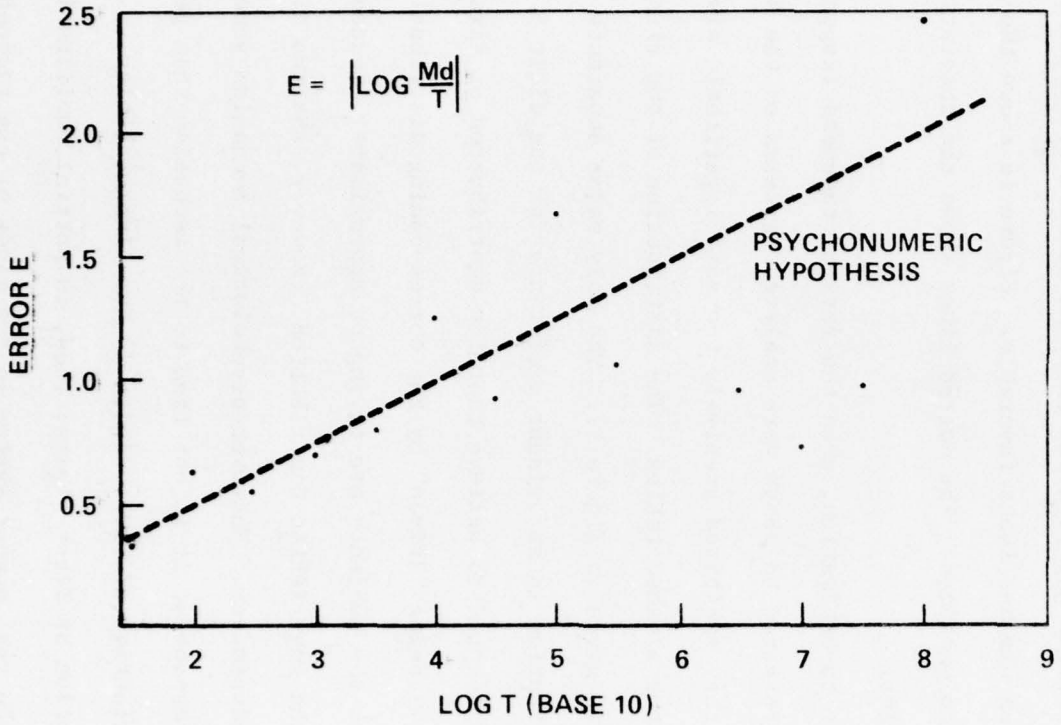


Figure 15. Average Error as Function of Log True

subsequent digits would also have the appropriate distribution — e.g., the frequency of d as a second digit would be proportional to $\sum_{i=1}^{10} [\log(10i + d + 1) - \log(10i + d)]$. I haven't had the opportunity to verify this hypothesis in detail — a relatively large body of data would be needed to generate stable statistics — but a quick try of a few thousand numbers selected more or less at random out of an almanac looks favorable. Figure 16 shows the distribution of second digits so obtained. The dotted lines shows the theoretically expected frequencies.

More relevant to estimation, when the several thousand responses to estimation questions referred to above were analyzed in terms of the distribution of first digits, they exhibited precisely the same logarithmic distribution as the data from the almanac tables. The distribution of the first digits of the responses is given in Figure 17. The only major departure from the theoretical distribution is an evident preference for the digit 5.

One might be tempted to believe that the distribution of first digits in the responses is being "driven" by the corresponding distribution in the true answers which the subjects are trying to approximate. Indeed, the true answers exhibit the logarithmic distribution. However, the two distributions are completely independent. Whatever psychological mechanism generates the distribution of responses, it is not tied to the mechanism that generates a logarithmic distribution of first digits in the almanac tables.

The distribution of first digits, then, is partial confirmation of the hypothesis that the real number system in the minds of the respondents is distributed like 1/x.

Some additional supporting evidence for the psychonumeric hypothesis will be discussed in the section dealing with group judgment and the theory

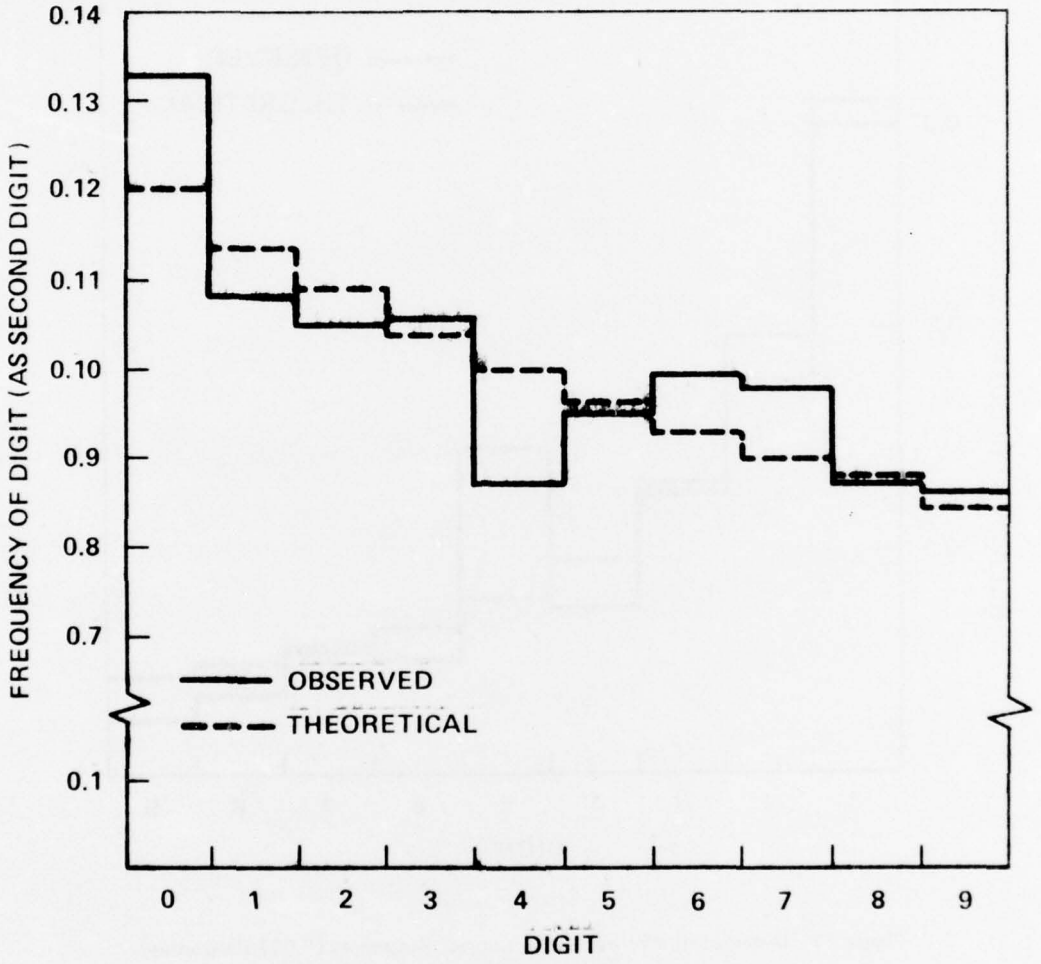


Figure 16. **Relative Frequency of Digits Occurring as Second Digits In Almanac Tables (3114 Numbers)**

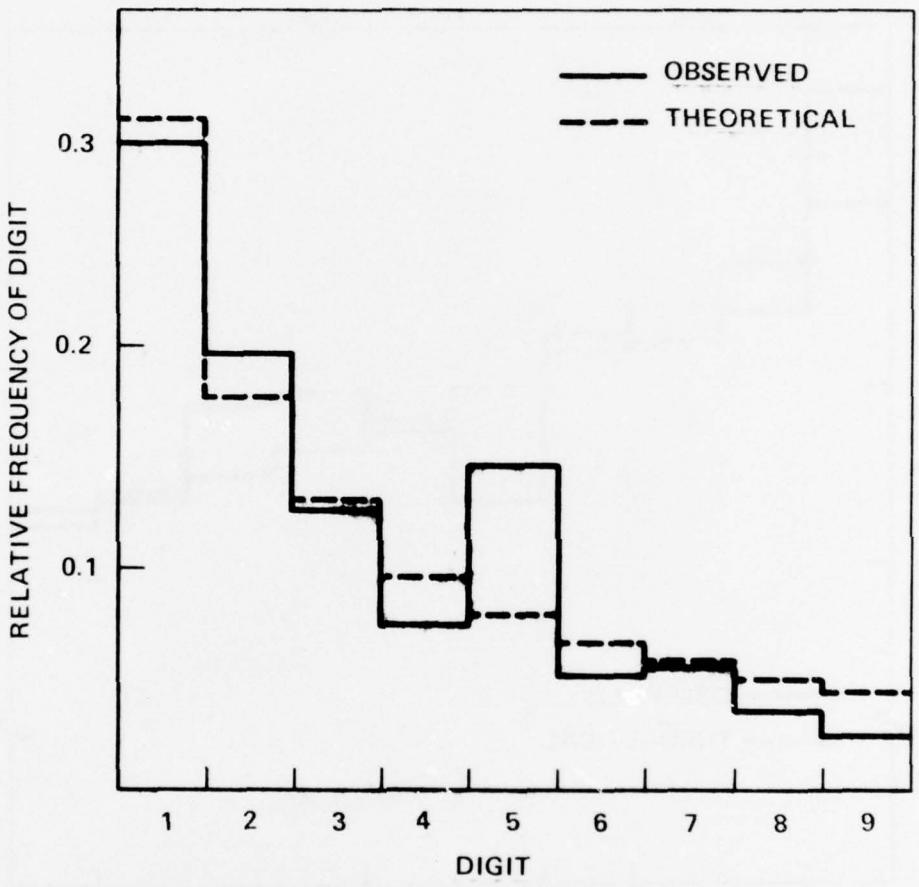


Figure 17. Distribution of First Digits, Subject Responses (5,037 Responses)

of errors, Chapter V. But to sum up what we have: If we assume that the individual scales his responses on the logarithm of the number he is trying to estimate, then we have a fairly straightforward explanation for the log normal distributions of responses, for the increase in standard deviations and errors with size, and for the logarithmic distribution of first digits in the responses. There does not appear to be a natural way to test the hypothesis directly. The direct question "how much bigger than 1 million is one billion" has too facile a response from arithmetic.

The psychonumeric hypothesis can be interpreted as asserting that the individual "thinks" in terms of the log transform scale, and thus needs, so to speak, a double translation process, first expressing the response space for the question in terms of the log scale, performing his estimate in this response space, and then retranslating the response into "ordinary numbers." Of course, if this were done literally, the power of computation available to the individual would have to be somewhat beyond the capabilities observed in lower division mathematics courses. The problem of representing such "internal scales" has been treated by Anderson.[21]

For our purposes, it is useful to express the psychonumeric hypotheses explicitly. This can be done by considering the logarithmic form of Equation (7). If we let lower case letters stand for the log transform of the corresponding upper-case letters, i.e., $r = \log R$, $t = \log T$, $\varepsilon = \log \epsilon$, etc., then we have

$$r = t + b + \varepsilon \tag{8}$$

If we assume ε is normally distributed, then the density function for r is

$$D(r) = 1 \Big/ \sqrt{2\pi}\sigma \, e^{-\frac{(r-t-b)^2}{2\sigma^2}} \tag{9}$$

where σ is the standard deviation of the error component. Equation (9) asserts that the distribution D(R) of the overt response of the subject is

$$D(R) = \frac{1}{R} D(r) \tag{10}$$

Equation (9) looks a good deal more intricate than it is; it asserts that the distribution of individual responses, if we could observe it, would look like the distribution in Figure 13.

The distribution defined by Equation (9) is not the same as the individual's subjective probability distribution on the quantity being estimated. The subjective theory of probability expounded in the previous section has the consequence that the individual has a subjective probability distribution on any quantity (at least any which he contemplates). It seems likely that there is some relationship between the response distribution and this subjective probability distribution, but neither the theory of errors nor subjective probability theory provides a formal link. If an experimental procedure could be devised which identified the response distribution sharply, it would doubtless be highly informative to make a comparison between the two types of distribution. One would expect a rough correlation between the standard deviations of the two, and it seems likely that the means would correspond closely to each other.

An instructive investigation arises from forming a hybrid between subjective probability theory and the theory of errors. In the section on calibration, it was pointed out that most individuals are poorly calibrated. This is usually interpreted as indicating that the individual <u>overestimates</u> his information. This interpretation is strengthened by testing the individual's propensity to bet on his estimates. Slovic reports that most of

his subjects were willing to bet on their estimates, even those with extreme odds.[22] But the situation may be more complicated. Suppose we assume that the individual's probability estimates are subject to the same sort of random error as magnitude estimates, so that formula (6) applies. It turns out that this assumption leads to calibration curves very much like Figure 8.

The assumption that the random error is normally distributed is inappropriate for the restricted interval of probabilities. Other types of error function can be postulated; in the analysis below I investigate two possibilities, a beta function and the negative exponential.

Using formula (6) rather than formula (7) implies that the mean of the individual's probability judgments is the "true" probability. There are many investigators in the field of probability estimates who deny the existence of "true" or objective probabilities for the kind of question where calibration is of interest. I'll bypass that sticky point by making a less troublesome assumption; namely, the assumption that if the individual always reported the mean of his response distribution, then he would be fully calibrated. Although this appears to be a strongly "favorable" assumption, it will turn out that the subject will be poorly calibrated when the random error is included.

Case 1: Beta distribution. Consider an individual who selects a response R out of a distribution of the form

$$D(R) = aR^{T}(1 - R) \tag{11}$$

where T is determined by the mean, $\bar{R}$. In this case

$$T(\bar{R}) = \frac{3\bar{R} - 1}{1 - \bar{R}}$$

a is a normalizing constant, which is also a function of the mean

$$a(\bar{R}) = (T(\bar{R}) + 1)\ (T(\bar{R}) + 2)$$

Equation (11) is the form appropriate for $\bar{R} \geq .5$; for $\bar{R} \leq .5$ the appropriate form is

$$D(R) = aR(1 - R)^T \tag{12}$$

An example of Equation (11) is given in Figure 18 where $\bar{R} = .35$.

If we assume that the individual is posed a large set of questions for which the objective probabilities are uniformly distributed between 0 and 1, we can calculate the resulting calibration curve. The results are presented in Figure 19.

Case 2: Negative exponential distribution. If we assume that the shape of the distribution is affected by the degree of certainty of the individual in his estimate, then the distribution should become "flatter" as the degree of uncertainty increases. A rather extreme case is afforded by the negative exponential

$$D(R) = ae^{-bR} \tag{13}$$

where both a and b are functions of the mean $\bar{R}$. An example of Equation (13) for the case $\bar{R} = .25$ is given in Figure 20.

The negative exponential is the maximum entropy distribution where only the mean is known. Thus it is the "flattest possible" curve given the assumption of realism for the mean. The calibration curve resulting from this distribution is displayed in Figure 21.

The resemblance of the two calibration curves to those observed empirically is rather striking. As might be expected, the maximum entropy distribution leads to poorer calibration than the beta distribution. Nevertheless, it should be pointed out that the resulting calibration curves for either the beta distribution or the negative exponential are still "better" than the

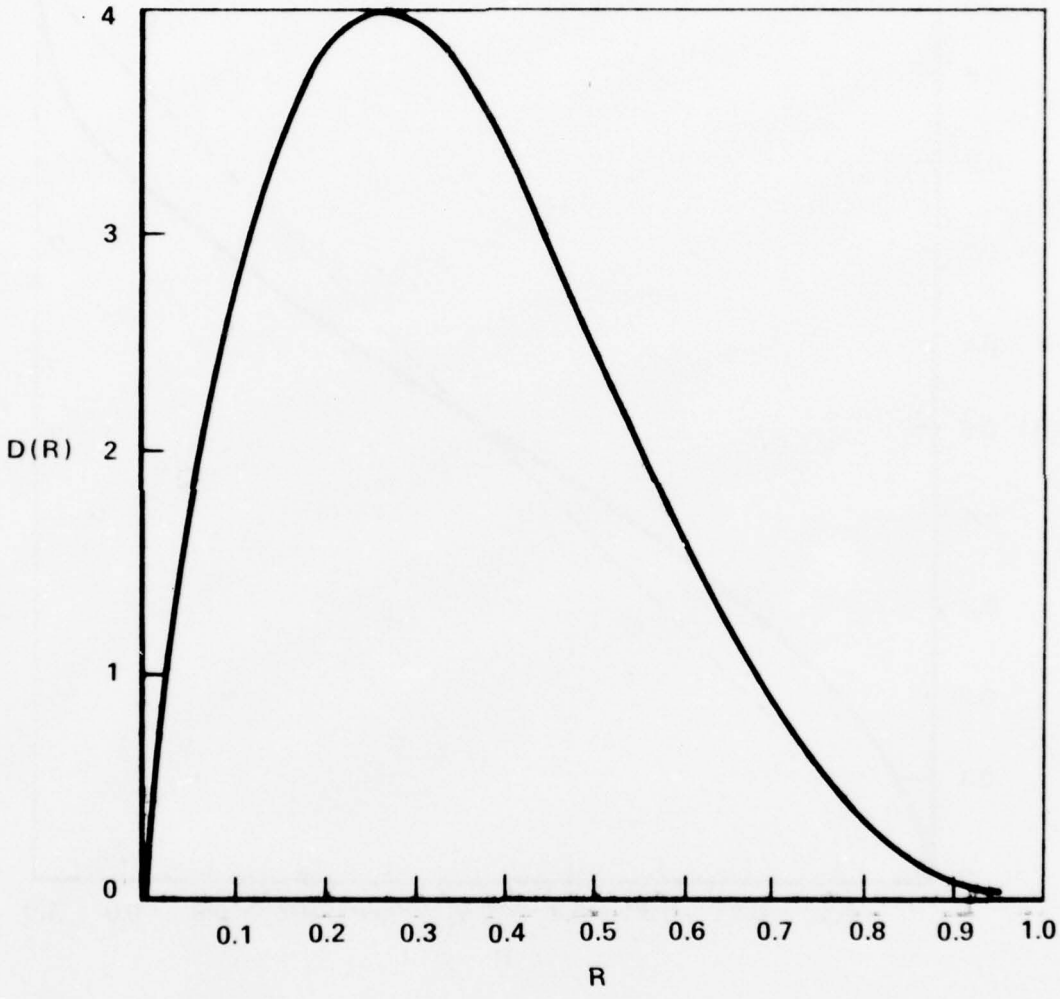


Figure 18. Beta Reliability Distribution $D(R) = R(1-R)^{2.7143}$, $\bar{R} = 0.35$

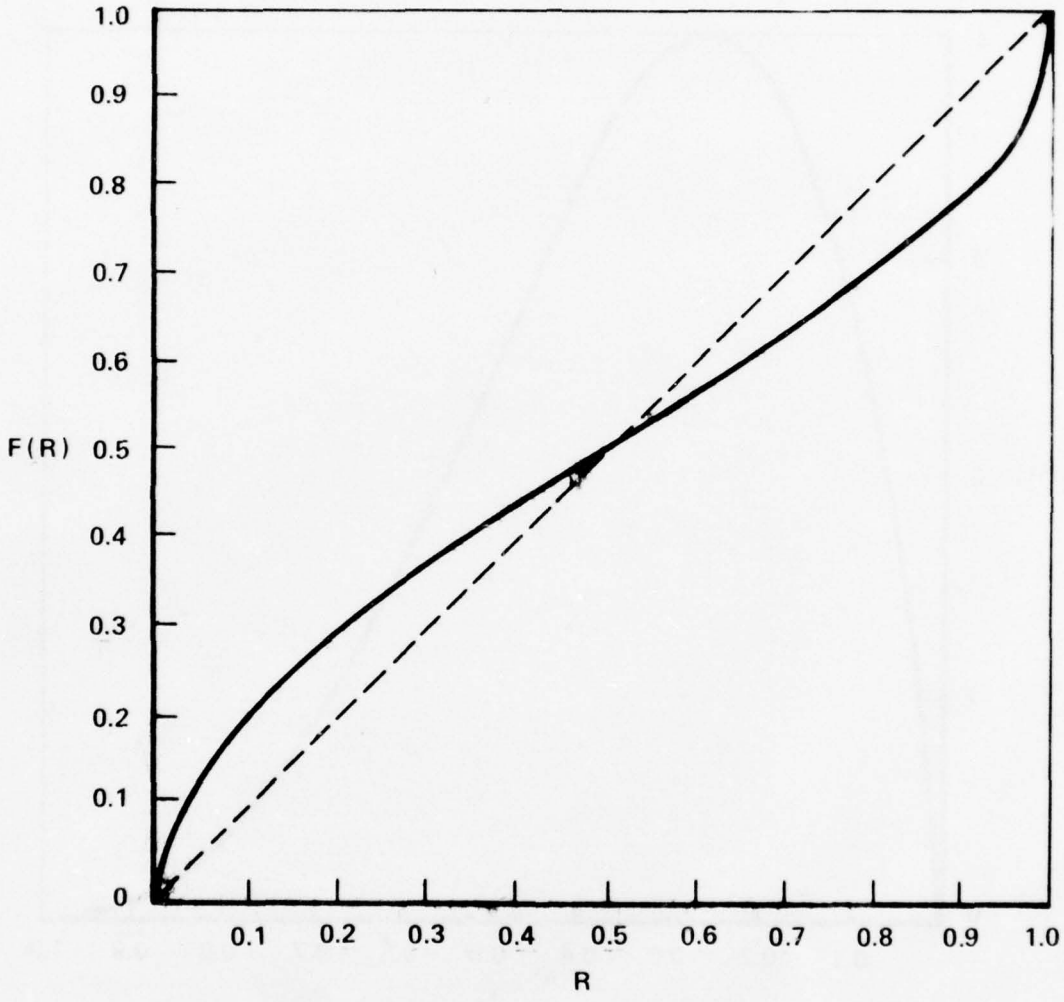


Figure 19. Computed Calibration Curve
Beta Reliability Distribution

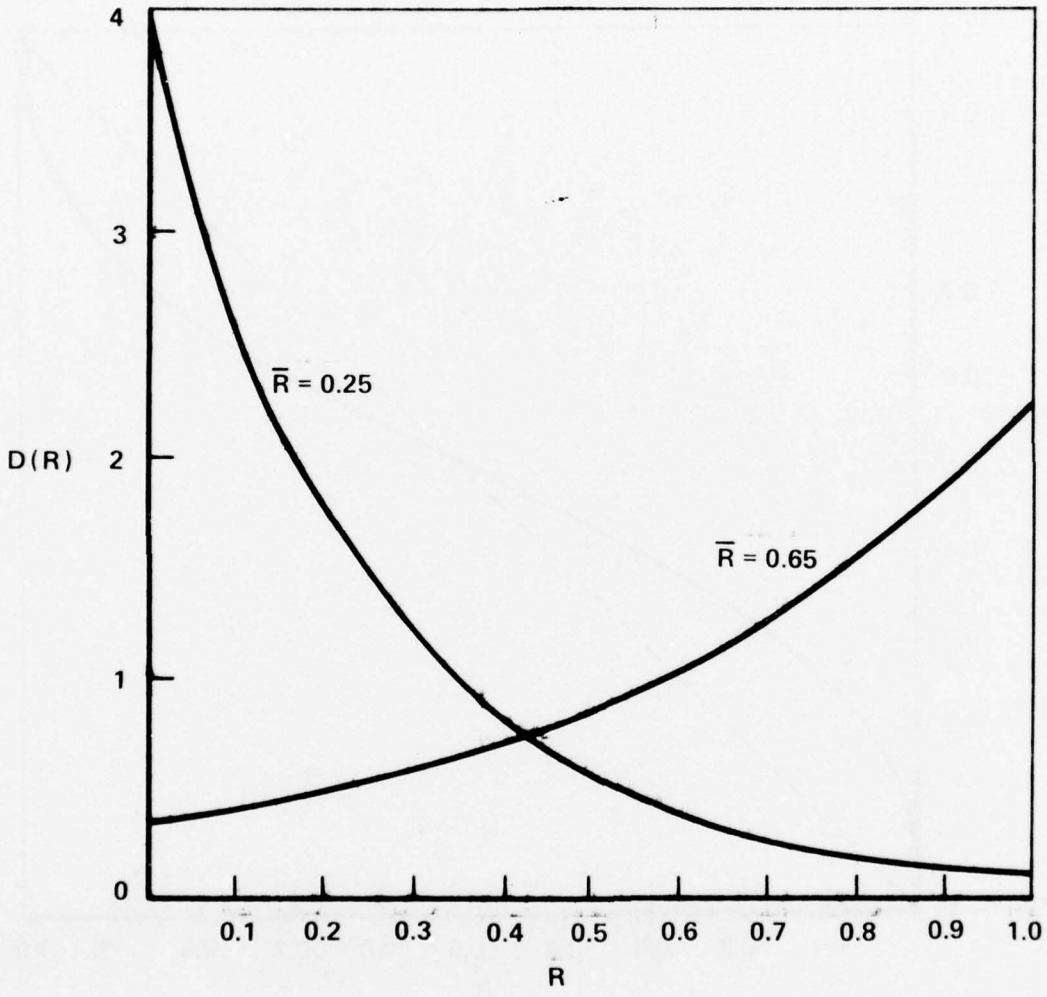


Figure 20. Maximum Entropy Reliability Distribution

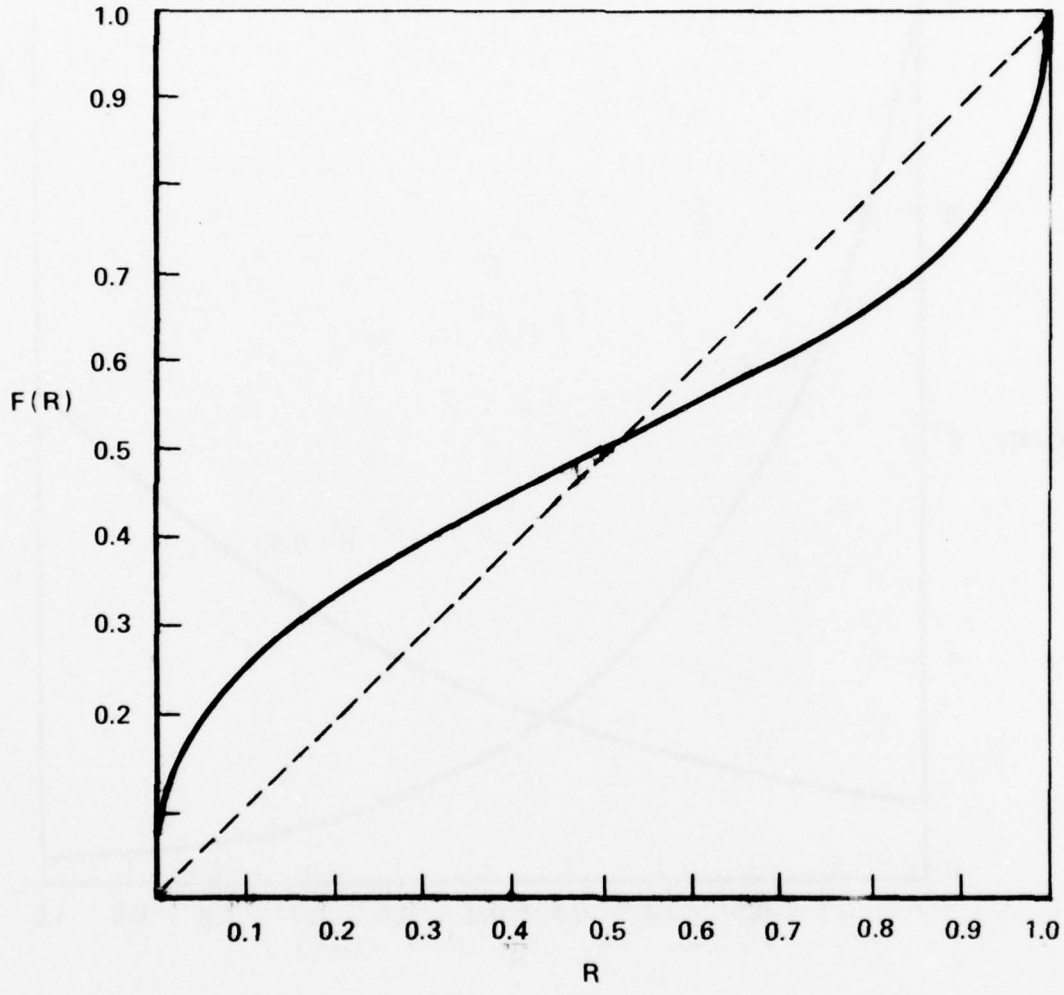


Figure 21. Computed Calibration Curve
Negative Exponential Reliability Distribution

empirically observed curve Figure 9. It seems likely that bias is playing a role in Figure 9 as well as random error.

Figures 19, 21 are especially interesting in that they are derived from an assumption which in a sense is the reverse of the usual interpretation. The assumption is that the individual is highly uncertain and this leads to variability in his responses. Rather than being overconfident, he is (behaviorally) just the reverse.

There is a peculiar nursery rhyme quality about this analysis; the situation is much like Jack Horner, thrusting his thumb at random into the response pie, pulling out a response, and then saying "What a good boy am I — that's just right!"

The analysis presented here seems to undercut further the notion of calibration as a feasible procedure for correcting an individual's probability judgments. For any given judgment, the "appropriate" correction would not be the empirically established F(R), but rather the unknown mean, m, of the individual's response distribution.

CHAPTER III. FIGURES OF MERIT

1. Types of Scores

It was stressed in the Introduction that an essential aspect of the resolution of disagreement is the formulation of figures of merit or scoring rules. Some measure of performance is needed to apply the Emerson principle — to show that one resolution procedure generates a judgment of greater excellence than another. The present chapter is concerned with figures of merit for factual statements. Value judgments will be taken up in Chapter VI.

There is a bewildering variety of scoring procedures to be found in the literature. There does not appear to be a general theory encompassing all of these and their obvious extensions. One reason is that scores can be handmaidens to a variety of kinds of excellence. One large and not very well organized class of scores has evolved around the function of specifying scientific excellence--i.e., furnishing criteria for the decision "statement x is scientifically acceptable." Another large body of procedures has grown up around psychological measurement — using test scores to attach descriptive indices such as I.Q.'s to individuals. Scores can be used as motivating devices as with school grades, or National Football League standings. And scores can perform multiple roles, such as being constructs for model building. The Gross National Product is a figure of merit for the economy, but it is also a basic notion in macro-economic theories.

Another complexity is the fact that scores themselves can be subjected to evaluation. A given scoring procedure may or may not be a good measure, depending on the role it is intended to play. A case in point is the theory of scoring as applied to psychological measurement. There is (from the point

of view of the present effort) an interesting ambiguity that pervades most of this subject. The usual function of a psychological test is to attach a number to an individual, denoting an aptitude, a trait, or perhaps an attitude. The function of attaching a figure of merit to the responses is secondary. In intelligence testing, for example, the test constructor presumably knows the answers to the questions. For some tests, like the intelligence test, attaching an independent objective score to the items is an intermediate step to the basic intent. For other tests, such as personality "inventories" or attitude scales, there is no objectively correct answer to the items. Even for those tests where there is an objectively correct answer it is becoming clear that utilizing data for responses which are not correct may be more diagnostic than simply counting correct answers.[1]

Very generally speaking, given a response R (to a factual question) and given the true response T, a score measures the discrepancy between R and T; i.e., there is a function S(R,T) which expresses the degree to which R approximates T on some criterion. S will thus depend on the form of R and T, as well as the role of R in a decision process. The following list of frequently used scores is not intended to be exhaustive; however, the list is representative of the range of scoring methods that have played a role in decision analysis.

Types of Scores

1. Binary scores
2. Distance scores
3. Scaled distance scores
4. Correlations
5. Probability scores
6. Decisional scores

Binary Scores. The simplest kind of binary score is a direct comparison between R and T, e.g., 1 (or "true") for the case $R = T$, and 0 (or "false") for $R \neq T$. This is the scoring scheme used in traditional two-valued logic. 1 is, of course, excellent, and 0 is bad. Most of the rules of deductive logic are concerned with transformations which preserve the score 1. Straightforward extensions of this type of score are obtained by "acceptable level of approximation" forms, e.g., $S(R,T) = 1$ if $|R - T| \leq c$, otherwise $S(R,T) = 0$, where c is a constant defining a region around T that is "close enough."

The true-false score is relatively unambiguous when applied to singular estimates like "The diameter of the moon is 2160 miles." Difficulties arise when the score is applied to general statements such as "Human beings have a life span of less than 130 years," where no exhaustive list of cases can be displayed. An implicit requirement for most scoring rules is finite applicability. As a result, scoring rules are usually restricted to singular estimates, or at most a finite set of singular estimates. This restriction is followed in the present book. It clearly leaves unexamined a significant realm of estimates which guide decisions, namely, judgmental generalizations. In part this gap is narrowed by considering correlations as types of scores, and by probabilistic scores which have a hazy connection with generalization. But neither confront head on the issue of attaching figures of merit to general statements.

Distance Scores. A more general kind of score can be defined if R and T are elements in a metric space. A metric space is defined as a set of elements where the distance D(x,y) between each pair of elements x,y is defined, and D(x,y) fulfills the conditions:

D1. $D(x,y) \geq 0$, and $D(x,y) = 0$ if and only if $x = y$.

D2. $D(x,y) = D(y,x)$

D3. $D(x,y) + D(y,z) \geq D(x,z)$

For ordinary magnitudes, $D(x,y) = |x - y|$. For example, the distance between 5° Fahrenheit and 30° Fahrenheit is 25°. However, the notion of a metric space is very general and applies to a wide variety of types of quantities and types of distance measures.* Given a metric space, we can define $S(R,T) = -D(R,T)$, the negative sign to indicate that smaller distances are more excellent. $D(R,T)$ is the common definition of error.

For many purposes, the distance squared is a more convenient measure. For single quantities, the distance involves the absolute magnitude, which is difficult to manipulate; the squared distance is more tractable. For multidimensional quantities, the distance squared has a number of additional convenient features. In n-dimensional euclidean space, the distance is defined as

$$D(x,y) = \left(\sum_i (x_i - y_i)^2 \right)^{1/2}$$

where x_i and y_i are the components of x and y on dimension i. The square root is awkward, but in addition, $D(x,y)^2$ decomposes directly into the sum of the squared differences on each coordinate — a very useful property.

The distance squared is associated with a number of indices which are of basic utility in statistics, such as the variance and least squares approximations. A useful relationship is the following:

$$1/n \sum_i D(R_i,T)^2 = 1/n \sum_i D(R_i,\bar{R})^2 + D(\bar{R},T)^2 \qquad (1)$$

$$= \mathrm{Var}(R) + D(\bar{R},T)^2 \qquad (1a)$$

*An elementary treatment, along with many considerations relevant to the present chapter and Chapter V, is presented in Kemmeny and Snell.[2]

This relationship holds for a set of responses $\{R_i\}$ where $\bar{R}$ is the mean, and Var(R) is the variance of the responses. The formula states that the average squared distance of a set of responses to the true is just the variance of the responses, plus the squared distance of the mean to the true. In the language of error, the average of the squared error of the individual responses is just the variance of the individual responses plus the squared error of the mean.*

Scaled Distances. It is often not the absolute size of an error that is relevant, but the comparative size. This becomes especially important when accuracy on different questions is being compared, or when an average score over many questions is computed. For example, if the question is "How many gallons of beer were consumed by the American public in 1970?" if the answer is off by one million gallons, that is only an error of one-tenth of one percent. But if the question is, "How many deaths due to accidents occurred in the United States?" and the answer is off by one million, that is an error of 2000 percent. Clearly, the latter error is "much worse." To compare accuracy on these two questions, it seems natural to normalize the responses to the size of the true answer, i.e., to define the error as $|R - T|/T$. This particular normalization works only for ratio scales, where

*Actually, T, and the R_i, can be any real numbers as far as (1) is concerned; the formula does not depend upon the fact that T is the true answer to a question for which the R_i are responses. A similar remark holds for many of the formulae in this book, where the notation is more restrictive than the content of the statements.

there is a well-defined zero. For example, if a temperature is being estimated, and T = 32° Fahrenheit, and the estimate is 41°, the scaled error is 9/32 = 6%. However, if the problem is transformed to the Celsius scale, the scaled error is 5/0 = ∞. For other purposes, $D(R,T)^2/T$ may be more appropriate.

More intricate forms of rescaling are often useful. The psychonumeric hypothesis, for example, suggests that a logarithmic transformation of R and T may be more in accord with perceived size than R and T unscaled. The logarithmic transformation has been employed in a number of group judgment studies with a slightly different justification.[3] Thus, if the quantity in question is measured by a ratio scale with a natural zero, but no upper bound (such as length, height, age, etc.,) then for a given T the range of underestimates is fixed, namely, estimates between 0 and T, while the potential range of overestimates is unlimited, from T to ∞. In actual practice, the range of potential overestimates is usually not quite so grand, but it still may be much larger than the range of underestimates. The distance measure D(R,T) = log R/T = log R - log T evens out these two ranges. $-\infty \leq \log R - \log T \leq \infty$.

More generally, a transformation G(R), where G is monotonically increasing in R, may be employed. The figure of merit then becomes D(G(R),G(T)). One frequently employed transformation in statistics, where a set of estimates of the same quantity is elicited, is the so-called normal score or z-score

$$G(R) = (R - \bar{R})/s_R. \qquad (2)$$

The use of the standard deviation as a normalizing factor is widespread in statistics. In fact, the standard deviation itself has sometimes been

employed as a figure of merit — the so-called standard error. There is a large family of statistical figures of merit, associated with the notion of statistical significance. It will not be possible to deal with this family in the present exposition. Some of the statistical measures are closely related to probabilistic figures of merit to be discussed below.

Correlations. One statistical figure of merit closely related to the factor model approach to estimation is the correlation. In the case of the factor model, the object of interest (for the investigator) is the model, i.e., the process by which estimates (of a given kind of quantity) are generated, not a single estimate. Thus an individual can be asked to generate a set of responses $\{R_j\}$ each with a different true response T_j. The question is, then, how closely the set $\{R_j\}$ matches the set $\{T_j\}$. One widely employed measure is the average error, $1/n \sum_j D(R_j,T_j)$. More frequently, the average squared error, $1/n \sum_j D(R_j,T_j)^2$ is employed for reasons given above. It is also quite common to first compute z-scores for the R's and T's.

The average squared error has one drawback. If all the R's contain a large bias, even if the R's match the T's well otherwise, the average squared error will obscure the fact. A measure which overlooks the bias is the correlation, usually defined as

$$\rho_{RT} = \frac{1}{n} \sum_j \frac{(R_j - \bar{R})\ (T_j - \bar{T})}{s_R s_T} \tag{3}$$

If we set $R' = (R - \bar{R})/s_R$ and $T' = (T - \bar{T})/s_T$, (3) becomes

$$\rho_{RT} = 1/n \sum_j R'_j T'_j \tag{4}$$

As might be expected, there is a close relationship between the average squared error and the correlation. This relationship is given by the formula

$$1/n \sum_j D(R_j,T_j)^2 = \mathrm{Var}(R) + \mathrm{Var}(T) - 2s_R s_T \rho_{RT} + (\bar{R} - \bar{T})^2 \qquad (5)$$

Correlation has received a "bad press" ranging from the frequently iterated statement that correlations do not display causal relations, only "associations," to recent contentions that correlations are inherently weak measures.[4] This contention has been pressed most strongly for multiple correlation where several variables are involved. Roughly speaking, the idea is that if several variables, f_i are used to predict a given quantity T, then if each of the variables individually are positively correlated with T ("monotonically related" is the more common expression) a high correlation will obtain for about any "reasonable" linear combination of the variables; thus the correlation is uninformative about the structure of the model $T = F(f_1,\ldots,f_n)$.

The first objection — the non-causal import of correlation — is not particularly troublesome when a correlation is being used as a score, rather than as an adjunct to theory building. For the general case of non-perceptual estimates there is no presumption that an estimate is causally related to the phenomena being described. However, the fact that correlation ignores bias is somewhat more serious. For theoretical investigations, demonstrating that a significant correlation between estimates and true answers exists does not imply that the responses will be particularly close to the true answers, only that they will covary in a reasonable way.

The second objection is more to the point. If a relatively good score can be guaranteed beforehand simply by the structure of the estimate and the

way the score is computed, the score may not be a useful measure of the excellence of the estimate. This issue will be explored more thoroughly in the next chapter on nominal judgments.

2. Probabilistic Scores

Scores for probability estimates warrant a section of their own. The conceptual problems involved are much deeper than those associated with magnitude estimates. There has been something like a major breakthrough in the past decade or so in the formal theory of probabilistic scores, but the significance of this theory for decision analysis is still under lively exploration.

In the spirit of distance scores, if an individual estimates that the probability of an event E is R, and the true probability is P, then $D(R,P) = |R - P|$ or $D(R,P)^2 = (R - P)^2$ might appear to be reasonable figures of merit for R. The only difficulty with this measure is that in practice the true probability P is usually unknown. However, there is a way of obtaining a closely associated measure, without knowing P, by using the notion of expectation. This can be illustrated by the expectation of the characteristic function for an event. The characteristic function C(E) for the event E is equal to 1 if E occurs, and equal to 0 if it does not occur. The association of the characteristic function and the truth value of the statement "E occurs" is clearly quite close. Now the expectation of the characteristic function is just equal to the probability of E. In symbols, $Ex(C(E)) = P(E)$. This statement can be made without knowing P(E).

Generalizing this notion, we can look for a score which has the property that the expectation of the score for R minus the expectation of the score for P is just the distance squared between R and P. In symbols, we can try

to design a score S(R) such that $Ex(S(R)) = (R - P)^2$. This goal is, in fact achievable. The quadratic score described below has this property. However, it turns out that a somewhat weaker requirement leads to a more fruitful theory.

No matter how we want to measure the discrepancy between R and P, we clearly would like the measure to be a minimum when R equals P. In symbols, we would like $Ex(S(R)) - Ex(S(P))$ to be a minimum at $R = P$. In somewhat more expressive notation, we would like a score function S(R,E), which assigns a score given that E occurs, where R is the estimated probability of E. For estimates of the probability distribution $R = (R_1,\ldots,R_m)$ for a partition $E = (E_1,\ldots,E_m)$ of U, we can write the score as S(R,j), the score given that event E_j occurs, and R is the estimate. Our discrepancy condition then becomes

$$\sum_j P_j S(R,j) \leq \sum_j P_j S(P,j) \tag{6}$$

This formulation allows for the possibility that the score associated with event E_j may depend on the entire distribution R, rather than just on the estimate R_j of the probability for the event E_j.

(6) in one variant or another, and in a plethora of interpretations, has formed the basis for a large proportion of recent investigations in probabilistic scores.[5] The condition, of course, does not guarantee that there is a function S(R,j) that fulfills (6); however, as it turns out, there is a large family of such functions. A number of examples are listed later on. The remarkable thing about (6) is that despite the simplicity of its origin, it imposes a number of properties on scores which are desirable in light of their potential role in decisions.

The family of scoring rules characterized by (6) has been called proper scores, admissible scores, or reproducing scores, the latter from the fact that for a probability distribution P, the maximum expectation is obtained when R = P.

The desirable properties of scores fulfilling (6) include the following: (a) The score is operational, that is, it can be assigned on the basis of a single instance. If a weather forecaster says "The probability of rain tomorrow is R" and it doesn't rain, an index can be attached to his forecast without waiting for a thousand forecasts. How useful that single score may be is another matter. (b) The score rewards the forecast for accuracy; i.e., the expectation increases as the report R gets closer to the actual probability. (c) With a small additional assumption, the score rewards the forecast for definiteness; i.e., the expected score increases as R tends toward probability one for some alternative, and toward zero on the others. (d) The score rewards a forecaster for honesty. If the forecaster believes Q and asserts R, then his subjective expectation is a maximum when Q = R. This last property has been used as a basis for imposing (6) by many investigators who are dubious of "objective" probabilities. (e) The score rewards the estimator for increasing his information concerning the events before formulating his report; i.e., his expected score is greater if based on more information. (f) S(R,j) can be employed as a figure of merit for general statements of the form "All A's are B's" if this is translated as P(B/A) = 1.

The last property suggests a simple resolution to a long standing controversy concerning the usefulness of a probability logic. Although a number of attempts have been made to formulate the probabilistic analogue of traditional two-valued logic, none of these have caught the imagination either of

logicians or of decision analysts. The reason would appear to be that these attempts have been based on the assumption that the appropriate analogue for the two-valued truth values true and false in the case of probability statements is just the probability. Conceiving of truth values as scores, the appropriate analogue in the case of probability statements would be the score as defined by (6). Some additional comments on this possibility will be made in the next section.

Properties (b) and (c) are established in the following results. We first note that the expectation "averages out" the individual alternative events, thus we can abbreviate $\sum_j P_j S(R,j)$ by $G(P,R)$, and $\sum_j P_j S(P,j)$ by $H(P)$. With this notation, (6) becomes $G(P,R) \leq H(P)$. A fundamental property of scoring functions defined by (6) is that $H(P)$ is convex; that is, $H(aP + (1-a)P') \leq aH(P) + (1-a)H(P')$ where $0 \leq a \leq 1$. In words, H is convex, if the average of H at two different points is greater than the value of H for the average of the two points. A convex function is one such that a line (or hyperplane) tangent to the function at some point always remains below the function.

Theorem 1. $H(P)$ is convex

Proof: Let $P'' = aP + (1-a)P'$, whence

$$H(P'') = \sum_j (aP_j + (1-a)P'_j)S(P'',j)$$

$$= a \sum_j P_j S(S(P'',j) + (1-a) \sum_j P'_j S(P'',j)$$

$$= aG(P,P'') + (1-a)G(P',P'')$$

From (6) $G(P,P'') \leq H(P)$ and $G(P',P'') \leq H(P')$, whence

$H(P'') \leq aH(P) + (1-a)H(P')$

Theorem 2. $S(R,j)$ is a maximum when $R_j = 1$.

Proof: If $R_j = 1$, $H(R) = S(R,j) \geq G(R,R') = S(R',j)$

D1. R is more accurate than R' if $R = aR' + (1-a)P$, where $0 \leq a \leq 1$.

This is a somewhat restricted definition of more accurate. In a sense, each specific score rule defines its own special brand of accuracy (closeness to P). However, there is no question but that if R is on a ray between R' and P, R is closer to P.

Theorem 3. If R is more accurate than R', then $G(P,R) \geq G(P,R')$.

Proof: Since $R = aR' + (1-a)P$, we can write

$P = (R - aR')/(1-a)$. Thus $G(P,R) =$

$1/(1-a)(G(R,R) - aG(R',R))$. Similarly,

$G(P,R') = 1/(1-a)(G(R,R') - aG(R',R'))$.

Since $G(R,R) \geq G(R,R')$ and $G(R',R') \geq G(R',R)$,

the result follows.

D2. A score function $S(R,j)$ will be called normal if, when R is an equipartion (all R_j equal), $S(R,j) = S(R,k)$ for all j and k.

Theorem 4. If G is a normal scoring system, H(P) is a minimum at the equipartion, $P_j = 1/m$ for all j.

Proof: Denote the equipartion by $\tilde{P}$. From D2,

$H(\tilde{P}) = \sum_j 1/mS(\tilde{P},j) = S(\tilde{P},j)$. $G(R,\tilde{P}) = \sum_j R_j S(\tilde{P},j) =$

$S(\tilde{P},j) = H(\tilde{P}) \leq G(R,R) = H(R)$.

D3. R is more definite than R' if $R' = aR + (1-a)\tilde{R}$, $0 \leq a \leq 1$.

If R is farther away from a uniform distribution than R' in the sense that R' lies on a ray between R and $\tilde{R}$, clearly some of the R_j are greater than R'_j, whereas for those $R'_j < 1/n$, $R_k < R'_k$.

Theorem 5. If G is normal and R is more definite than R', $H(R) \geq H(R')$.

Proof: $H(R') \leq aH(R) + (1-a)H(\tilde{R})$, and since $H(\tilde{R})$ is a minimum, $H(R') \leq H(R)$.

Theorems 3-5 can be tightened up somewhat if further restrictions are placed on S(R,j), for example, if it is assumed that S(R,j), is symmetrical. However, they are sufficient to give content to the assertion that any (normal) proper score rewards the estimator for definiteness and for accuracy.

Consider a set $\mathscr{R}$ of R's such that, except for R_k for some k, the remaining R_i's are in proportion; that is, $R_i = a_i(1-R_k)$, $i \neq k$, $\sum_{i \neq k} a_i = 1$, $0 \leq a_i \leq 1$. We can call such a set determined by k.

Theorem 6. If $\mathscr{R}$ is determined by k, then for R in $\mathscr{R}$ S(R,k) is monotonic in R_k.

Proof: Let R, and R' be members of $\mathscr{R}$ where $R_k > R'_k$.

From (6) we have

$$\sum_j R_j S(R',j) \leq \sum_j R_j S(R,j)$$

$$\sum_j R'_j S(R,j) \leq \sum_j R'_j S(R',j)$$

Whence $\sum_j (R_j - R'_j)(S(R', j) - S(R,j)) \leq 0$, that is

$$(R_k - R'_k)(S(R',k) - S(R,k)) + (R'_k - R_k) \sum_i a_i (S(R',i) - S(R,i)) \leq 0$$

By assumption, $R_k - R'_k > 0$, thus if $S(R',k) - S(R,k) \geq 0$,

$\sum_i a_i (S(R',i) - S(R,k)) \geq 0$. But this implies

$\sum_j R_j S(R',j) \geq \sum_j R_j S(R,j)$, contrary to (6).

Corollary 1. For a two alternative score rule, S(R,j) is monotonic in R_j.

Proof: Immediate since any two-alternative R is determined by R_j.

Theorem 6 establishes a useful monotonicy property for any proper scoring function $S(R,j)$.

With regard to property (e), suppose an estimator is faced with the option either of making his judgment based on what he knows, or of obtaining further information. Can anything be said about his expected score taking one or the other option? You would probably expect that either the score obtained with additional information is better than the score without, or something is seriously wrong with the score rule. As it turns out, condition (6) is sufficient to show that any proper score rule has this desirable property.* In order to demonstrate this, we need some additional notation. Let R_i be as usual the estimated probability for event E_i. Assume that "obtaining additional information" can be described by another partition $\{I_j\}$ on U, to be interpreted as the alternative states of the world that could be identified by a particular information search. Given that the search specifies I_j as the state of the world, the individual can then estimate the probability $(R_i|I_j) = R_{ij}$ for the event E_i.

Evaluating the two options before the fact — i.e., before the decision to seek further information or not — there exists a certain probability that the search will identify each of the I_j as the state of the world. We can designate these probabilities as $P(I_j)$. They are not represented as estimates since the general result will be independent of these probabilities. However, if the decision were being made in a practical context, it would be necessary to estimate these probabilities in order to determine whether the additional

*The analysis that follows is a generalization of a result due to Raiffa.[7]

information justified whatever costs were involved. With the additional notation we can assert

Theorem 7. $\sum_i R_i S(R,i) \leq \sum_j P(I_j) \sum_i R_{ij} S(R_j,i)$

Proof: From the rule of elimination, F(3), Chap. II, $R_i = \sum_j P(I_j) R_{ij}$. Substituting for R_i in the left side of the theorem we get

$$\sum_i \sum_j P(I_j) R_{ij} S(R,i) = \sum_j P(I_j) \sum_i R_{ij} S(R,i)$$

By (6) $\sum_i R_{ij} S(R,i) \leq \sum_i R_{ij} S(R_j,i)$ for every j.

Whence the theorem follows.

This result is quite general. It applies to any scoring rule that fulfills (6). As we have seen, the result is independent of the probabilities $P(I_j)$. One assumption is concealed in the notation, namely that the estimates R_i and R_{ij} "act like probabilities," and in particular $P(I_j)$ and R_{ij} combine in the proper way.

This theorem is the analogue for scoring rules of the refinement theorem of Marschak.[8] The refinement theorem states that the only condition under which an information pattern will lead to an improved payoff over another, independently of the payoff function or the specific probability assignment, is if the first is a refinement of the second, i.e., if the first is a partition of the second. In Theorem 7 we have assumed that the partition I_j is a refinement of the (implicit) information pattern available to the individual.

Theorem 7 has a number of applications. It substantiates the intuitive attitude that additional information is always a good thing. This will be investigated further in Chap. V. It is a general schema that can be exploited to show that two heads are better than one, as is done in Chap. V. Finally, it

illuminates a puzzle that can plague the evaluation of a specific piece of information. The theorem holds before the fact, i.e., before a specific I_j is established as the state of the world. After the fact, the expected score from a given I_j may be smaller than the expected score without I_j. It is only on the average that the expectation is increased. Although the present conceptual apparatus is not quite sufficient to express this notion completely, it is germane to point out that any specific piece of information may decrease the accuracy of an estimate. We can predict with complete confidence that in the normal course of events, once in a while a solid, relevant piece of information will decrease the accuracy of an estimate. This point flies in the face of standard lore which holds that increased information always improves estimates.

A list of the more frequently employed proper scoring procedures would include:

1. "Scientific"*: $S(R,j) = 1$ if R_j is the maximum of the $\{R_i\}$, otherwise zero.

This simple scoring scheme can be interpreted as the justification for the ordinary scoring of objective tests, i.e., counting 1 for each correct answer and 0 for each incorrect answer. Presumably, the testee checks the answer that he thinks is most likely to be correct.

2. Brier Score. Defined only for a two-alternative estimate.

$$S(R,j) = (1-R_j)^2. \quad H(R) = R(1-R).$$

This score, devised by Brier,[10] has been extensively used by the U.S. Weather Bureau to evaluate the probabilistic estimates of weather forecasters. It is slightly anomalous in that a lower score indicates better performance.

*The name is suggested by Marshak.[8]

3. Quadratic Score. $S(R,j) = 2R_j - \sum_k R_k^2$

Brown reports that the quadratic score is the only one for which the difference between the expected score of a "perfect" forecaster — i.e., one that announces P — and one that announces R is a function solely of P-R.[11] Note that $H(P) - G(P,R) = \sum_j (P_j - R_j)^2$. A complete graph of the quadratic score for two alternatives is presented in Fig. 29, Chap. IV.

4. Spherical Score. $S(R,j) = R_j / \sqrt{\sum_i R_i^2}$

The spherical score rule is notable for the fact that it is not concave. Fig. 22 is a plot of the spherical score for the two alternative case. $H(R) = \sqrt{\sum_i R_i^2}$.

5. Logarthmic Score. $S(R,j) = \log R_j$

The logarithmic score rule has a number of properties which set it apart from the others. (a) It is the only rule which depends solely on the probability reported for the event that occurs. (b) It is the only rule which is additive over successive estimates. (c) It is a close analogue of the Shannon entropy. Note that $H(R) = \sum_j R_j \log R_j$. (d) It is the only score rule which is invariant over logically equivalent estimates. These properties will be explored more thoroughly in the next section.

The above list of proper score rules is a thin sampling of the range of scoring methods that can be devised fulfilling (6). Further examples will be discussed in Section 4. The fact that none of them has demonstrated an overwhelming superiority can be interpreted in either of two ways: (1) the field is still immature. (2) There are many different roles that a scoring procedure can play in decision analysis and no one of these dominates the others. I am inclined to follow the second interpretation.

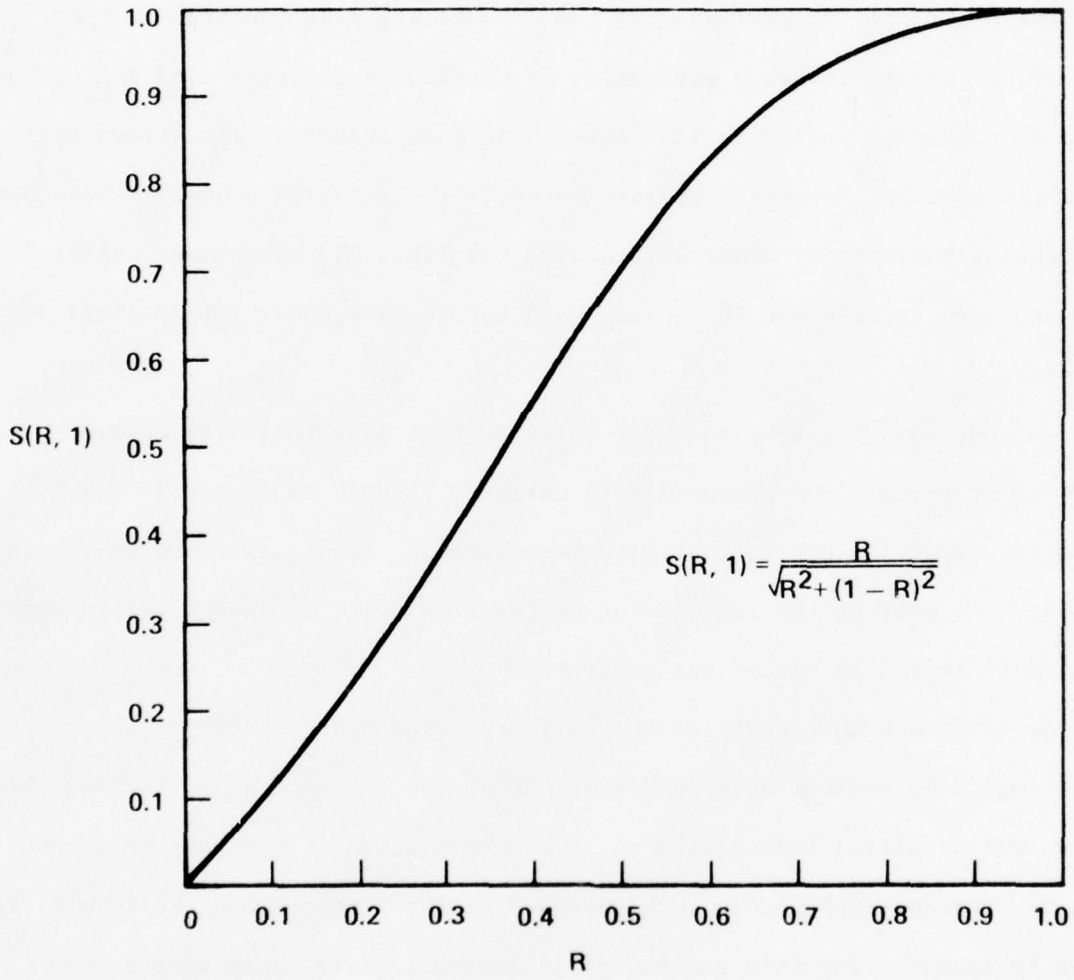


Figure 22. Spherical Score Rule for Binary Events

3. Equivalent Estimates

The preceding section approached probabilistic figures of merit from the standpoint of reproducing scores — those which have a maximum expectation when the estimate coincides with the objective probability. Although satisfying in many ways, the approach has the drawback that it has a restricted range of application, namely to estimates of the probability distribution on a partition of U. Straightforward extensions to continuous distributions exist, but these also have an analogous limitation. In many practical situations estimates are required for other logical forms, e.g., relative probabilities, disjunctive or conjunctive combinations, and the like. In many cases, initial estimates are transformed in various ways before they enter into a final decision.

Another way of making the same point is that many logically equivalent forms exist for a given probabilistic estimate. There is no simple way to accompany these transformations with corresponding transformations on scores. In addition, many of the reproducing scores give quite different values when applied to logically equivalent estimates.

We could add the condition that a reproducing score should be invariant under logically equivalent transformations of the estimate and determine the scores which fulfill this limitation. However, it turns out that this condition is extremely strong, so much so that it almost determines the form of the score by itself. For this reason, it is instructive to begin with a more general kind of estimate and examine the consequences of the equivalence condition.

We first generalize the notion of an estimate from the specification of a probability distribution on a partition to a probability tree. In practice, few estimates are solitary; they generally occur in a sequence. Marketing

estimates for a business enterprise are iterated at more or less regular intervals. Professional weather men issue a steady stream of forecasts of tomorrow's weather. Development managers periodically revise estimates of likely completion dates for projects, and the like.

This type of iterated estimate can be modelled by a probability tree. Starting at some initial point, the set of (near term) potential events can be displayed as branches with corresponding probabilities. Events that might issue from each of the initial branches can be represented by further branching, and so on. An elementary example for weather forecasts is given in Fig. 23. Two possible states for tomorrow, rain and no rain, expand into four possible states for day after tomorrow, rain following rain, no rain following rain, etc. The probabilities of given weather states day after tomorrow will depend on tomorrow's weather. The probability of rain day after tomorrow (at least in Southern California) is higher if it rains tomorrow than if it doesn't.

The branching structure contains the notion of relative probability. It also contains the notion of conjunctive and disjunctive events. The event "rain day after tomorrow following rain tomorrow" is a conjunctive event. The event "rain tomorrow" is a disjunctive event in the context of the tree; it is equivalent to the event "rain tomorrow and rain day after tomorrow, or rain tomorrow and no rain day after tomorrow."

The elements of a probability tree estimate are a set $K = \{o,x,y,z,w,\ldots\}$ of nodes (events). o is the origin (base) of the tree. Defined on K is a relation xLy, meaning x is the immediate predecessor of y. o is the only node with no immediate predecessor. The ancestral relation xL*y is defined as: there is a sequence $x_1,\ldots,x_n$ of nodes, $x = x_1$ and $y = x_n$, and $x_i L x_{i+1}$ for all $i < n$.

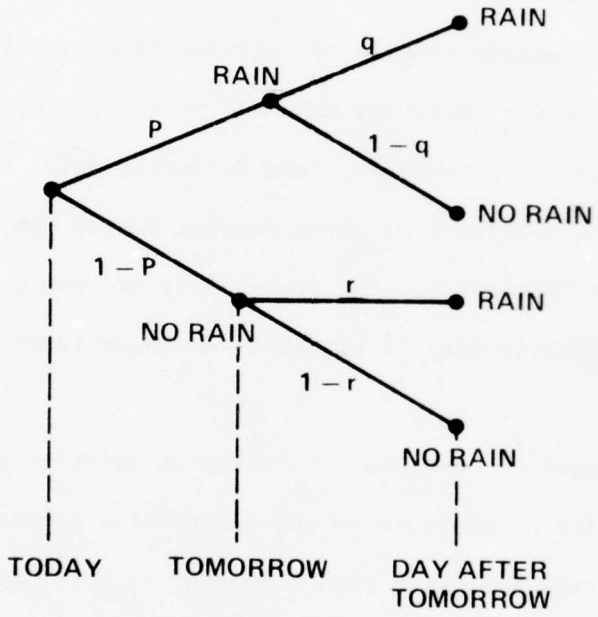


Figure 23. Meteorological Probability Tree

L(x) designates the immediate predecessor of x. B(x) denotes the set of nodes with the same immediate predecessor as x. L*(x) denotes the set of y such that yL*x plus x, i.e., L*(x) is the set of "ancestors" of x including x. Defined on the branches B(y) following a given node x is a probability distribution, where the probability of a given branch, y will be designated by P(y), and the distribution itself will be designated by $\hat{P}(y)$. Thus, if z is a member of B(y), $\hat{P}(y) = \hat{P}(z)$. This structure is illustrated in Fig. 24. Finally, $P^*(x) = \prod_{yL^*(x)} P(y)$. P*(x) is the product of the probabilities of all the nodes on the path leading from o to x. P(o) = 1.

Two probability trees K and K' are defined as being equivalent if there is a 1-1 correspondence between the endpoints of the two trees, and if x, x' are corresponding endpoints, then P*(x) = P*(x'). The normal form of a tree K is the tree K' which consists of a single stage, with as many branches as there are endpoints in K, and where for every endpoint x in K there is a corresponding branch x' in K' with P(x') = P*(x). It is an immediate consequence of this definition that two trees K_1 and K_2 are equivalent if and only if their corresponding normal forms K_1' and K_2' are equivalent.

To define a probabilistic score for an estimate K, we assume there is a function S(x) defined on each node of the tree. S(o) = 0. In non-technical terms, S is a reward, paid upon the occurrance of x. Thus, S could be a score assigned to the weather forecaster after the verification of each forecast, or it could be a fee paid to a marketing consultant after the close of each forecast period.

The expected score ES(K) of the estimate K is defined as

$$ES(K) = \sum_x P^*(x)S(x) \qquad (7)$$

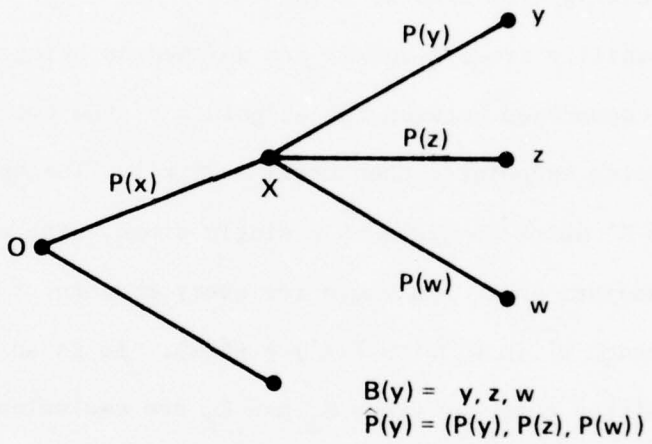


Figure 24. Illustration of Local Probability Distribution In Probability Tree

Four conditions complete the theory:

C1. $S(x)$ is a function of $\hat{P}(x)$.

C2. $S(x)$ is normal, i.e., if y is a member of $B(x)$ and $\hat{P}(y)$ is a uniform distribution, then $S(y) = S(x)$.

C3. If K is equivalent to K', then $ES(K) = ES(K')$.

C4. $S(x)$ is continuous in $\hat{P}(x)$.

C1 imposes the condition that the score is determined locally, depending only on the probability distribution on the fellow branches of a given branch. C2 is the analogue of D2 for reproducing scores. The essential condition is C3. If two estimates are equivalent, they will generate equal expected scores.

Theorem 8. The only function $S(x)$ fulfilling C1-C3 is

$$S(x) = c \log P(x).$$

Proof: If $P(x)$ is a uniform distribution, then from C1 and C2, we can write $S(x) = S(1/n)$ where n is the number of alternatives in $B(x)$. Consider a two-stage tree, K, where there are n branches in the first stage, and m branches at each second stage. Designate the nodes of the first stage as x_i and the endpoints of the second stage as x_{ij}. Assume that $\hat{P}(x_i)$ is a uniform distribution. Thus $S(x_i) = S(1/n)$ and $S(x_{ij}) = S(1/m)$.

The normal form of K has $n \times m$ endpoints, x'_{ij}, with a uniform distribution $S(x'_{ij}) = S(1/mn)$. C3 requires that $\sum_i P(x_i)S(x_i) + n\sum_i P(x_{ij})S(x_{ij}) = \sum_{ij} P(x'_{ij})S(x'_{ij})$. Thus $n(1/nS(1/n)) + n(1/n(m(1/m\ S(1/m)))) = nm(1/nm\ S(1/nm))$. Whence $S(1/n) + S(1/m) = S(1/nm)$. The only continuous function with this property is $c \log P(x)$.* This proves the theorem for uniform distributions.

*This fact is "well-known." For the curious, a proof is given in Appendix I.

To prove the result for non-uniform distributions, consider a tree K consisting of two stages; where the probability distribution on the branches of the first stage is not uniform. (Cf. Fig. 25). Assume that P(i) is of the form $a_i / \sum_i a_i$, with a_i an integer. At each endpoint i of the first stage there are a_i branches. Assume that $\hat{P}(ij)$ is uniform, i.e., $P(ij) = 1/a_i$. There are thus $\sum_i a_i$ endpoints to the two-stage tree, and $P^*(ij) = (a_i/\sum_i a_i)\ 1/a_i = 1/\sum_i a_i$. The normal form K' of K is thus a single stage with $\sum_i a_i$ endpoints and with a uniform probability distribution. Thus, from our previous result, $ES(K') = \log(1/\sum_i a_i)$.

$$ES(K) = \sum_i (a_j/\sum_i a_i) S(\hat{P}(j),j) + \sum_j (a_j/\sum_i a_i)\ \log(1/a_j)$$

Equating ES(K) and ES(K') we obtain

$$\sum_j a_j/\sum_i a_i\ S(\hat{P}(j),\ j) = \sum (a_j/\sum_i a_i)\ \log\ (a_j/\sum_i a_i)$$

Invoking continuity, we arrive at

$$\sum_j P_j S(P,j) - \sum_j P_j\ \log\ P_j$$

which was to be proved.

The two basic assumptions leading to this result are C1, the score for a given node is a function of the probability distribution on the fellow braches of the node, and (7), the assumption of conditional additivity. That these two quite general conditions could specify the form of the score precisely without invoking any ordering conditions is quite surprising. Of course, in order to use the score for a figure of merit, the constant c must be assigned, and depending on the sign of c, the score can increase with increasing probabilities, or decrease. However, the question of sign appears to be secondary. The Brier score is one in which a small score is desirable, but that creates no great confusion in interpretation.

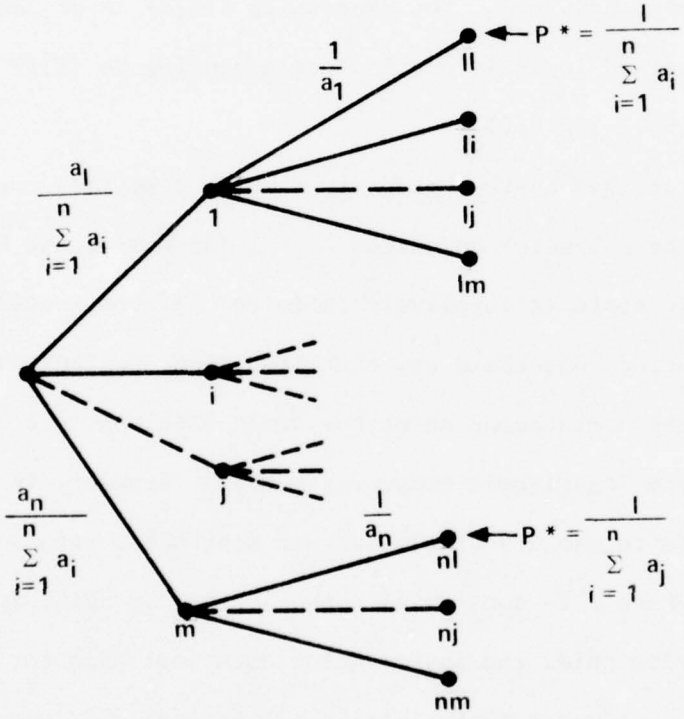


Figure 25. Probability Tree with Non-uniform Probabilities in First Stage

The logarithmic score would thus, at first glance, appear to be the only reasonable candidate for the analogue of truth values for probabilistic statements. It has one nice formal property which can be stated as

$$S(E.F) = S(E) + S(F) \text{ if } E \text{ and } F \text{ are independent}$$

$$S(E.F) = S(E) + S(F|E) = S(F) + S(E|F) \text{ otherwise} \qquad (8)$$

Unfortunately, there is no similar neat expression for $S(E \vee F)$ or for $S(\bar{E})$. They can, of course, be computed from the probabilities of the components, but not in a simple functional form. The expression $S(E|F)$ is of some interest. There is no two-valued logical function corresponding to $(E|F)$. $S(E|F) = \log P(E|F) = \log P(E.F) - \log P(F)$.

The logarithmic score has another property that is sometimes considered a drawback, namely if the estimator announces $R_j = 1$ for some event E_j, and E_j does not occur, then his score is negatively infinite. If one wanted to be moralistic about the matter, one could say that it served the individual right — no one can assert a statement about the world with absolute certainty, and that is just what the logarithmic score recommends. However, in practice, it is somewhat of a bore to qualify highly certain statements with some such hedge as $P(E) = 1 - \varepsilon$ where ε is some small number. In many applications of the logarithmic scoring rule, the investigator does just this for the estimator; that is, the estimated probabilities are truncated at some point close to 1 and close to zero. There is something slightly unsatisfactory about this tactic, since the computed score may be highly sensitive to the truncation point.

The analogy of Theorem 8 with Shannon's theorem that the entropy $-K \sum_i p_i \log p_i$ is the only function that is continuous and conditionally additive for probabilitistic message sources, is quite close. The basic

difference is the starting point. An identical theorem could be proved for information, if, for example one started by defining a notion of amount of information in a message (rather than the uncertainty of a source) and assumed that the expected information from a source had the form (7).

Despite all the advantages that the logarithmic score has going for it, it has not received the overwhelming endorsement of the community interested in evaluating estimates. The advantages do not appear to add up to any dramatic practical consequences. Some relevant substantive considerations will be raised in the following chapter.

4. Decisional Scores

The scores that have been discussed so far could be called informational; they concern the degree to which a response is correct. In the decision context, there is a more natural criterion, namely, to what extent does a response improve the outcome of the decision? To deal with this question in full generality it would be necessary to first develop the theory of utility. For simplicity in this section I will assume that the outcome of a decision can be assessed on a value scale which is linear in probabilities, i.e., a utility function.

Referring back to Fig. 1 in the Introduction, a decision can be characterized by a set of actions A_i, a set of events, E_j, and a matrix of outcomes O_{ij}. Given a utility function $U(O_{ij}) = U_{ij}$ and a probability assignment $R(E_j) = R_j$, the expected utility for action A_i, $U(A_i) = U_i$, is just $\sum_j R_j U_{ij}$. The decision rule normally associated with this analysis is, choose the action A_i that maximizes U_i. We first show that any decision matrix, with the maximize-expected-utility decision rule is a proper scoring rule for the probability estimate R. Define $U^*(R,j)$ as the U_{ij} for the action A_i that

maximizes $\sum_j R_j U_{ij}$. By definition

$$\sum_j R_j U^*(R,j) \geq \sum_j R_j U^*(R';j) \tag{9}$$

(9) is precisely of the form (6) with $S(R,j) = U^*(R,j)$.

It is thus possible to apply most of the general properties of proper score rules derived in the previous section to any decision matrix. However, most decision matrices are not normal. Thus, Theorems 4 and 5 do not hold in general for decisional score rules.

The converse of the statement that any decision matrix defines a proper score rule also holds; i.e., any proper score rule can be represented by a decision matrix. This follows trivially if the actions A_i are defined to be the report of a probability distribution R on a set of events $\{E_j\}$ and U_{Rj} is defined as $S(R,j)$. The triviality is, of course, that the optimal action for any assignment R of probabilities to the events is just R itself.

A somewhat more revealing exposition of the converse can be made if we start off with a general decision space X, i.e., X describes the potential actions. Consider any set of functions $f_i(x)$, $i = 1, \ldots, m$. This set of functions defines a proper scoring rule for a probability estimate $R = (R_1,\ldots,R_m)$ under the rule select the x such that $\sum_j R_j f_j(x)$ is maximized. This is really just another way of saying (9), with $f_i(x) = U_{xi}$. If the f_i are differentiable;

$$\sum_i R_i \frac{\partial}{\partial x_j} f_i(x) = 0$$

Call the solution to this system of equations $x^*(R)$, the x that maximizes $\sum_j R_j f_i(x)$ given R. Then $f_i(x^*(R)) = g_i(R)$ is the proper score rule defined by the set of functions f_i. If the f_i are initially a proper score rule, $g_i(R) = f_i(R)$.

For example, if X is two-dimensional, so that it can be specified by the single parameter x, $0 \leq x \leq 1$, and $f_1(x) = \sqrt{x}$, $f_2(x) = \sqrt{1-x}$, differentiating and performing the substitution gives $g(R) = R/\sqrt{R^2 + (1-R)^2}$ — the spherical scoring rule. If $f(x) = x^2$, $g(R)$ is the Brier scoring rule. Since, on this approach, there are no restrictions on the form of the functions f_i (other than differentiability), it is convenient way to derive a wide variety of informational scoring rules and at the same time obtain some insight into the nature of the kind of decision which is being made with that rule; i.e., the kind of payoff function which is (implicitly) being maximized in applying the rule.

A somewhat more intricate application of the approach can be made if the notion of repetitive decisions is introduced. In the section on equivalent estimates it was pointed out that many kinds of estimates are iterated in a fairly routine manner. In effect, this is a symptom of the fact that many decisions are iterated rather routinely. For simplicity, consider a prototype decision matrix U_{ij}, with a fixed number of rows and columns, but with varying values. We can conceive of each such matrix as a member of a sequence of decision problems where the form of the decision remains the same, but the U_{ij}, and the relevant probabilities, change from case to case. For each case, the decision maker selects the action A_i which is optimal for his estimate R of the probabilities. The decision function will map the space of matrices onto a partition U_i, where U_i is the set of matrices for which the action i is optimal. Strictly speaking the U_i need not form a partition — for some matrices more than one action may be optimal; but for simplicity we assume that such matrices are assigned to only one set U_i.

If we assume that there is a joint distribution D(U) on the space of matrices, and assume in addition that the decision maker's estimate R is independent of this distribution, then the expected payoff is

$$\sum_i \int_{U_i} \sum_j R_j U_{ij} \, D(U) \tag{10}$$

Or, interchanging the summation signs

$$\sum_j R_j \sum_i \int_{U_i} U_{ij} D(U) \tag{11}$$

Thus, we can set

$$S(R,j) = \sum_i \int_{U_i} U_{ij} D(U) \tag{12}$$

(12) is the decisional score rule defined by the sequence of the decisions with "random" matrices.

As a simple example, suppose we have an individual who engages in frequent bets on a binary event, e.g., win-lose types of bets on athletic contests. Each bet has the decision matrix

	E	$\bar{E}$
1. Bet on E	(1-u)/u	-1
2. Bet on $\bar{E}$	-1	u/(1-u)

Where the outcomes are the appropriate odds for a bet on an event with probability u. Thus, we assume some one offers the stated odds and the individual can choose which side to bet on, with a standard bet of 1. Following the maximization rule, the individual would bet on E if his subjective probability for E is greater than u, otherwise he would bet on $\bar{E}$.

We obtain a strategically equivalent matrix if we add 1 to each entry, giving the matrix

	E	$\bar{E}$
1. Bet on E	$1/u$	0
2. Bet on $\bar{E}$	0	$1/(1-u)$

Now suppose the individual is presented with a sequence of such bets, where the distribution of u in the sequence is D(u). In this case U_1 is just $0 \leq u \leq R$, and U_2 is $R \leq u \leq 1$. Since $U_{12} = U_{21} = 0$, we have from (12)

$$S(R,1) = \int_0^R D(u)/u$$

$$S(R,2) = \int_R^1 D(u)/(1-u) \tag{13}$$

Thus, the gambling sequence generates a variety of score rules depending on D(u). For example, if D(u) = constant, the logarithmic score rule ensues. If D(u) = ku(1-u), the spherical rule is generated, and so on.

The sequential model shows that a score rule may look very different derived from a single decision compared with one derived from a sequence of similar decisions. The informational score rules, which may seem irrelevant in the case of a specific decision can make a great deal of sense if the given decision is embedded in a sequence.

From this point of view, the logarithmic score rule would be appropriate for the "complete ignorance" situation where any given "opportunity" — the betting odds parameter u — is as likely as any other. The spherical rule would be appropriate if the distribution of "opportunities" is peaked about $u = \frac{1}{2}$, with relatively few at the extremes, and so on. For many of the simpler kinds of decisions — especially for betting decisions — the distribution of opportunities can be obtained empirically. For such cases, the appropriate score rule could be computed from the data. An investigation of this topic

would shed some light on the role of various informational score rules in guiding decision procedures. This point will be expanded in Chapter V in discussing the evaluation of group probability judgments by informational score rules.

Decisional score rules have been called "piece of the action" rules by Savage, in the context of rewarding a consultant for his advice. One way to reward a consultant is to pay him a certain proportion of the profit his advice creates for the enterprise. Raiffa calls decisional score rules "naturally imputed" rules, on the grounds that they derive directly from a decision problem. Some of the issues implied by these terms will be taken up in the next section on scores as motivators.

5. Motivational Role of Scores

The basic emphasis in this chapter has been on scores as measures of excellence, essentially accuracy, as measured by the distance from the true answer, or less directly, the regret – the difference in expected utility achievable by the correct judgment and a given estimate. Lurking in the background has been the notion that scores also can act as motivators; an individual will tend to maximize his expected score. Thus, it is usually thought that school grades are both a measure of the performance of a student, and also a spur to better performance. In many traditional economic texts, wages are treated primarily as rewards which motivate workers to perform requisite tasks. The boundary line between scores and rewards (or reinforcing agents, to use the Skinnerian term) is thus quite fuzzy.

What brings this topic alive is the possibility that a given scoring scheme may backfire. What appears to be a completely reasonable measure of excellence can induce behavior that is quite contrary to what was intended in formulating the score. Suppose there is a specific kind of response Q which

is desired. Consider a function S which rewards an individual with S(R,T) if the individual responds with R and criterion T applies. Such a function will motivate the individual to perform Q if his anticipated reward F(Q,R,T), computed from S(R,T), is a maximum when R = Q.*

An example might clarify the role of these two functions. The following example is due to Marschak.[12] Suppose we have an individual who has an article for sale—e.g., a house. He sets some value on this article, which we may as well think of as its worth to him in money. Most bargaining procedures tend to motivate the individual (initially at least) to set a higher price on the article than its worth to him. Can an exchange process be designed that would motivate the individual to reveal his "true price"? The following scheme will do it. The potential buyer submits a sealed bid, T. Without knowing the bid, the seller announces a price, R. The bid is then compared with the price. If the price is less than or equal to the bid, the exchange is made at the bid amount. In this case the desired behavior is announcing the true price, Q. S(R,T) = T if $R \leq T$, otherwise 0. F(Q,R,T) = T - Q if $R \leq T$, otherwise 0.

Figure 26 displays the situation. Potential asking prices R define the vertical scale and potential bids T define the horizontal scale. To the left of the 45° line, $R \leq T$, no exchange takes place, and S(R,T) = 0. Similarly, F(Q,R,T) = 0 in this region. In region A, F is negative. In the region to the right of the 45° line, and beyond T = Q, F is positive, and independent of

---

*If this scheme were to be taken seriously as a practical technique for eliciting the response Q, a number of other conditions would be imposed on the functions S and F. Among these would be that S represents the total "sum of rewards" involved in the situation; that F is apparent to the individual; that the maximum is not too flat; that the reward is appropriately timed with respect to the behavior; and so on. However, for purposes of theoretical investigation, the usual issue is the appropriateness of a given response R, not the probability that it will be elicited in fact. Thus, these practical conditions are commonly omitted.

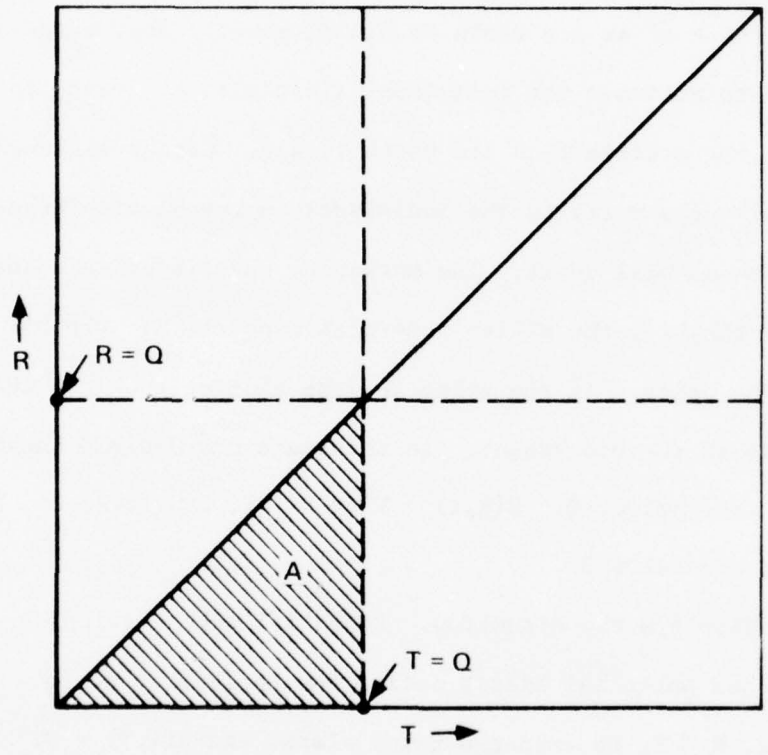


Figure 26. Bidding Procedure to Motivate Honesty

R. Thus, for any T, $F(Q,R,T) \leq F(Q,Q,T)$. The seller cannot lose if he announces Q, and he may lose if he announces an $R \neq Q$. In standard parlance, announcing Q dominates any other announcement.

The example of a voting scheme described in Chap. VI, Sec. 5 employing random selection of a final slate, is another illustration of a scoring scheme which generates a desired kind of estimate — in this case the honest rating of a candidate.

An informative example of a presumed proper scoring scheme which fails rather miserably can be found in test grading. It is common practice in scoring objective examinations to do what is called "correcting for guessing." The basic assumption is that the student can get the right answer by guessing at least 50% of the time. To discourage the student from guessing, his recorded score is computed as $C = W/(a-1)$, where C is the number correct, W is the number incorrect, and a is the number of alternative answers to the question. For true-false questions, the score is just C - W (rights minus wrongs). The speech which accompanies this score goes like this: if the student guesses, and guesses wrong, he will be "punished" by subtracting a point from his score.

It is a simple exercise to show that this scheme is futile (providing the assumption on which it is based is correct.) Suppose the probability that a given student will give the correct answer on question j is q_j. His expected score, if he responds to every question is $\sum_j (q_j - (1 - q_j))$ $= 2 \sum_j q_j - n$, where n is the number of questions in the test. If he responds to only m of the questions (for whatever reason), his expected score is $2\sum_i^m q_j - m$. The difference between the two scores is $2 \sum_{j=m}^{n} q_j - (n-m)$.

Since the basic assumption is that $q_j \geq \frac{1}{2}$, $2 \sum_{j=m}^{n} q_j \geq n-m$, and the difference is positive. Thus, the student never loses (on the average) by guessing.*

In the following chapter we will see that the assumption that the student always has an expectation of at least $\frac{1}{2}$ in guessing the answer to a question is just false, and that, in fact, this myth covers a major gap in the theory of test design. For the present discussion, however, the moral is that a scoring system which is in widespread use and which is intended to motivate a kind of behaviour, simply doesn't do its job.

Many Bayesians who object to the notion of "correct probability" for single events, prefer to approach the theory of probabilistic scores from the aim of motivating the estimator to be honest. Thus, if the individual believes Q is the probability distribution on a set of events, he might be motivated to report something different, depending on how he is rewarded for his estimate. The condition on the score that will motivate honesty is just

$$\sum_j Q_j \, S(R,j) \leq \sum_j Q_j S(Q,j) \qquad (14)$$

That is, the individual's subjective expectation is a maximum when he reports his believed probabilities. Fourteen is formally identical to 6, with Q replacing P. Hence 14 leads to the same family of scoring rules as 6.

It is easy to formulate probability scores which appear reasonable, but which violate 14. If $S(R,j) = R_j$, i.e., the score is just the reported probability of the event that occurs, several intuitively reasonable conditions are met. The score increases with the reported probability. H(R) is convex, in fact $H(R) = \sum_j R_j^2$, which is identical to H(R) for the quadratic

*Of course, he may increase the variance of his score by responding to questions where q_j is close to ½. But I have yet to see a justification for the procedure which invokes a trade-off between expected score and variance.

scoring rule. Nevertheless, $S(R,j) = R_j$ does not fulfill 14, and it is easy to see that if this scoring rule is used, the individual will always report either 0 or 1 for the probability of an event. His expected score $\sum_j Q_j S(R,j) = \sum_j Q_j R_j$ is maximized when he reports 1 for the event with maximum Q and 0 for all the others.

A somewhat more amusing case is $S(R,j) = R_j^2$. This might look at first sight like a variant of the quadratic score. However, the expression $QR^2 + (1-Q)(1-R)^2$ is a maximum precisely when $R = (1-Q)$; the score motivates the individual to report the "opposite" of what he believes!

The non-intuitive nature of these "anomalies" suggests a second look at the distance scores introduced in Sect. 1. Although from the standpoint of measuring discrepancies between a report and a true answer they appear impeccable, what can be said for them from the viewpoint of motivating responses?

A subject answering a question in the laboratory with the instuctions "make as good an estimate as you can," must be guided by some rough idea as to what the experimenter thinks is a "good answer." Or lacking any guidance in the instructions, he can only proceed on what he thinks is a good answer. By now it should be clear that there is no well defined content to the term "good answer." Suppose, for example, that the subject thinks that the proportional score is reasonable — he would like, if possible, to make a small percentage error. Thus, in a loose way, he is trying to minimize $|R - T|/T$. If we assume he has a subjective probability distribution D(T) on T, then he will seek to minimize his expectation of his proportional error; that is, he will try to minimize

$$\int \frac{|R - T|}{T} D(T)$$

with the integral extending over his subjective range for T. If the minimum is computed, it turns out to occur at R^* where

$$\int^{R^*} \frac{|R-T|}{T} D(T) = \int_{R^*} \frac{|R-T|}{T} D(T)$$

R* could be called the proportional median. It is an unusual statistic. Generally, it is quite small, smaller even than the harmonic mean. For example, if D(T) is a uniform distribution between a and b, then $R^* = \sqrt{ab}$. Suppose a = 10, b = 100. The mean of this distribution is 55, the geometric mean is 47.5, the harmonic mean is 39.01, and R*, the proportional median is 31.62. If the subjects in this experiment were realistic in the sense that the average of their subjective probability distributions correspond rather well with the true answers, then a large majority of their answers would appear to be underestimates.

In an extensive series of experiments at the Rand Corporation, with college student subjects and general information type questions, a majority of the responses were in fact underestimates.[13] The 65th percentile was a better estimator of the true answer than either the mean or the geometric mean.

The assumption that the subjects were expressing the proportional median of their subjective probability distributions appears a little too drastic. Suppose we invoke the theory of errors model and make the following assumptions:

(1) The subjects have a subjective probability distribution D(T) on the potential answers.

(2) Following the psychonumeric hypothesis, D(T) is log normal.

(3) The subjects are roughly realistic — i.e., the actual answer $A = e^{\int \log T D(T)}$

(4) The individual evaluates his answers with the error-squared score; i.e., he tries to minimize $\int \frac{(R-T)^2}{T} D(T)$.

The last assumption implies that the individual will respond with the harmonic mean of his distribution. The harmonic mean of a log normal distribution can be computed readily, it is $e^{\mu-\frac{1}{2}\sigma^2}$ where as in Chapter II, Sec. 7 μ is the mean of the log transform distribution and σ is its standard deviation. Thus, the harmonic mean occurs at the 31st percentile. This is to be compared with the observed 35th percentile. Restating the point, in the Rand data, about 65% of the responses were underestimates. On the present theory, about 69% would be underestimates. Considering the fact that the Rand data are based on a variety of sample sizes (ranging from 13 to 29) on different questions, and that we are smearing the averages over a large population with a small number of questions, the figure does not appear out of line.

The data, as analyzed is not sufficient to assert with high condidence that the subjects were indeed responding with the harmonic mean of their subjective probability distributions. However, the results are highly suggestive. In particular, they suggest that possibly much of the apparent bias observed in probability estimates could be due to completely reasonable behavior on the part of the subjects. They may be responding to an implicit scoring rule which has consequences that the experimenter has not anticipated.

In the present context, we can examine the implied response for a variety of scoring rules. These are obtained by minimizing on R the integrals

$$\int S(R,T)D(T)$$

Minimum expected score responses for various

(magnitude estimation) scores

1.	$\|R-T\|$	Median
2.	$(R-T)^2$	Mean
3.	$\|R-T\|/T$	Harmonic Median
4.	$(R-T)^2/T$	Harmonic Mean
5.	$\|\text{Log } R - \text{Log } T\|$	Geometric Median
6	$(\text{Log } R - \text{Log } T)^2$	Geometric Mean

If we apply the preceding type of analysis to probabilistic scores we arrive at a surprising result. Consider an individual who has a subjective distribution D(P) over the range K of possible objective probabilities P. His expected score, for response R will be

$$\int_K \left[\sum_j P_j S(R,j)\right] D(P) = \sum_j \bar{P}_j S(R,j) \qquad (15)$$

where $\bar{P}_j$ designates the average of the P_j's. According to 6, 15 is maximized when $R = \bar{P}$. This result is independent of the kind of score and follows from the linearity of G(P,R) in P. This result becomes of importance when one tries to use probabilistic scores as a method of clarifying the notion of uncertainty, expressed, e.g., as a higher level distribution on the estimates.

An instructive application of these ideas can be found in the analysis of the payment of a consultant. Suppose an expert has been hired by an enterprise to furnish inputs to a decision problem. Theorem 7 states that the expert will collect whatever additional information is available, i.e., within

the bounds of feasibility, he will try to make himself more expert on the specific problem facing the enterprise. Whatever the form of payment, it will be to his advantage to become more knowledgeable.

However, suppose the relationship between the consultant and the enterprise is of one fairly common sort, where the expert is expected to help structure the problem -- that is, he is expected to give advice concerning the relevant events to be taken into account, potential actions, and the like. To simplify the point, suppose the consultant is merely asked to suggest the relevant events, and of course, furnish a probability assignment for them. If the consultant is paid according to one of the informational score rules, then it will be to his advantage to make the list of events as small as possible. Put in informal language, it will be to his advantage to learn as much as possible, and to tell his client as little as possible! This results from the fact that informational score rules are roughly speaking monotonic in the probabilities. By choosing a smaller event list, the consultant raises the probabilities of the predicted events, and hence his expected score. For example, if the expert is a geologist who is asked to forecast the probability of an earthquake in Southern California over the next twenty years, and he has the choice between estimating the probability of earthquakes in a number of magnitude classes, or of a simple dichotomy like major or minor, then an informational score rule would motivate him to select the second forecast.*

*In actual practice, rewards for expert judgment are highly complex. A consultant may prize his reputation more than money, and reputation may depend more on precision of estimates than on accuracy. Of course, complex reward situations of this sort may not be a proper scoring situation; the expert may find his greatest reward in lying.

Decisional score rules do not have the disadvantage just discussed. Since the decisional score rule depends on the decision matrix as well as on the estimated probabilities, there is no gain in simplifying the event list. In fact, for those cases where refinement of the event list can lead to increased expected payoff, if the consultant is rewarded with a "piece of the action," he will be motivated to generate the appropriate refinement of the events.

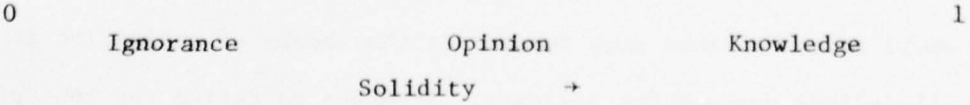
CHAPTER IV. NOMINAL JUDGMENTS

1. The Spectrum of Uncertainty

In Chapter II several theories of individual estimates were expounded, and in Chapter III a variety of ways of measuring the excellence of such estimates were explored. Underlying both chapters was the assumption that human judgment can be used as a surrogate for data or theories, at least in a decision context. Experimental investigation of human judgment has shown that it is frequently very bad. A basic issue to be explored in this chapter is the question whether there are situations in which it is preferrable to replace human judgment with something that is in a sense even weaker, namely a form of estimate that can be called a nominal judgment.

By a nominal judgment is meant an assertion based on a quasi-logical rule, such as the principle of indifference or the law of insufficient reason. Typical examples are the assumption of equal a priori probabilities in some kinds of statistical inference, or the equal weighting of individual responses inherent in using a simple average to express the group response in a Delphi exercise. Estimates of this sort are clearly a third kind of judgment, very different in origin from statements based on "hard" data, or on "intuition." They have led a sort of demi-monde existence in the past, frequently employed, but not with a clear conscience.

Before attempting to probe the foundations of this sort of judgment, it is useful to look at a related topic, namely the range of excellence that estimates can exhibit. We can display this range with a crude scale



The quantity being scaled here is roughly the degree of verification, or the amount of evidence, that exists for a statement. Because there have been many attempts to generate a formal definition for this scale, none of which to my knowledge have succeeded, I prefer to leave the notion informal and call the quantity "solidity." Thus, judgments at the right end of the scale are well-established and solid, those at the left end are unsubstantiated or "flimsy."

At the far right are the well verified generalizations of natural science and the equally solid generalizations of common sense, like "Unsupported objects fall," or "Food is necessary to sustain life." At the far left are statements for which there is no evidence whatsoever, but equally, no contrary evidence. In Chapter II, it was pointed out that it is not clear whether statements with zero solidity exist other than in theory. Just to understand a sentence probably requires some relevant information. There is an extensive literature, beginning at least as early as Hume, emphasizing the fact that the ideal of a completely certain factual statement is just that; all empirical statements are incompletely verified. In a similar way, the notion of a statement for which no relevant information exists is probably an idealization.

In the middle of the range are assertions for which there is some evidence, but not enough to inspire high confidence. I have called this range "opinion." The term doesn't seem to have caught on for this application, possibly because of its negative connotations.

At all events, it is this middle range which appears most appropriate for applying the theories of estimation of Chapter II.

It would be an enormous step forward in the theory of estimation if there were a well-defined measure for solidity. Attempts to define the concept based on formal logic, such as Keynes' logical theory of probability[1] and Carnap's

efforts to specify a degree of confirmation[2] have proven unsatisfactory. Attempts to equate the scale with probability run afoul of a number of problems, some of which will be discussed below in the section on uncertainty. The scale is clearly related to the notion of degree of confidence, defined in various ways in statistics, but most of these require special assumptions (such as randomized sampling) which limit their application to instances of formal data collection.

One relatively obvious tactic is to use a judgmental scale of solidity; e.g., to elicit a confidence rating along with each estimate. The present evidence suggests that individuals are no better at estimating the solidity of a judgment than they are at making the original estimate. The illusion of certainty phenomenon studied by Slovic, Lichtenstein, and others[3] is a clear case in point. In the series of experiments with college students and almanac questions described in Chapter II, the subjects were asked to rate their answers, usually on a scale from 1 to 5, where 1 meant "I'm just guessing" and 5 meant "I know the answer." The correlation between these self-ratings and log error was -.25. The negative sign is in the right direction, but the size of the correlation is not impressive.

For high confidence statements (knowledge), there is no basic difficulty. The rule for using such statements in decisions is simply assume the statement is true, and act accordingly. For statements in the middle range (opinion) there are a number of open issues, but the rule make the best estimate you can and act accordingly, appears to have general acceptance. In this range, expressing uncertainty with estimated probabilities is gaining credibility among decision analysts. Serious conceptual problems arise in formulating rules for incorporating statements at the low end of the solidity scale (ignorance)

in decisions. It is statements of this sort for which nominal judgments, rather than estimates, are probably appropriate.

2. Uncertainty and Probability

There is a long-standing controversy concerning the question whether probability completely encodes the notion of uncertainty. The distinction between uncertainty (or lack of information) and risk (probability) has been around at least since the writings of Frank Knight.[4] However, there has been no clear consensus on the subject among those interested in decision analysis. Those who favor a subjectivist theory of probability have been inclined to reject the distinction, on the grounds that when an individual makes a probability judgment he is quantifying his degree of uncertainty. On this view, an individual who says "The probability of event E is one-half" is saying "I haven't the foggiest notion whether E will happen or not." Some objectivists have also rejected the distinction, notably Reichenbach who remained convinced that the notion of probability was flexible enough to cover all instances of incomplete information.

Objections to the identification of uncertainty with probability have been raised by Allais, and Ellsberg, who contend that the theory of subjective probability does not describe the behavior of individuals making choices under incomplete information. This topic will be expanded in Section 6.

On the face of it, the subjectivist position is hard to maintain. Consider the assertion, "The probability of event E is one-half." As an example, suppose there are two coins, one of which is well-known to the estimator. Let's say he has flipped it many times, and has very good reason to believe it is a fair coin. The other is an exotic object with which he has had no prior experience. There is no contradiction on the subjectivist theory of probability to suppose that he asserts "The probability of heads is

one-half" for each of the coins. And yet he may make the assertion for the first coin with high confidence, and the assertion about the second coin with little or no confidence. In this case, the ascription of probability one-half to an event occurs with two very different states of knowledge.

The coin example shows that a probability judgement cannot, by itself, express the degree of certainty with which the judgement is asserted. This point is in accord with the view expressed in Chapter II that probabilities are properties of the world, not of the estimator's state of knowledge. To pursue an example raised then, an estimate of the height of a tree can be made with any degree of solidity, and the solidity is unrelated to the number expressing the height. Similarly, a probability estimate can be asserted with any degree of solidity, and the solidity is not directly related to the numerical probability.

The argument appears to require that an additional index other than probability be formulated to express the solidity scale. In statistics, it is common practice to attach a number, the significance level, to a derived statistic. In scientific applications the practice is to "suspend belief" if the significance level is too low. This scientific procedure is not much help to a decision maker if the statistic is relevant to a pending decision, and the significance level is below the accepted criterion. A low significance level does not imply that the opposite of the hypothesis has a high significance level.

The significance level is a special case of one suggestion for extending the notion of probability to include uncertainty, namely to introduce probabilities of a higher level. Thus, associated with any given estimate we can conceive of a probability distribution for that probability. If the second level distribution has a low dispersion, the estimate is relatively solid.

If the upper level distribution is flat, then the estimate is uncertain. This suggestion is illustrated in Figure 27. The second-level distribution for the familiar coin is peaked around one-half, while the second-level distribution for the unfamiliar coin is relatively uniform.

Second-level probability distributions clearly do not eliminate the problem, because the second level distribution may also be uncertain. Taken seriously, the suggestion then leads to a third level, and so on. The only one I know who whole-heartedly accepted the implied infinite set of higher level distributions is Reichenbach.[5] Most other investigators have felt that an infinite sequence of higher level distributions creates many more problems than it solves.

Nevertheless, the notion of a higher level distribution is useful for exploring some kinds of relationships between uncertainty and probability. For example, if we assume that a higher-level distribution approximately expresses uncertainty, then we can assert a coupling between the dispersion of the higher level distribution and the probability. This comes about because the range of probability estimates is constrained between zero and one. Assume that the individual asserts the average of his second-level distribution as his first level response. Then it is impossible to assert a probability close to one with high uncertainty. For example, if the upper level distribution has the form $(n+1)p^n$, and the average is .95, then n = 18, and σ = .0475. On the other hand, if the average is around .5, then the second level distribution is not constrained; it can be about anything.

More generally, consider any event space U. If U consists of a discrete set of events $\{E_j\}$, then the totality of the possible probability distributions on U is described by the simplex $\sum_j P_j = 1$. The individual may have a certain

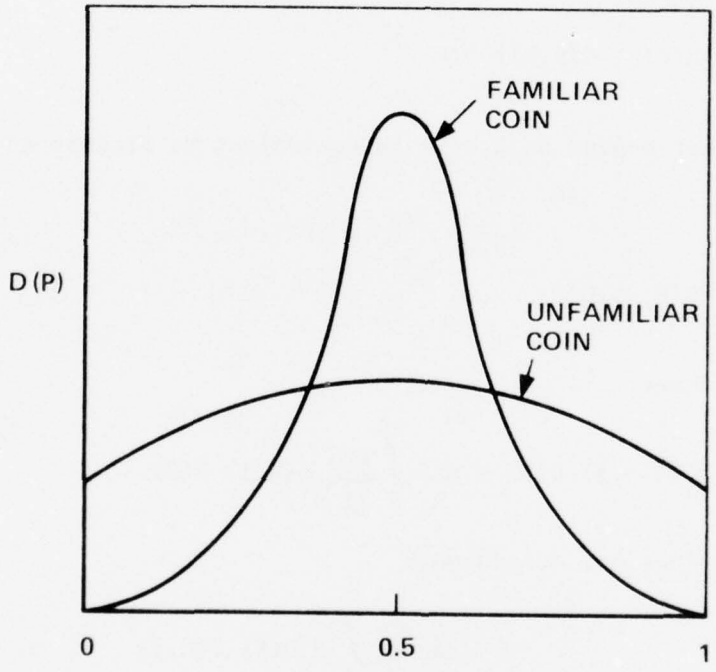


Figure 27. Second Level Probability Interpretation of Uncertainty

amount of information about U, which can be expressed by saying that he knows that the probability distributions on U are limited to a certain class K. By assumption, some distribution P, in K is the true distribution. If the individual asserts R = P, then his expected score will be H(P). If he does not know P and asserts some R ≠ P, then his loss will be H(P) - G(P,R). If there is an upper level distribution D(P) on the distributions in K, then his expected loss will be $\int_K \Big(H(P) - G(P,R)\Big)\ D(P)$.

The expected loss will be minimized by selecting an R* to give

$$\min_R \int_K \Big(H(P) - G(P,R)\Big)\ D(P) \tag{1}$$

Since H(P) does not depend on R, (1) is equivalent to finding an R* which gives

$$\max_R \int_K G(P,R)\ D(P) \tag{2}$$

Expanding (2) we have

$$\int_K \sum_j P_j S(R^*,j)\ D(P) = \max_R \int_K \sum_j P_j S(R,j)\ D(P)$$

And since S(R,j) does not depend on P

$$= \max_R \sum_j \int_K P_j D(P)\ S(R,j)$$

$$= \max_R \sum_j \bar{P}_j S(R,j) \tag{3}$$

where $\bar{P}$ is the average of P over K.

From the definition of a proper score, R* = $\bar{P}$. This result is quite general. It does not depend on the form of G(P,R), other than it be a proper score. It does not depend on the nature of the class K, other than that

G(P,R) D(P) be integrable over K. If K is the total simplex, $\sum_j P_j = 1$, and D(P) is the uniform distribution, then $R^* = \tilde{P}$, the uniform distribution on U.

We can call the prescription assert $\bar{P}$, the minimum loss rule for estimates with incomplete information. It is, in a way, a generalization of the principle of indifference. For decisional score rules, with matrix U_{ij}, it recommends selecting the action A_i such that $\sum_j \bar{P}_j U_{ij}$ is a maximum. For the case that K is the total simplex, $\sum_j P_j = 1$, it recommends, in effect, selecting the action with the highest row sum.

Given the assumption that upper level distributions are a reasonable approximation to uncertainty and that a uniform upper level distribution is a sufficient description of complete ignorance, the min loss rule appears rather inevitable. It has a number of attractive features in addition to those already mentioned. The average distribution $\bar{P}$ is relatively easy to compute. If a decision matrix has extreme values — "catastrophes" or "windfalls" — the rule neither ignores them, nor is obsessed with them. And, to anticipate the next section, in the "complete ignorance" form — a uniform upper level distribution on K — the rule is a hedge against bias. However, the assumption that upper level distributions are a sufficient approximation for incomplete information remains to be established.

3. Counterprediction

A phenomenon that gives some additional insights into the nature of uncertainty is counterprediction. Consider an individual who, if he asserts R, then you are better off to believe not-R. According to the theory of probability expressed in Chapter II, there is no such individual. Presumably, if anyone would be better off to believe not-R, then, in particular, the individual himself would be better off; so he also should, if his best estimate is R, believe not-R, which is some kind of contradiction.

Nevertheless, there is good evidence that the phenomenon of counter-prediction is fairly common. In the theory of psychological test construction, there is a concept called the difficulty of an item. The difficulty, for a given population, is defined as the probability that a member of that population will get the right answer, as diagrammed in Figure 28.[6] The interesting feature of this scale is that it covers the full range between 0 and 1. For those items with difficulty greater than the vertical line at d in Figure 28 where the probability drops below one half, a typical member of the population is a counter-predictor; you would be better off to reject his answer.

The second feature of interest for this scale is that there are examples of items with difficulty greater than d. And the classification of such items is well-defined in the sense that if they are scaled on a random sample of the population, then a different sample will exhibit the same proportion of individuals who get the correct answer. So far as I know there is no general theory for such questions in the sense that they can be identified without first trying them out on a sample of respondents.

The third interesting feature of the difficulty scale is that for questions with difficulty greater than d, the individual would do better if he reached in his pocket, drew out a coin, and flipped it to obtain his answer (assuming a true-false question). For such questions, a good fair coin is better than guessing. This is the hole in test theory that I referred to in the discussion of the rights-minus-wrongs score. If the individual could identify those questions for which he was a counterpredictor, he would do better by relying on a chance mechanism. He also would do better with the rights-minus-wrongs score by not answering such questions, since his expectation is negative. However, in order for the rights-minus-wrongs score to be

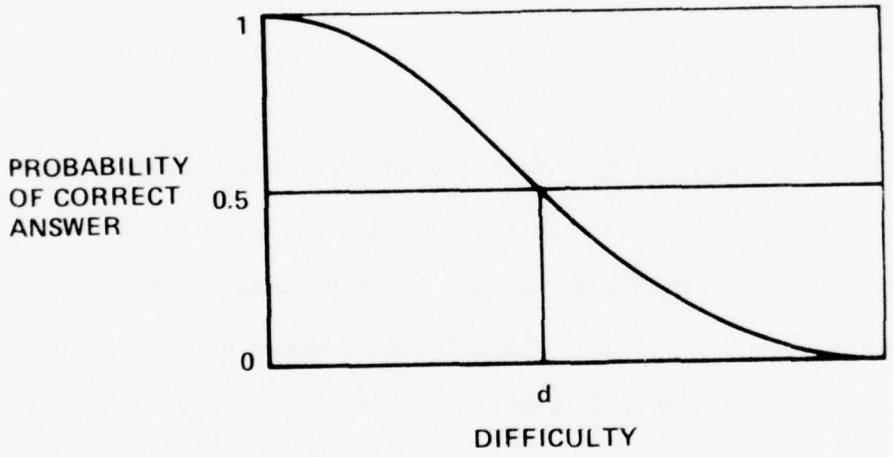


Figure 28. Scale of Difficulty for a Question

effective, the individual must be able to identify those questions for which he is a counterpredictor.

There have been a number of experiments exploring this issue, namely experiments concerned with the realism of subjects in making probability estimates. The data of Capen, Figure 9, Chapter II, is typical. The average quadratic score for the 5160 responses in this data is .55. The expected quadratic score for a complete ignorance response — i.e., for a response of .5 to every question — is .5. Thus, the individuals on the average did a little better than chance. However, for the roughly 40% of the responses where answers of .7, .8, and .9 were given, the average score was .46, distinctly worse than chance.

Figure 29 is a graph of the expected quadratic score, where R is the individual's response, and P is the objective probability. Only half of the graph is presented, since the other half is just a mirror image. .5 has been subtracted from the values to display the difference between the expected score and what would be expected from the response .5. Thus, along the horizontal line P = .5, and on the slanting line bounding the filled in area, the difference is 0. It is no surprise that the expected score is negative if the individual reports a probability greater than one half for events where the objective probability is less than one half. However, the stippled area shows a region where, despite the fact that both the report and the objective probability is greater than one half, the individual's score is still worse than chance. In the stippled area, the individual is a counterpredictor in the weak sense that he would achieve a higher score if he said, "I don't know"— i.e., responded with .5.

Figure 30 is a similar graph for the logarithmic scoring rule. Again, the stippled area is the region where both the report and the objective probability

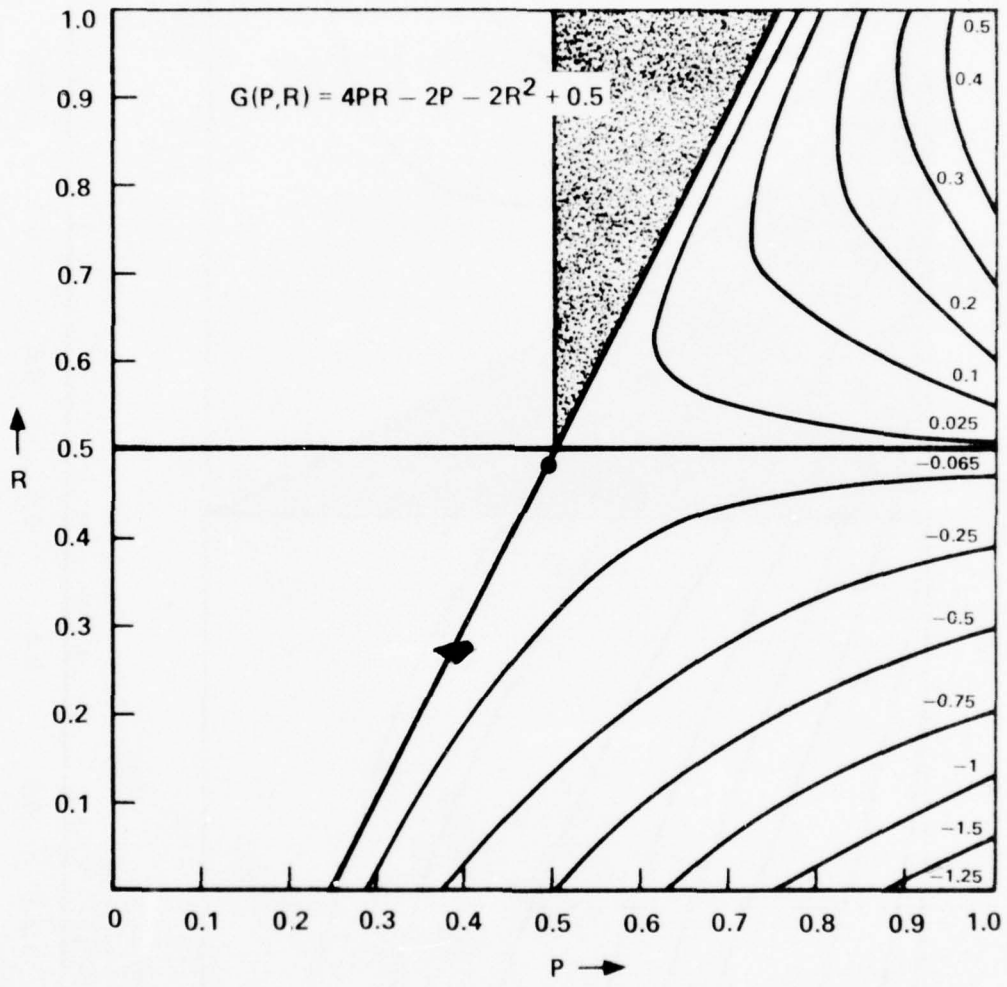


Figure 29. Expected Normalized Quadradic Score

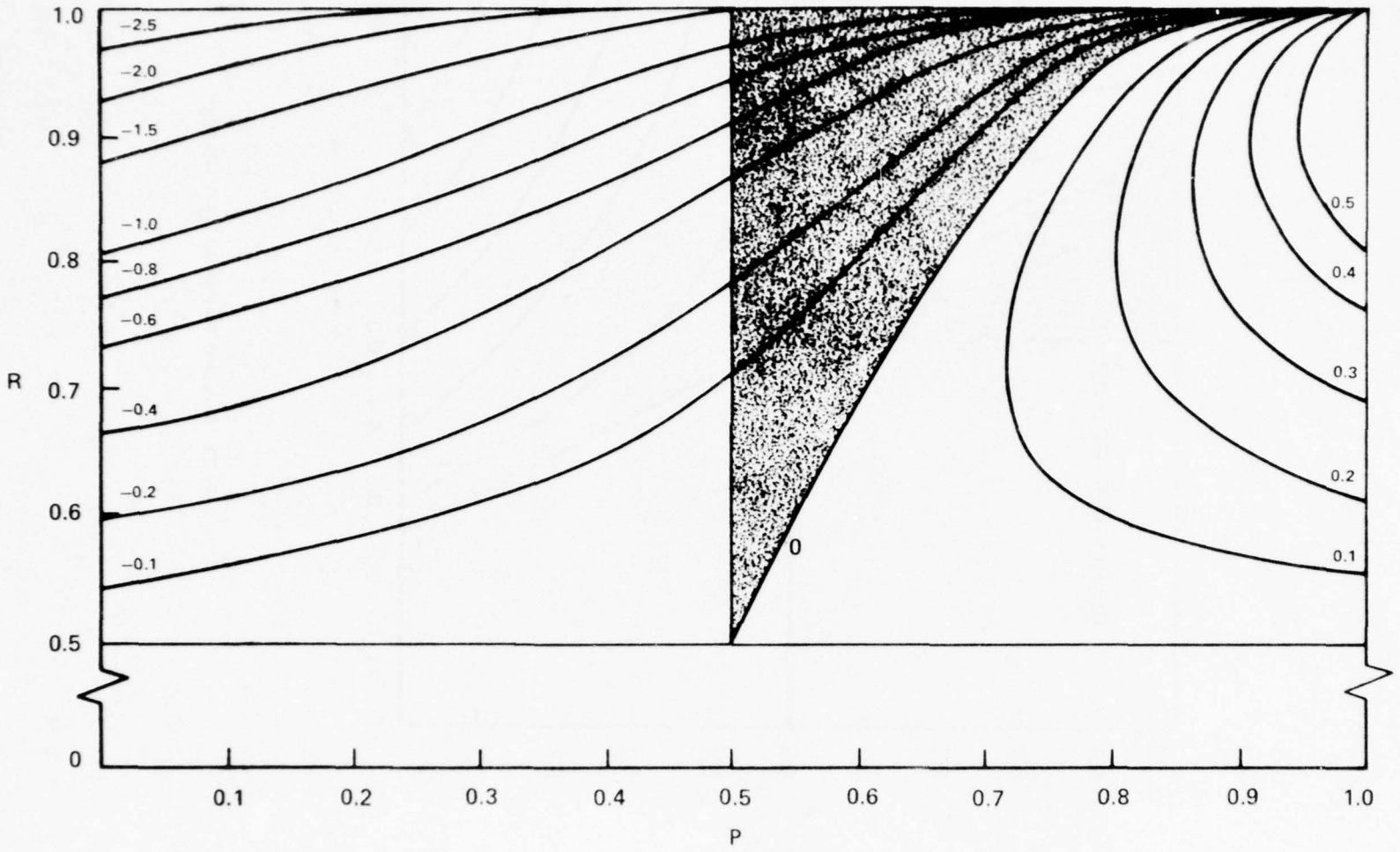


Figure 30. Expected Normalized Log Score
$G(P,R) = P \log R + (1 - P) \log (1 - R) - \log 0.5$

are greater than .5 and yet the expected score is less than the complete ignorance score. Although the general features of the two graphs are the same, they indicate that to some extent, the question whether a given response is counterpredictive depends on the score rule. For example, R = .75, P = .95 is counterpredictive for the logarithmic score rule, but not for the quadratic rule.

Figure 31 shows a similar graph for the scientific score rule. The picture is quite different than for the logarithmic and quadratic rules. There is no region where R and P are both greater than .5 and the expected score is less than for R = .5. The expected score is discontinuous at R = .5. Finally, Figure 32 shows the expected decisional score for the inset decision matrix. Here the expected payoff for any R less than 2/3 is precisely the same as for R = .5, hence the normalized score is zero for this region. Here the anomalous region where both P and R are greater than .5, but the normalized score is less than 0 (stippled) has a different locus than for the informational scores. The expected score is not defined for R = 2/3, since the decision-maker can select either action A_1 or action A_2. The numbers attached to R = 2/3 are a "safe score" where the two actions are mixed in a ratio 1/3, 2/3. The scientific score rule is essentially a decisional score rule with the matrix

	E	not-E
A_1	0	1
A_2	1	0

The structural differences between Figures 31 and 32 stem mainly from the fact that the matrix for the scientific score rule is symmetric, whereas the matrix for Figure 32 is not.

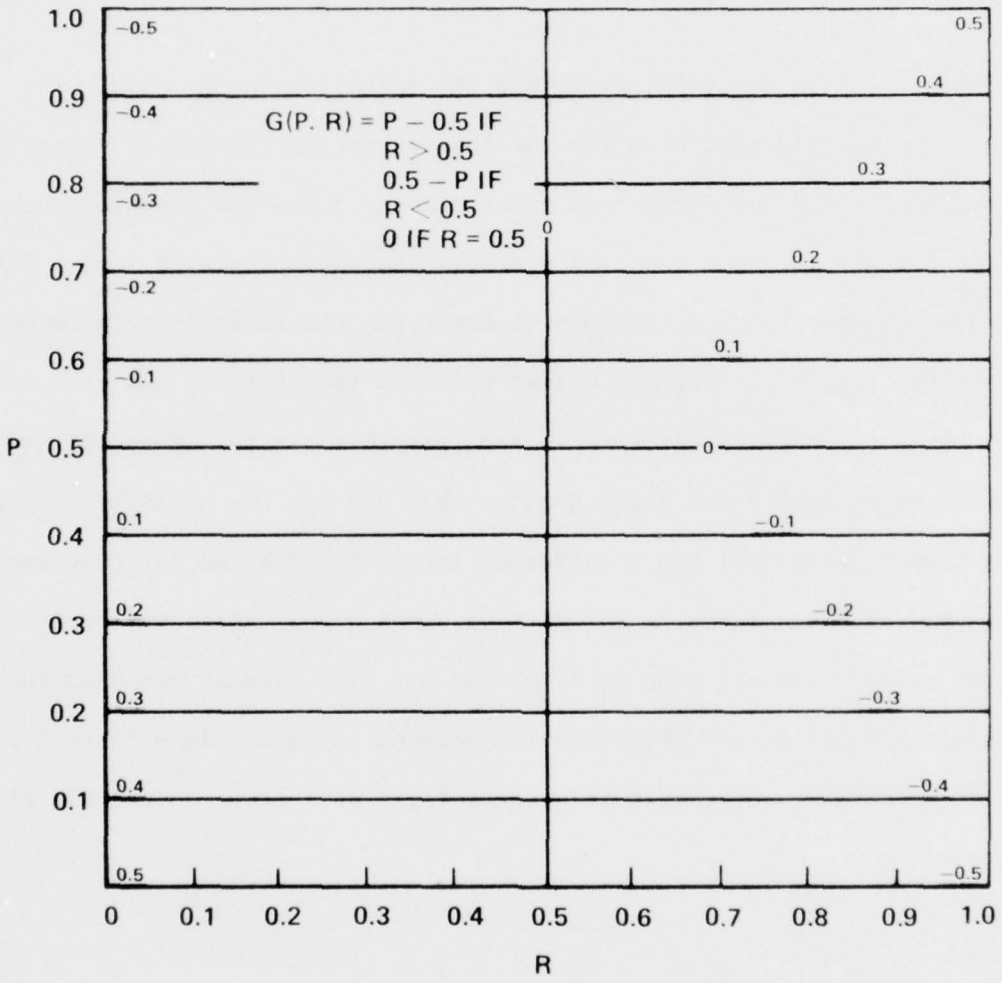


Figure 31. Expected Normalized Scientific Score

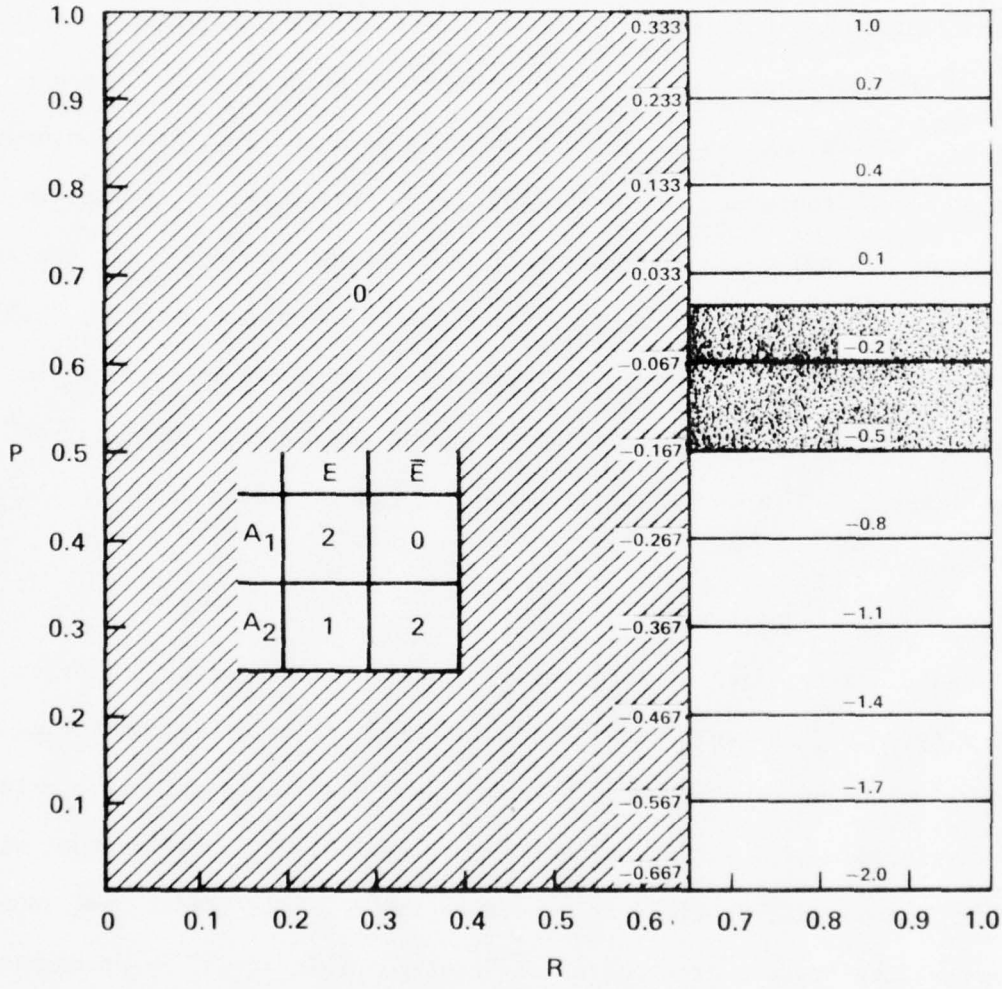


Figure 32. Expected Normalized Decisional Score

The phenomenon of counterprediction appears to offer a relatively sharp criterion for demarcating the area of ignorance within which nominal judgments are to be preferred to intuitive estimates. Almost tautologically, if a given question is one for which a given individual will make a counterpredictive estimate, then that individual would be advised to make a "weaker" estimate. There are several considerations which keep this point from being a pure tautology. As we have seen, whether or not a given estimate is counterpredictive depends on the score rule employed. More seriously, it depends on the form of the nominal estimate with which it is being compared. For the examples above, the comparison was with the "I don't know" estimate R = .5. Although this is a well-known and intuitively appealing criterion it assumes that a uniform distribution is the proper interpretation of "total uncertainty" or "total ignorance." A number of well-known paradoxes casts doubt on this interpretation.

4. Paradoxes of Uniformity

Traditionally, lack of information has been linked with the notion of uniformity, e.g., the well known "Bayesian" assumption of uniform prior probabilities in the absence of further knowledge. Most applications of rules such as the principle of insufficient reason, or the rule of indifference, wind up with uniform distributions on some event space. The traditional foundation of probability based on the notion of "equi-possible cases" -- probability is the ratio between the number of favorable cases to all possible cases, assuming they are all equally possible — has the same flavor.

Applied indiscriminately, these prescriptions can lead to simple paradoxes. Perhaps the simplest is the fact that distributions arrived at by these rules are not invariant over different ways of partititioning the event

space. If I am completely ignorant about the weather tomorrow, simple application of the principle of indifference would give probability 1/2 to rain, and probability 1/2 to not-rain. However, I am equally ignorant as to whether it will rain, snow, or be fair, in which case the probability of rain is 1/3. Almost any probability for rain less than 1/2 can be derived by selecting other possible partitions of the states of the weather.

An analogous puzzle arises when dealing with distributions on continuous quantities. If complete ignorance about a given quantity is taken to be a uniform distribution (over a given interval), then the distribution of the logarithm of the quantity (about which at least as much ignorance would be expected) is by no means uniform.

A similar difficulty applies to the min loss rule. In the complete ignorance case, $\bar{P}$ is a function of the selected partition of U.

The fact that various nominal rules for assigning probabilities are not invariant under changes in the partition of U indicates that concepts such as uniform distributions, or even uniform distributions of distributions, are not a complete explication of the notion of total ignorance or total uncertainty. Most attempts to define a logical measure for solidity are based on the assumption that there is some absolute partition — "atomic events" — which cannot be further subdivided. If there were such an irreducible partition, then complete ignorance could be defined by a uniform distribution on that partition. Unfortunately, there does not appear to be a meaningful criterion for specifying such atomic events.

5. Maximum Entropy and Minimum Score

A rule which, at first sight, appears to be similar to the min loss rule, has been receiving increasing attention by statisticians and information theorists. The rule goes under various names; perhaps the most popular is the

principle of maximum entropy. The rule can be formulated as: given an event space U, and certain a-priori information I concerning U, the most reasonable extension of I to a complete probability distribution D on U is that distribution which maximizes the entropy of D given I. The entropy of a discrete distribution P_j on a set of events E_j is just $-\sum_j P_j \log P_j$. The corresponding expression for a continuous distribution, D(x) is $-\int D(x) \log D(x)$. Entropy is a basic notion in the theory of communication as developed by Claude Shannon. As we saw in the discussion of probabilistic scores, this definition of entropy is equivalent to the negative of the expected logarithmic score. Hence, the principle of maximum entropy could be restated as, minimize the expected logarithmic score, given I.

The maximum entropy rule can be said to be reasonable in two ways: (a) It is a hedge against bias. If we define a counterpredictor as one who makes an estimate with a lower expected log score than the complete ignorance score, the maximum entropy rule will assure that the resulting estimate is not counterpredictive. (b) A maximum entropy estimate is the "weakest" assumption possible given I -- i.e., it adds the least possible information to I of any estimate. The second statement must be taken with some caution. The maximum entropy rule is just as sensitive to the chosen partition as any other rule. In addition, if the negative entropy is interpreted as a probabilistic score, then other score rules may generate different estimates. This point will be elaborated later in this section.

The principle of maximum entropy is often recommended in the context of generating a priori distributions for Bayesian inference, when the a priori probabilities are incompletely known.[7] Traditionally, one specialized form of this recommendation has been the ascription of a uniform distribution when the a priori probabilities are unknown. This rule has been controversial, but

difficult to either establish or kill. Some statisticians like A.J. Fisher have found the rule outrageous, and have rejected the use of a priori probabilities in statistical inference.[8] It is my impression that of late, along with a general spread of the subjective theory of probability, there has been an increasing willingness to employ both the notion of a priori probabilities, and the assignment of uniform distributions when the a priori probabilities are unknown.

One frequently employed justification of both subjective a priori probabilities, and uniform distributions with incomplete information, is the tacit assumption that use of these "devices" will be limited to the case of inference where some objective data is available, e.g., from an experiment or an observation. The function of the rule is to "get the inference started." It is an elementary exercise to demonstrate that as the amount of objective data increases, the influence of the a priori assumptions rapidly declines -- "the a-posteriori overwhelms the a-priori".[9] It is not clear whether any advocate of uniform priors would recommend uniform distributions for cases where no objective data is available.

In the case of complete lack of information, the principle of maximum entropy implies a uniform distribution for a discrete event space. Theorem 4, Chapter II, states that for any normal proper score rule, H(P) is a minimum for a uniform distribution; and since the logarithmic score rule is normal, the theorem applies to it in particular. A uniform distribution cannot be defined for a continuous quantity with an infinite range. On the infinite real line, any finite uniform function has an infinite integral. Thus, there is no way to impose the maximum entropy rule for continuous quantities without assuming some minimum amount of information, e.g., restricting the distribution

to a finite interval, or assuming that some property of the distribution such as the mean is known. For example, in a two alternative event space, if it is known that the probability of a given alternative is greater than or equal to a given bound, e.g., $P(E) \geq .7$, then the minimum expected score distribution (for normal scores) is just $P(E) = .7$.

As another example, consider the case of a three-alternative event space $U = (E_1, E_2, E_3)$. Assume a random variable X defined on U, such that $X = 3$ if E_1 occurs, $X = 1$ if E_2 occurs, and $X = 0$ if E_3 occurs. Assume that the information I is that the average of X is 1.5. We can ask, what is the minimum score distribution on U, given I? Call the minimum score distribution P*. There are three equations expressing the situation.

1. $\sum_i P(E_i) = 1$

2. $\sum_i P(E_i)X_i = 1.5$

3. $H(P^*) \leq H(P)$, for any P fulfilling 1 and 2.

To simplify the notation, let $p = P(E_1)$, $q = P(E_2)$, whence $P(E_3) = 1-p-q$. If we use the quadratic score for variety, we have

$$H(P) = p^2 + q^2 + (1-p-q)^2$$

From 1 and 2,

$$3p + q + 0(1-p-q) = 1.5$$

whence, $q = 1.5 - 3p$. Substituting in the expression for H(P)

$$H(P) = 14p^2 - 11p + 2.5$$

Taking the derivative with respect to p, and setting it equal to 0, we find $p = 11/28$, whence $q = 9/28$. The second derivative is positive, verifying that the point is a minimum. Figure 33 illustrates the computation. Equation 1 limits the possibilities to the triangular simplex; equation 2 further limits

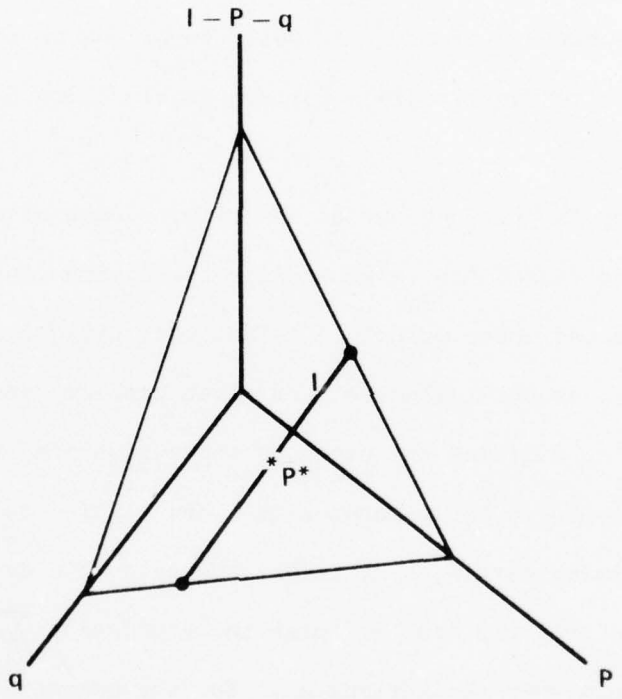


Figure 33. Information Set I for $\overline{X}$ = 1.5

the possibilities to the line labelled I, and equation 3 selects P* on this line.

Maximum entropy distributions have been computed for a variety of types of distributions and typical kinds of prior information. For example, the maximum entropy distribution for a quantity with an infinite range in each direction, with known mean and known standard deviation is just the familiar normal (Gaussian) distribution.[10] It would doubtless be instructive to compute minimum score distributions for comparable cases for other types of proper scores.

Since entropy is just one out of an unlimited set of score rules, it would appear to be reasonable to generalize the maximum entropy principle to a minimum expected score principle. That is: Given that a particular score rule $S(R,j)$ has been selected in a given problem, and given prior information I, then minimize the expected score, assuming information I.

This rule is particularly interesting when applied to decisional scores. If we have a decision matrix, U_{ij}, as we have seen, the decisional score rule is defined by the prescription, maximize the expression $\sum_j P_j U_{ij}$ where the maximization occurs over the actions A_i. For the moment assuming no information about the contingencies E_j, the minimum expected score rule is defined by

$$\min_R \max_i \sum_j R_j U_{ij} \tag{4}$$

Where the minimization on R is over the entire simplex $\sum_j R_j = 1$.

To take a simple example, consider the decision matrix

	E	$\bar{E}$
A	3	1
B	0	5

If we have no idea whatsoever concerning the probability of E, (4) recommends positing that the probability of E is the R that minimizes $\max_i \sum_j R_j U_{ij}$. The analysis is diagrammed in Fig. 34a. For $0 \leq R < 4/7$, B generates the higher expected utility. For $4/7 < R \leq 1$, A takes over. The rule recommends assuming that R = 4/7, the probability that produces the minimum of the maximum expectations. The expected return on this assumption is 3 x 4/7 + 1 x 3/7 = 15/7 = 0 x 4/7 + 5 x 3/7.

Fig. 34a illustrates a general point, namely, that for score rules which are not normal, the min score principle does not generate a uniform distribution, even for the case of no information.

The min score rule allows us to invoke some results of zero-sum game theory. If we think of R as the analogue of the mixed strategy of an opponent, then the min score rule is the analogue of the min max solution of two-person, zero-sum games. If we introduce the possibility of mixed actions on the part of the decision maker, then the basic theorem of two person zero-sum games can be used to establish the result that a mixed action for the decision maker exists which guarantees -- at least in terms of expectation -- the min score expected utility.

Let S_i designate a mixed action for the decision maker; that is $\sum_i S_i = 1$, and S_i is the probability with which action i is selected. Then

$$\min_R \max_i \sum_j R_j U_{ij} = \max_S \min_j \sum_i S_i U_{ij} \qquad (5)$$

(5) is a direct consequence of the fundamental theorem of game theory.[11] The result is illustrated in Fig. 34b. The abscissa is now S, the probability with which action A is selected. For $0 \leq S < 5/7$, the minimum expectation is from $\bar{E}$. The maximum of the minimum expectations occurs at S = 5/7. If the decision maker chooses A with frequency 5/7 and B with frequency 2/7, then

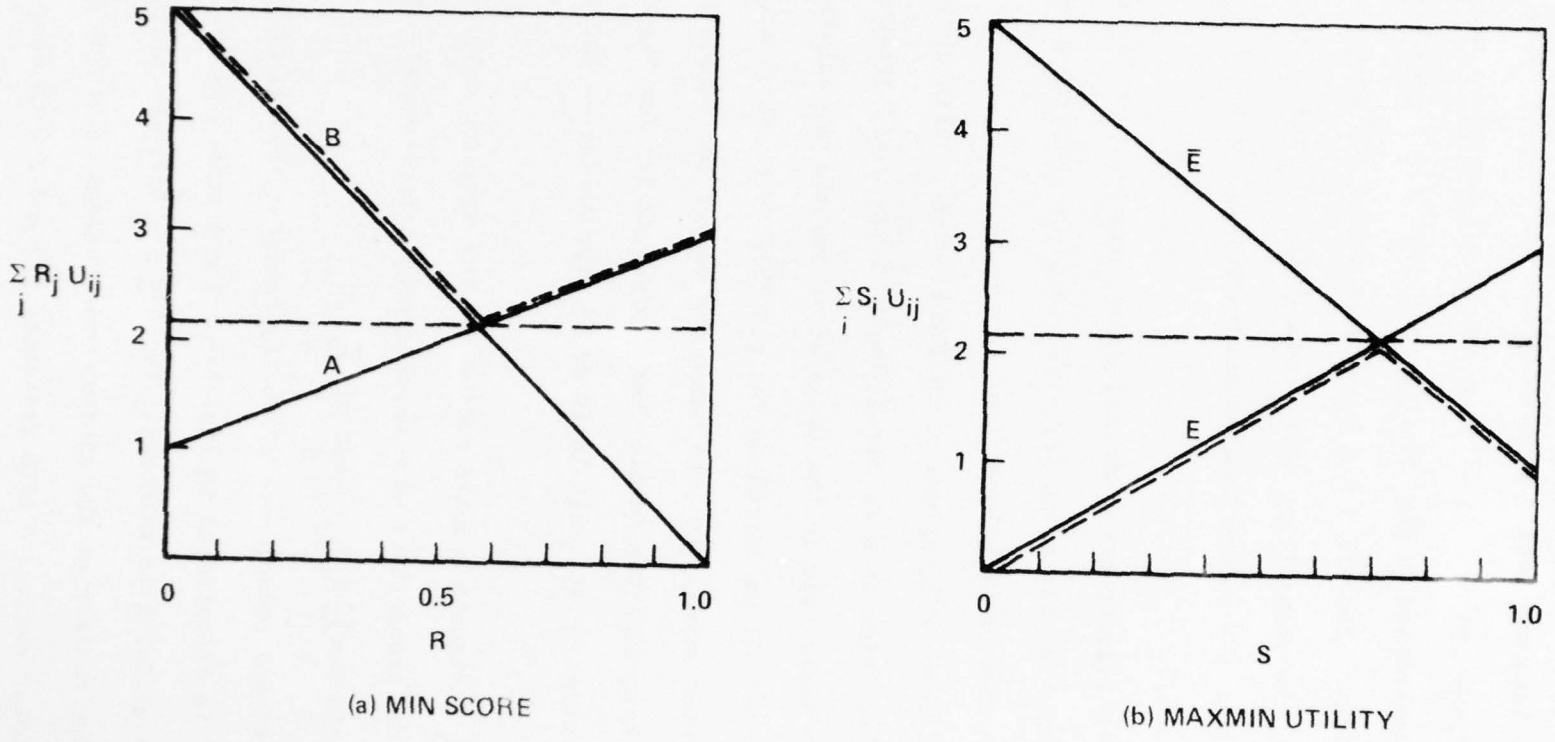


Figure 34. Comparison of Min Score and Maxmin Utility Analysis

his expected payoff is 15/7, whatever the probability of E. As we saw above, 15/7 is also the expected return on the min score assumption R = 4/7.

Through the route of the min expected score rule, we have arrived at a suggestion which is fairly old as such things go in the theory of decisions under uncertainty, namely the min max rule. Although the suggestion appeared shortly after the publication of von Neumann and Morgenstern's basic work on game theory, it has not received general acceptance as a "solution" to the problem of decisions with no information. Ironically, part of the reason is that it works so well for zero-sum game theory. The objection is that the rule does the equivalent of casting nature in the role of an inimical opponent, i.e., something which is "striving" to minimize the decision maker's rewards. Since all of physics implies that nature is neutral, the min max rule appears more pessimistic than necessary.

It is not clear that biology carries the same message. If the competitive interpretation of natural selection is accepted, there is some reason for assuming that the living part of nature is hostile. For example, plants secrete poisons or grow thorns to discourage the tendency of animals to eat them. But that would not be relevant to the estimation, e.g., of the probability of an earthquake.

Another potential objection which is a good deal stronger, is that if there are saddlepoints in the matrix, then the rule ignores potential outcomes which are highly favorable, and not ruled out by the information (or lack thereof) of the decisionmaker. For example, if the matrix is

	E	$\bar{E}$
A	1	2
B	0	x

then the rule says select action A, no matter what x is. If x is 10^6, with the pay off in dollars, a decisionmaker would probably be tempted to try B.

The min score rule addresses the same issue as the min loss rule derived in Section 2. They generate different estimates, even with informational score rules. Consider the two elementary examples discussed earlier. For the two alternative cases where P(E) is between .7 and 1, the min score rule gives R* = .7, whereas the min loss rule gives R* = .85, the average of .7 and 1. In the three alternative example, with the average of the random variable = 1.5, the min score rule gives R* = (11/28,9/28,8/28) whereas the min loss rule gives R* = (.375,.375,.25). In the case of a simple decision matrix with no constraints on the probabilities, the min loss rule gives the uninspired recommendation R* = $\tilde{P}$, the uniform distribution. Thus, for the matrix of Fig. 34 the min loss rule selects action B, rather than the mixed action selected by the min score rule.

The min score rule has the peculiarity that it recommends minimizing whatever reward function is operative in the decision. That feature has a non-intuitive flavor. Normally, in decision theory, rules are formulated so as to maximize some reward (or to minimize a loss). The min loss rule is more in the "mainstream." As we have seen, it is equivalent to maximizing the average expected score. If the min score rule is applied to the formulation of a prior distribution as the first step in a Bayesian inference, there is perhaps some justification for the minimization, in that it will "contaminate" the remainder of the inference as little as possible. However, it is not clear how one might justify the minimization if the resulting estimate is to be used directly in a decision.

Since the min loss rule is independent of the score rule employed, and because it involves maximizing the score, it would appear to be the preferred

rule for estimating "unknown" distributions for decisions. The following section examines some experimental data relevant to this conclusion.

6. Uncertainty and Choice Behavior

Perhaps the most persuasive evidence that there is a distinction between uncertainty and probability is a set of experiments which appear to show that individual choice behavior under uncertainty is incompatible with the postulates of subjective probability. Some of these have been triggered by the arguments of Ellsberg concerning the appropriateness of the sure-thing postulate for choices with incomplete information.[12] Others have derived from the work of Allais.[13]

To take up the Allais paradox first. It is my impression that this type of puzzle can be resolved within the ambit of personal probability theory without invoking any new notions. The Allais puzzle goes like this. Suppose I ask, which would you rather have, one million dollars for sure, or five million dollars with a probability of .8? Most individuals who have been asked this question have little difficulty deciding they would prefer the million for sure. (I don't think any billionaires have been among the respondents.) Now suppose I ask, which would you rather have, one million dollars with a probability of .1 or 5 million dollars with a probability of .08? Most individuals would prefer the 5 million dollars with probability .08. On the face of it, this violates the standard prescription that individuals should maximize their expected utility. There is no pair of utilities for one million and five million dollars which will rationalize these choices. To see this, let the utility of one million dollars be U(1) and the utility of five million dollars be U(5). We want $U(1) > .8U(5)$ and $.1U(1) < .08U(5)$. Multiplying both sides of the second inequality by 10 we get $U(1) < .8U(5)$.

However, the puzzle as usually presented overlooks a well known judgmental phenomenon, namely the influence of context. If someone offers me a choice between a million dollars for sure, and some probabilistic outcome, if I believe the individual, then in a quite reasonable sense, I have a million dollars. Whatever the status of my fortune before the offer, it has changed drastically. The actual choice is now between 0 (i.e., minus one million dollars) with the probability .2, and 5 million (i.e., plus four million dollars) with the probability .8. In short, the offer resets the zero of the decision situation to one million dollars. There is no puzzle in assuming that the loss of a million dollars would more than compensate for the .8 chance of getting four more million dollars.*

Suppose we assume a simple exponential utility function $U(x) = 1-e^{-x}$, where x is measured say in units of 1/4 million dollars. The extra four million is a gain of 16 units; the utility is essentially 1. The loss of a million is the loss of four units, and the utility is -.54. In this case, the individual would reject any option in which the probability of the five million is less than .98. Interestingly, with this utility function, the outcome is highly sensitive to the units. If the individual was thinking in terms of units of say $1000, the utility of a loss of a million dollars would be, for all practical purposes, negatively infinite.

The principle invoked in the preceding could be called the context dependent zero; that is, the zero point of the utility scale is dependent on the situation, including the options presently available.

---

*The pioneering paper of Friedman and Savage[14] on nonlinear utility of money as an explanation of various kinds of behavior which appear anomolous if the value of money is considered linear would then explain this, as well as other puzzles.

The same transformation of zero is not appropriate for the second choice situation, since now there is no assurance that the million is available.

The context dependent zero analysis does not appear to be adequate for the Ellsberg case. The Ellsberg type paradox involves a choice between rewards with known, and with unknown probabilities. A typical situation (taken from McKrimmon[15]) is that of an urn drawing, with three kinds of balls -- say red, blue, and green -- in which the proportion of red balls is specified, say 1/3, but the relative proportion of blue and green is not specified. Two different choices are proposed, e.g. those in Table 1.

Table 1

	1/3	2/3	
	Red	Blue	Green
A	$1000	0	0
B	0	$1000	0
A'	$1000	0	$1000
B'	0	$1000	$1000

Asked to choose between A and B, most subjects will choose A. Asked to choose between A' and B', most subjects will choose B'. As in the untransformed Allais example there is no choice of subjective probabilities for Blue and Green, and utilities for 0 and $1000 that will account for these choices. Roughly speaking, the subjects appear to prefer choices in which the rewards are more "surely known" even if probabilistic. The issue can be further sharpened by noting that the rewards if Green is drawn are identical for A and B and identical for A' and B'. Thus, whatever could make A preferable to B should also make A' preferrable to B'. In short, the observed choices violate the sure-thing principle, P6, Chapter II.

Situations with choice tasks like those in Table 1 have been tried in several experimental series.[16] The results from these experiments are all similar, the majority of subjects choose A and B'.

Although it is possible to find a kind of resulution for the Ellsberg paradox using the context dependent zero notion -- in this case assuming that the more certain of the options generates a new "expected zero" -- the result appears labored.

If the min score rule is applied to Table 1, since the probability of Red is fixed, the minimization is carried out for $0 \leq P(B) \leq 2/3$, where P(B) is the probability of blue. We can diagram the two choice situations as in Figs. 35a and b, where the abscissa is the (unknown) probability P(B) of blue. In Fig. 35a, if the individual selects A, then his expectation is a constant 1/3 x \$1000. If he selects B, then his expected outcome could lie anywhere along the line labeled B. If $P(B) < 1/3$, then his expectation is maximized if he chooses B. For $P(B) > 1/3$, his expectation is maximized if he chooses A. The minimum of the maximum outcomes occurs where the two lines cross at $P(B) = 1/3$.

However, choosing A assures obtaining the min max, since the expectation of A is a constant. Put in other terms, the pair (A,G) is a saddlepoint of the matrix

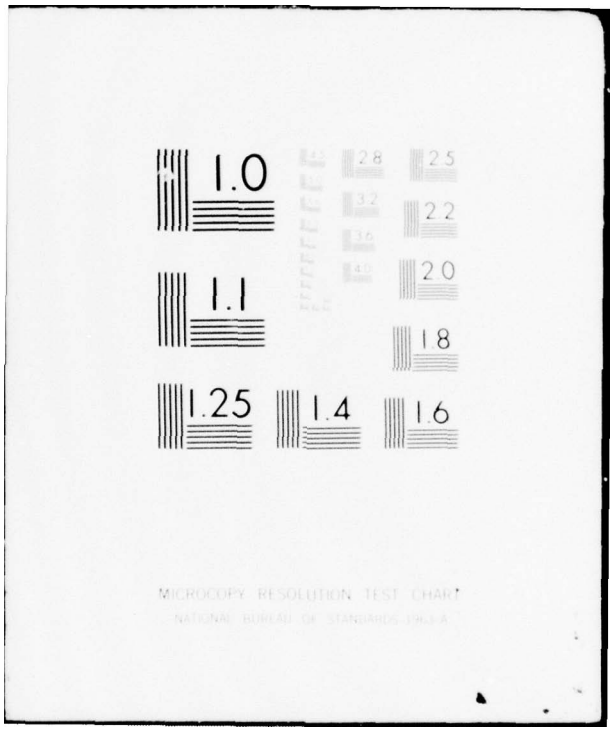
	B	G
A	1/3 \$1000	1/3 \$1000
B	\$1000	0

which is obtained by omitting the constant probability column Red from Table 1.

For the second option, Fig. 35b, it is B' which assures the min max outcome. The basic assymmetry between the two cases can thus be expressed by the statement that in both cases there is a pure (unmixed) action which

AD-A042 852 CALIFORNIA UNIV LOS ANGELES SCHOOL OF ENGINEERING A--ETC F/G 5/10
GROUP DECISION THEORY.(U)
JUL 77 N C DALKEY
UNCLASSIFIED UCLA-ENG-7749 N00014-69-A-0200-4056
NL

3 OF 4
AD A042852


1.0
1.1
1.25
1.4
1.6
2.8
2.5
3.2
2.2
3.6
2.0
4.0
1.8
MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS 1963-A

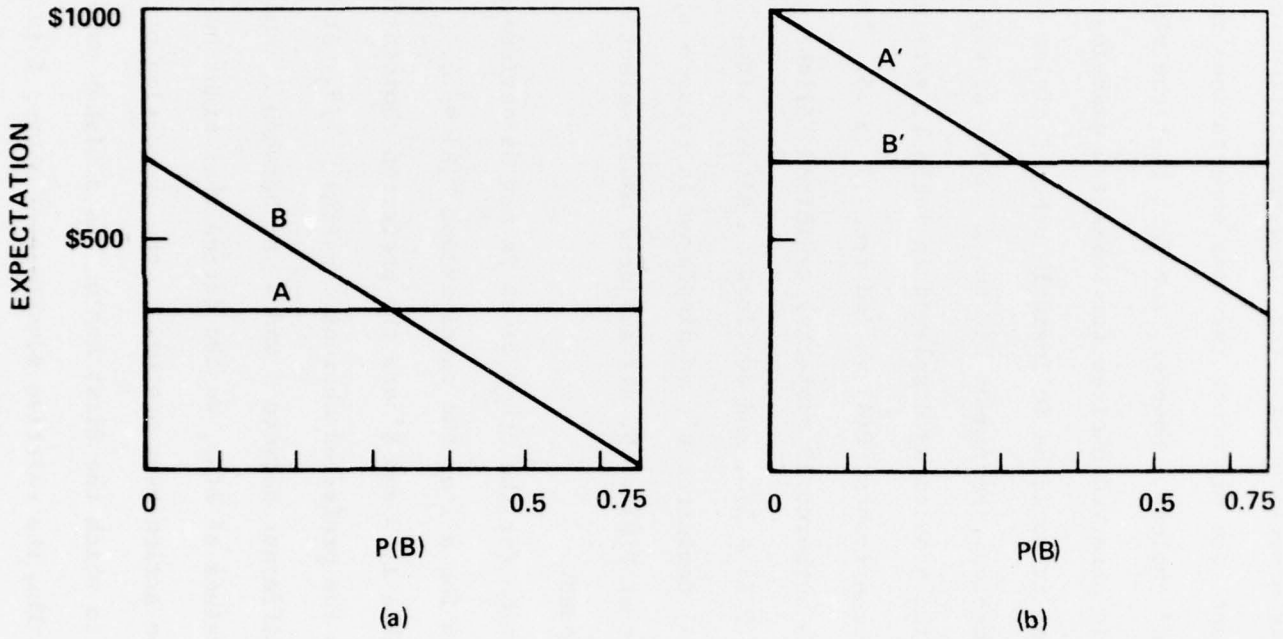


Figure 35. Decision with Incomplete Information P(R) = 1/3

guarantees the min max outcome, but in the first choice the pure action A is associated with the fixed probability for Red, whereas in the second choice, the pure action is associated with the fixed probability for Blue or Green.

To this extent, the min score (min max expectation) rule is in accord with the subjects' choices. However, the fact that the min score rule fits the experimentally observed choices for these two examples is by no means a demonstration that it would be followed in other decision problems with incomplete information. McKrimmon has run a number of experiments with the same basic decision problem as displayed in Table 1, varying the probability of Red. In his experiments, P(R) varied from .2 to .5. The extent to which his groups (which numbered 19 subjects) exhibited "Ellsberg type decisions" was a maximum at P(R) = 1/3, and declined on either side. At P(R) = .5, A dominates B and A' dominates B', as displayed in Figs. 36 a,b. Thus, we would expect that at P(R) = .5, all subjects would select A and A', which is what McKrimmon found.

The explanation for the other cases is not so obvious. A and B' are the min max solutions for all of the cases except P(R) = .5. If the min loss rule is applied to Table 1, B and B' are the preferred choices for P(R) > 1/3, and A and A' are the preferred choices for P(R) < 1/3. At P(R) = 1/3, the min loss rule is indifferent between A and B and between A' and B'.

For other values of P(R), we can define what might be called the "relative advantage" of one action over another as the proportion of the undetermined interval 1-P(R) in which the first action has a higher expected value than the second. In Fig. 35a, the relative advantage of A over B is 1/2. In Figs. 37a and b, the choices are diagrammed for the case P(R) = .2. The relative advantage of A over B is the ratio of the solid part of the line labelled A to the total line, in this case 1/4. The relative advantage of B' over A' is 3/4.

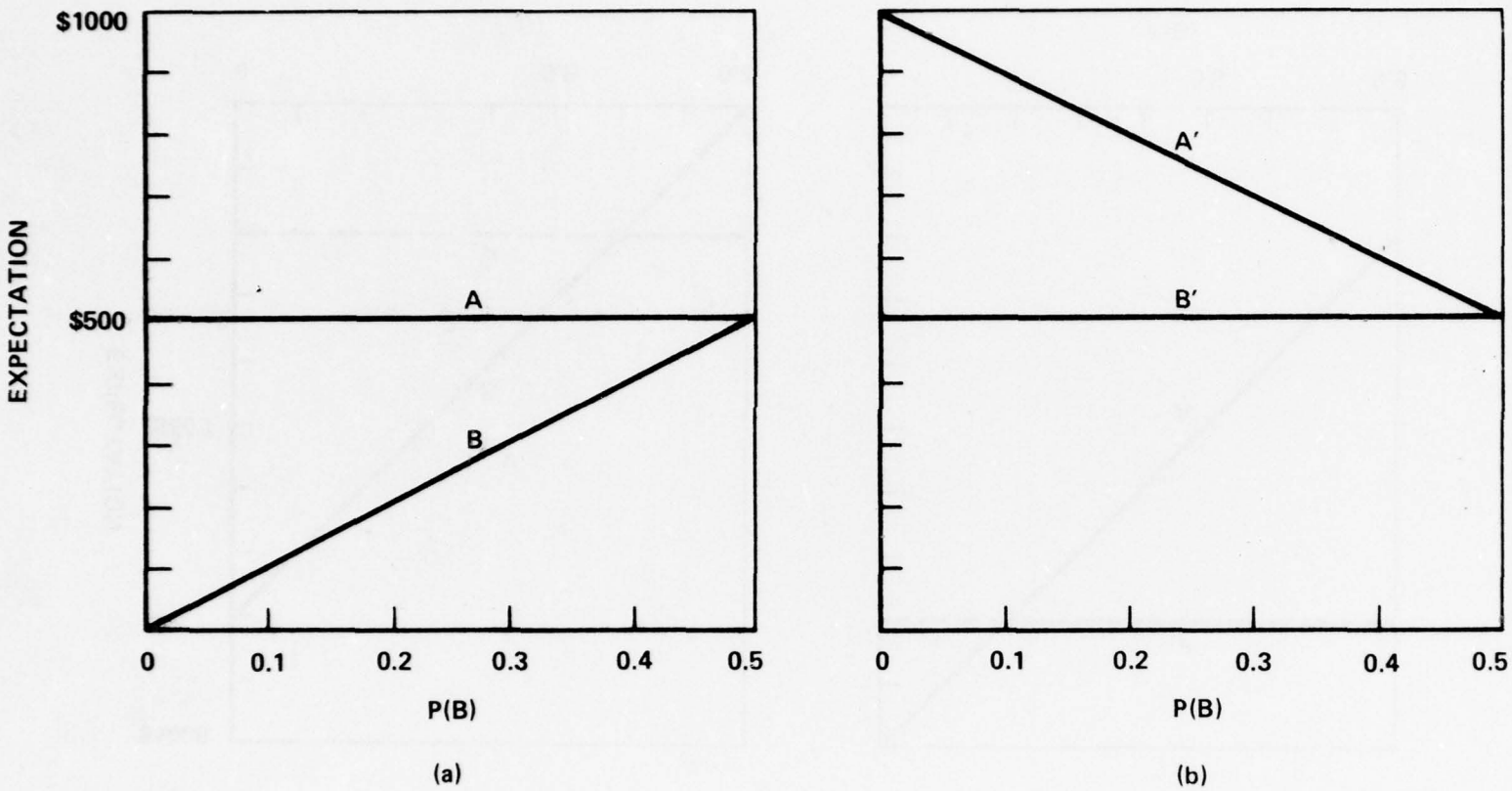


Figure 36. Decision with Incomplete Information, P(R) = 0.5

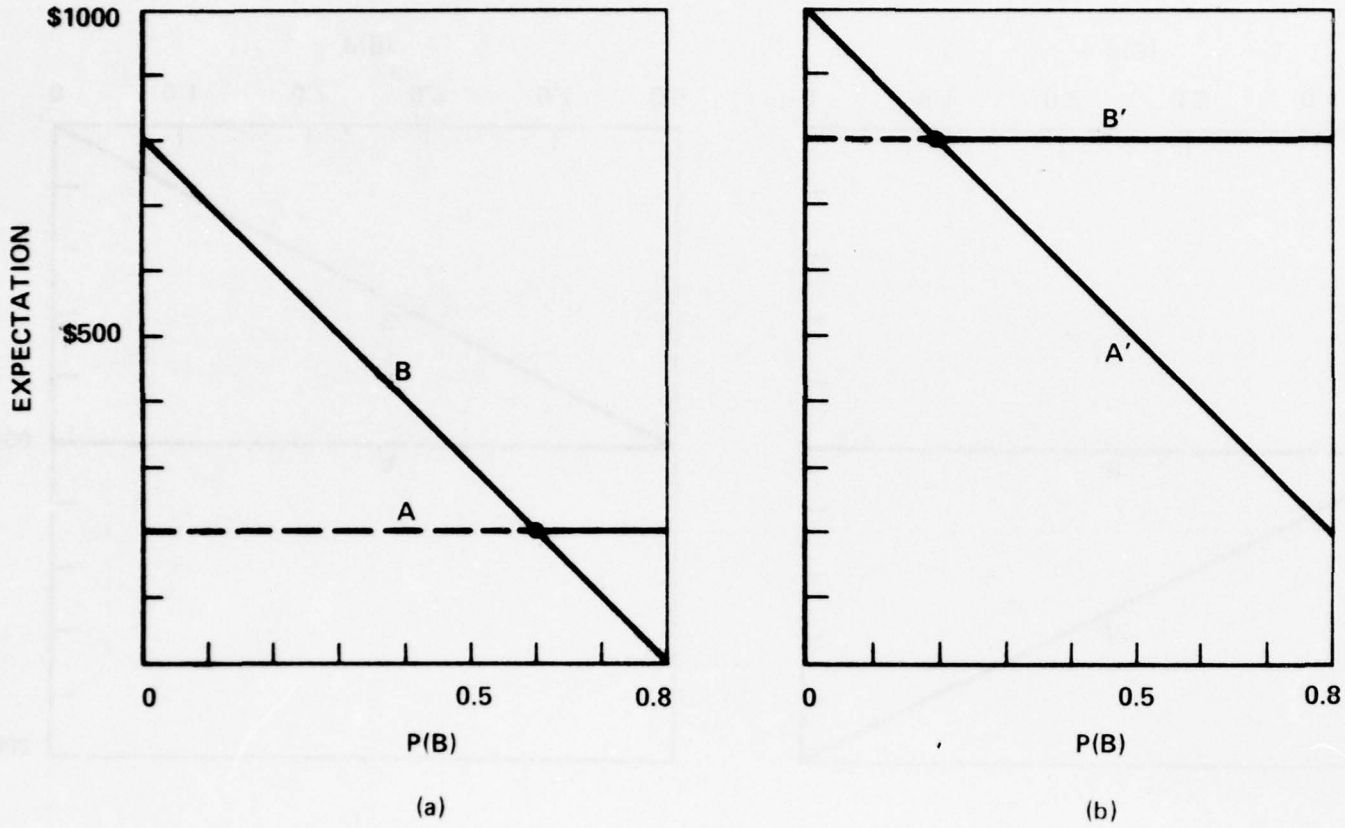


Figure 37. Relative Advantage for P(R) = 0.2

In Fig. 38 the McKrimmon data is plotted with the ordinate showing the proportion of times a given action was selected, and the abscissa showing the relative advantage of that action. The two sets of points are for A over B and B' over A'. A is the min max action in the first choice, and B' is the min max action in the second choice, except for the case P(R) = .5. The point at the origin is the result for P(R) = .5.

The proportion selecting a given action is nicely monotonic in the relative advantage, and the behavior of the two curves is surprisingly similar (note the two identical points), despite the fact that the two are in reverse order with respect to the value of P(R) involved. For example, the two identical points at (.33,.42) are for A over B at P(R) = .25 and for B' over A' at P(R) = .4.

The effect of the min-max property of A and B' shows up in the fact that the proportion selecting these two mounts close to 1 when the relative advantage is only .5. In the linear region between .2 and .5 on relative advantage, the data fit the hypothesis that the subjects are choosing between the min score and the min loss "solutions" in a ratio roughly about 2.7 times the relative advantage.

Without additional experimentation designed specifically to test the hypothesis inherent in Fig. 38, it would be hasty to call it more than suggestive. What does appear to be firm, from the published experimental data, is that the upper level distribution model, i.e. the min loss rule, does not completely fit observed choices under incomplete information, and neither does the min score rule.

7. Nominal Estimates with Factor Models

One of the drawbacks to general prescriptions like the min loss rule is the fact that they have a mainly negative import. The chief advantage appears

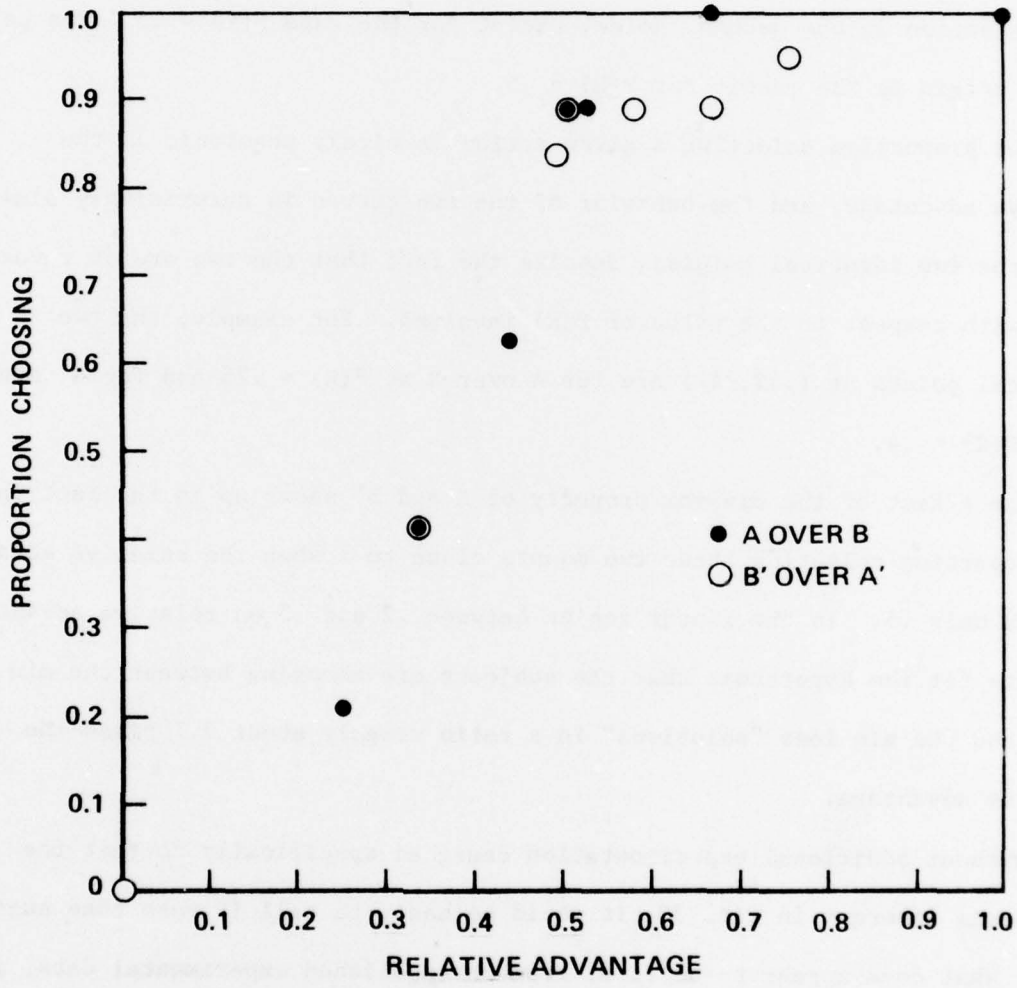


Figure 38. Proportion Choosing Given Action as Function of Relative Advantage

to be guarding against bias. Put another way, although such rules are "safe," they also appear to be "weak;" the naive expectation is that they would generate low returns if applied in practice. Part of the misapprehension here is an illusion concerning the excellence of everyday decisions. Although there has not been a general survey of the quality of decisions in industry, government and private affairs, the evidence is mounting that many decisions in everyday life would be dramatically improved if "weak" nominal estimates were substituted for the guesses and hunches which presently guide the decisions.

Some of the positive advantages of nominal estimation rules can be seen more clearly in the context of factor models. In the factor model approach to estimation, if we restrict attention to the elementary case in which the values of the factors are given, and the individual makes his estimate knowing these values, then the only task is to assign subjective weights to the factors and "perform the arithmetic." In practice, it seems unlikely that estimates are arrived at in this formal way. However, for many estimation tasks, it appears to be a reasonable approximation.

The figure of merit most often used for factor models is correlation. In the elementary case we are now considering, there are a set of objects $X = \{x,y,z....\}$ where each object is a vector $(x_1,...x_n)$ and there is some function $t(x)$ which defines the quantity to be estimated. The x_i are the factors. An individual estimates t by determining a set of weights a_i for the factors, and asserts $r(x) = \sum_i a_i x_i$.* We can compute the correlation between r and t as

$$\rho(r,t) = 1/m\sum\left(\sum_i a_i x_i - \overline{\sum_i a_i x_i}\right)(t - \bar{t})/s_r s_t \qquad (6)$$

*Normally, it would be necessary to add a constant, but since correlation is invariant under a linear transformation on the quantities being correlated, the constant will be omitted for simplicity.

In this case, the bar over an expression indicates the mean of that expression. m is the total number of cases. Simplifying (6), we obtain

$$\rho(r,t) = 1/s_r \sum_i a_i s_i \rho\ (x_i,t) \tag{7}$$

If we assume the variables have been normalized, so that $s_i = 1$, and furthermore assume that the x_i are uncorrelated, i.e., $\rho(x_i,x_j) = 0$ for all i and j, then (7) reduces to

$$\rho(r,t) = \sum_i \frac{a_i}{\sqrt{\sum_i a_i^2}}\ \rho(x_i,t) \tag{8}$$

For the case of uniform weights, $a_i = 1$ for all i, (8) becomes

$$\rho(r,t) = \frac{1}{\sqrt{n}} \sum_i \rho(x_i,t) \tag{9}$$

Thus, for this elementary case, an estimate based on uniform weights is better than the average correlation between the variables and the true answer by a factor of $\sqrt{n}$. If all the individual correlations are positive, then (9) indicates that the uniform weight estimate will improve with each additional variable. As a rough illustration, suppose each $\rho(x_i,t)$ is about .2, and $n = 5$, then $\rho(r,t) = .45$.

The topic can be explored a little further if we assume that $t(x) = \sum_i b_i x_i$ -- i.e., the function to be estimated is also linear. We now have (under the assumption that the variables are normalized and uncorrelated).

$$\rho(r,t) = \frac{\sum a_i b_i}{\sqrt{\sum a_i^2}\sqrt{\sum b_i^2}} \tag{10}$$

Since the b's are unknown, a reasonable requirement on the a's is that they maximize the expectation of $\rho(r,t)$ over the possible values of the b's. Thus we would like to find the a which generates

$$\max_a \int_B \frac{\sum_i a_i b_i}{\sqrt{\sum_i a_i^2}\sqrt{\sum_i b_i^2}} D(b) \tag{11}$$

where B is the set of possible b's and D(b) is a probability distribution on B. Since the correlation is invariant under linear transformations, there is no loss of generality in assuming that the b's have been normalized so that $\sum_i b_i = 1$. In the extreme case that B includes all possible sets of coefficients, i.e., B is the simplex $\sum_i b_i = 1$, and D(b) is taken to be uniform, we have

$$\max_a \frac{\sum a_i}{\sqrt{\sum_i a_i^2}} \int_B \frac{b_i}{\sqrt{\sum_i b_i^2}} \tag{12}$$

The expression $b_i/\sqrt{\sum_i b_i^2}$ is symmetric in i, as is the simplex $\sum_i b_i = 1$; thus, the integrals will be the same for all i. Hence

$$\int_B \frac{b_i}{\sqrt{\sum_i b_i^2}} = k\ 1/n \tag{13}$$

Whence, (12) becomes

$$\max_a k\ 1/n \frac{\sum a_i}{\sqrt{\sum_i a_i^2}} \tag{14}$$

Assuming the a's are also normalized, $a_i/\sqrt{\sum_i a_i^2}$ is just the spherical scoring rule, so (14) is formally equivalent to (6) Chapter III with $P_j = 1/n$. Hence, the maximum over a occurs when $a_i = 1/n$ for all i.

Thus, for the case of "complete ignorance" (interpreted here as a uniform distribution over the set of all possible b's) the maximum expected correlation is obtained when the estimated weights are uniform.

Having determined that the best estimate given "complete ignorance" is the set of uniform weights, there remains the question of just how good (or bad) it is. One way to measure this is to compute the expected correlation over the set of possible coefficients. This requires evaluating the integral

$$\bar{\rho}(r,t) = 1/\sqrt{n} \int_B \frac{1}{\sqrt{\sum b_i^2}} \tag{15}$$

For the case of two variables, the integral is quite straightforward. It is

$$1/\sqrt{2} \int_0^1 \frac{dx}{x^2 + (1-x)^2}$$

which yields

$$\bar{\rho}(r,t) = \tfrac{1}{2} \log \frac{2+1}{2-1} = .881$$

For three or more variables, the integral becomes somewhat more complex, taking the form

$$\bar{\rho}(r,t) = 1/\sqrt{n} \int_0^1 \int_0^{1-x} \cdots \int_0^{1-\sum_1^{n-1}} \frac{x_i \, dx_1 dx_2, \ldots, dx_{n-1}}{\sqrt{\sum_i x_i^2}} \tag{16}$$

A mixed analytic and numerical solution to (16) for three variables yields $\bar{\rho}(r,t) = .834$.

For three or more variables, there is a certain embarrassment in assuming the "complete ignorance" case, i.e., assuming B = total simplex. The assumption includes among the possible cases those in which all but one of the b's are zero. I'm inclined to think that if the problem is so poorly defined that it is necessary to take into account the possibility that t(x) is a function of only one of the variables, the model is not ready to be used in a

serious decision. A certain amount of arbitrariness enters in attempting to formulate a suitable restriction "a priori." A convenient restriction is to the case that no more than one of the b's is allowed to be zero. This restriction can be expressed by setting B equal to the inscribed hypersphere in the simplex. For the case of three variables, this restriction is illustrated in Fig. 39. The total set of possible b's consists of the triangular region $\sum_i b_i = 1$. The inscribed circle cuts off the extreme cases. For this B, the problem has spherical symmetry, and we have

$$\bar{\rho}(r,t) = \frac{1}{\sqrt{n}} \int_0^{\frac{1}{\sqrt{n(n+1)}}} \frac{S(r)dr}{V(S)\sqrt{r^2 + 1/n}} \tag{17}$$

where V(S) is the volume of the inscribed hypersphere with radius $\frac{1}{\sqrt{n(n+1)}}$ S(r) is the surface of the hypersphere with radius r, and n is the number of variables.

For n = 3, (17) gives $\bar{\rho}(r,t) = .90$, and for n = 4, $\bar{\rho}(r,t) = .915$. These two additional cases were as far as my patience in evaluating integrals lasted. The fact that the average ρ increases with n reflects in part the fact that the ratio of the volume of the inscribed hypersphere to the volume of the total simplex goes down with increasing n — more extreme cases are eliminated.

Some additional insight into the effectiveness of the equal weights approximation can be obtained by noting that $\sqrt{\sum_i b_i^2}$ is a measure of the dispersion of the b's. Setting $a_i = 1$ in (10), we obtain

$$\rho(r,t) = 1/\sqrt{n}\, \frac{\sum_i b_i}{\sqrt{\sum_i b_i^2}} \tag{18}$$

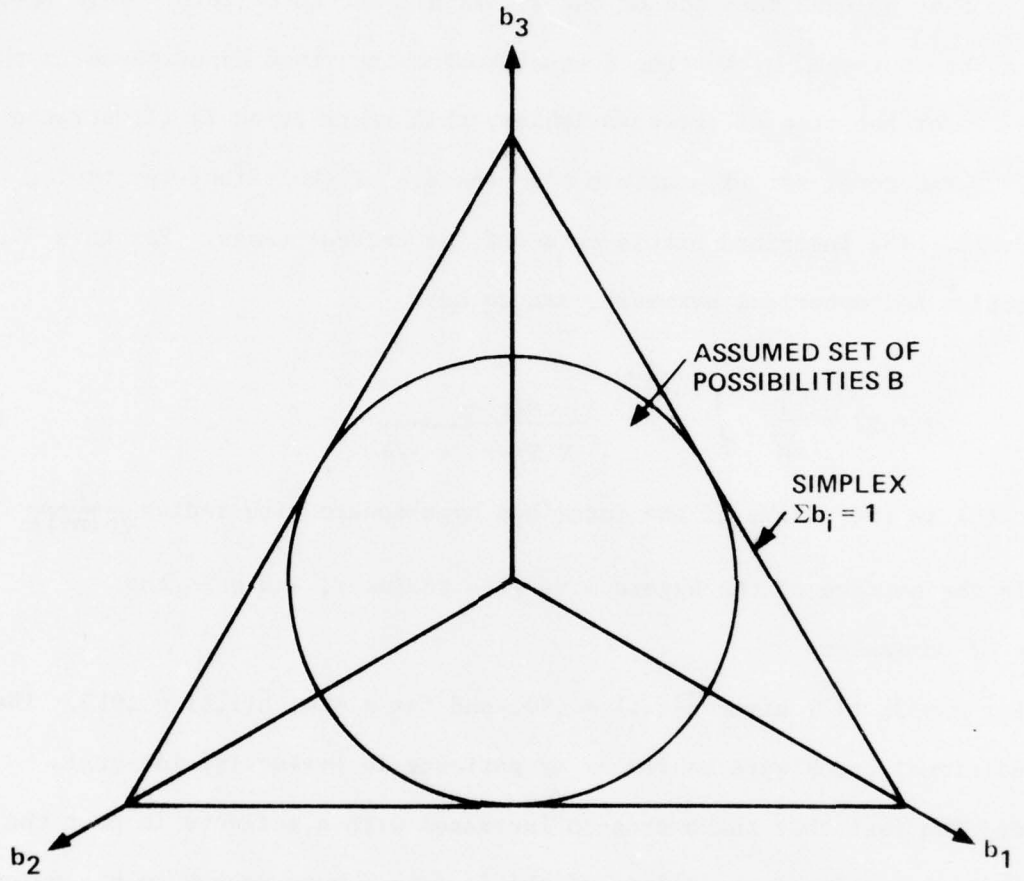


Figure 39. Constrained Possibility Set B for Computing Average Correlation

which is the correlation of the estimate with uniform weights with the true function with unknown weights.

For example, if $t = x + 2y$, $\rho(r,t)$ for uniform weights is .95. If $t = x - 2y + 3z$, $\rho(r,t) = .93$.

Rearranging terms in (18) we can write

$$\rho(r,t) = \frac{1}{\sqrt{\frac{s_b^2}{\bar{b}^2} + 1}} \tag{19}$$

where s_b^2 is the variance of the true weights, and $\bar{b}$ the average.[17]

Thus, for uncorrelated variables, the correlation between an equal-weight approximation $r(x)$ and any linear function $t(x)$ is determined by the relative variation of the coefficients of t, where the relative variation is defined as $s_b/\bar{b}$. If the distribution of the b's is "well behaved" then the correlation will be high. The worst case, of course, is the degenerate one where all but one of the coefficients are zero. In this case, $\rho(r,t) = 1/\sqrt{n}$. However, this worst case hardly appears to be of practical interest. If there is a non-trivial likelihood that t is a function of only one of the variables, then, as remarked earlier, introducing the approximation as a serious basis for a decision is at best "premature".

One useful reference case is that of a uniform distribution of weights on some interval. Whether a uniform distribution can be taken as a characterization of "ignorance" in the case of coefficients for linear models is not as obvious as it seems in the case of estimating a single quantity. For the coefficients, we are estimating a <u>set</u> of quantities. However, it is clear that a uniform distribution is one form of "low information" assumption about the coefficients. For any uniform distribution on a positive interval (u,v) $\bar{b} = 1/2(v+u)$ and $s_b^2 = 1/12(v-u)^2$. Thus

$$\frac{s_b^2}{\bar{b}^2} = 1/3 \frac{(v-u)^2}{(v+u)^2} \leq 1/3 \qquad (20)$$

For $s_b^2/\bar{b}^2 = 1/3$, from (19), $\rho(r,t) = .866$. (20) is independent of the size of the interval (u,v), and roughly independent of the number of coefficients--there have to be enough so that a uniform distribution can be approximated. For example, if the b's consist of a string of successive integers $1,2,\ldots,n$, then (20) holds approximately for any $n > 2$.

For any distribution of coefficients more favorable than a uniform distribution--e.g., if the coefficients tend to cluster about some intermediate value with only a few extreme values--$\rho(r,t)$ will be greater than .866.

Although the assumption of uniform weights is weak in the sense that it can be derived from the "complete ignorance" assumption of a uniform distribution on all possible sets of weights, the numerical examples show that it is not necessarily a poor assumption.

This conclusion is urged on empirical grounds by Robin Dawes.[18] He has compared the equal weight approximation to intuitive judgments of trained personnel in the estimation of grade point averages, and determination of degree of mental illness from personality tests. Over a number of extensive studies of these tasks, the equal weight approximation gave significantly higher correlations than the original estimates.

8. Theory of Information-Control

This section is somewhat of an aside with respect to the main theme of this chapter. The primary reason for including it is to indicate in a more fundamental way the interrelationship between the notion of certainty (or solidity) and the role of decision rules. In addition, the formal decision theory outlined below is a good deal more general than the theory embodied in

the probabilistic theory of estimation of Sec. 5, Chapter II. It offers a framework in which a wide variety of types of information can be incorporated in decision rules. However, most of the potentialities remain to be explored.

Any decision situation involves aspects which are clearly under the control of the decision maker, and other aspects which are clearly not under his control such as probabilistic events or actions which are under the control of other decision makers. In between are aspects which are not clearly one or the other. Bayesian decision theory assumes that there are only two classes, actions (controlled) and events (uncontrolled). In the following these distinctions will be blurred. It will be assumed that there is a set of aspects which are clearly under the control of the decision maker (which could be called capabilities) and the remaining aspects whose status is not well-defined initially, which might be called contingencies.

Control involves two properties, (a) the decision maker can implement a given option, and (b) he can select any alternative out of a set of options. The second, which might be called the "free-will" assumption, is the crucial one for the following theory. A decision model is, of course, not reality but a representation of reality as viewed by the decision maker. He can be mistaken about his capabilities. However, the ability to select one of the listed options is basic -- otherwise the whole exercise is a dream.

This assumption has been brought under fire for the case of high uncertainty. With insufficient information, it has been contended, an individual can find himself in a state which is at least as bad as that of Buridan's ass -- he simply can't make up his mind. Perhaps a better expression might be he can't make up his feelings. I'm not aware of a clear demonstration of this potential phenomenon in laboratory studies; but vacillation is a familiar concept in literary psychology, and "decisiveness" is a well-known trait

ascribed to successful managers. What has not been documented is the relationship between these traits and the information-control characteristics of the decision situation. What I propose to do is examine the consequences of assuming that an individual can always make a choice along with some of the more common assumptions concerning rationality of choice.

One additional piece of conceptual apparatus is introduced, namely mixed actions. For Bayesian decision theory, there is no gain involved in introducing mixed actions since, under Bayesian assumptions, the optimal action is always a pure strategy. In the following, we allow the possibility that if an individual cannot choose between two options, one reason might be that he would prefer a mixture of these two to either separately.

The formal model is simplified in several ways. Rather than starting with capabilities and composing these into potential actions, we assume that step taken and start with potential actions (or strategies). Similarly, the process of composing contingencies into joint occurrances will be bypassed and the model starts with a set of exclusive, possibly non-controlled states, which, for want of imagination on my part will also be called contingencies. The set of contingencies is assumed to be finite. An action x, will be represented by a vector, $x = (x_1,\ldots,x_n)$, where each component x_i specifies the utility to the decision maker of the outcome if x is taken and contingency i obtains. Thus, any two actions which engender the same vector of utilities are considered identical.

We thus have a set $X = \{x,y,z,\ldots\}$ of potential actions. This set is represented by an n-dimensional space, where n is the number of contingencies. It is assumed that X is a metric space, i.e., a distance function is defined which fulfills D1-D3 of Section 1, Chapter III. In addition, it is assumed that X is closed under mixtures. If x and y are any two points, then

$ax + (1-a)y$, $0 \leq a \leq 1$, is also a point of X. Actually there is no great loss of generality if it is assumed that X is ordinary Euclidean n-space. A decision problem consists of a subset, S, of X. A decision rule for X is a choice function C(S) which specifies a subset of S. Intuitively, C(S) identifies the members of S which are "preferred" over the other members.

There are two more or less obvious restrictions on C(S). We would not expect the decision maker to be able to make a choice if S were unbounded, i.e., if for every x in S he can find another which is preferrable to x. In the same vein, we would not expect a choice if there were sequences which, though bounded, had no limit; again, for any x, there could be a y preferable to x. Thus, we require that there by a C(S) only for those S's which are bounded from above, and which contain all their limit points.

A third condition is less common. C(S) is required only for S's which are convex -- i.e., if x and y are in S, then $ax + (1-a)y$, $0 \leq a \leq 1$, is in S. This is the condition which allows a decision maker -- if he chooses -- to reject both of two alternatives and select a mixture of the two instead. It has the small drawback that no pair of actions x and y can be compared directly. Any consideration of x and y requires taking into account all possible mixtures as well.

With these preliminaries, we can state the postulates governing the model.

H1. For every convex S which is bounded from above, and closed, C(S) exists. C(S) is a subset of S.

S is bounded from above if, for every i, there is a constant c_i, and for every x in S, $x_i < c_i$. S is closed if, for every sequence x_i, if x_i is in S and $x_i \to x$ as $i \to \infty$, then x is in S.

For the next postulates we need a definition. Intuitively, C(S) imposes a partial relationship on X in that if x is a member of C(S), it is to that extent preferred to the other members of S. This relationship will be indicated by $x \geq' y$.

DH1. $x \geq' y$ means there is an S and x belongs to C(S) and y is in S.

H2. Dominance. If $x_i \geq y_i$ for every i, then $x \geq' y$. If, in addition, $x_j > y_j$, then $x >' y$.

This is the old familiar postulate. The first inequality is a straightforward inequality on the components of the points x and y. The implied inequality is a preference relation between the points themselves. $x >' y$ simply means $x \geq' y$ and not $y \geq' x$.

$\geq^*$ will be used to designate the ancestral relation of $\geq'$.

DH2. $x \geq^* y$ means there is a sequence $x_1,\ldots,x_n$, $x = x_1$ and $y = x_n$ and $x_i \geq' y_{i+1}$, $i < n$.

That is, $x \geq^* y$ if there is a sequence of S's such that x_i is in $C(S_i)$ and x_{i+1} is in S_i.

H3. Acyclicity. If $x \geq^* y$, then not $y >' x$.

H3 requires that $\geq^*$ does not go around in a circle where at least one of the links is a strict inequality. It, in effect, enjoins the decision maker from engaging in a series of decisions in each of which he thinks he is bettering himself, and winding up accepting an alternative he had previously rejected. This postulate is closely related to the independence of irrelevant alternatives axiom that will be discussed in Chapter VI on group values.

H4. Continuity. If $x >^* y >^* z$, there are numbers a and b, $0 < a < 1$, $0 < b < 1$, such that $x >^* ax + (1-a)z >^* y >^* bx + (1-b)z >^* z$.

H4 is a familiar condition in decision theory. It guarantees that "similar problems generate similar decisions."

For the final condition, we need some additional notions. Let P_x designate the class of y's such that $y \geq^* x$, and Q_x the class of y's such that $x \geq^* y$ i.e., P_x is all the points that are "preferred to x" in the sense that some chain of choices leads from y to x, and conversely, Q_x is the set of points that x is preferred to. A positive ray R is a line where if x and y are points on R, $|x_i - y_i| > 0$, and $x_i - y_i$ has the same sign for all i. It is easy to show that for any positive ray R, and any point x not on R, R intersects P_x and Q_x and there exists the greatest lower bound (g.l.b.) of P_x on R and the lowest upper bound (l.u.b.) of Q_x on R.

H5. Archimedean. For any positive ray R, and any x not on R, g.l.b. P_x coincides with l.u.b. Q_x on R.

H5 is another kind of continuity axiom, in this case, for the boundaries of P_x and Q_x. In the form of H5 the condition is perhaps a little heavy handed; but it avoids having to define derivatives on the preference field and assuming some limit on the local variation of those derivatives.[19]

H1-H5 are sufficient to demonstrate the theorem

Theorem H1. There is a complete order on X, compatible with $\geq^*$, and C(S) consists of the maximal points in S with respect to this complete order.

The proof of Theorem H1 is fairly intricate, and thus has been relegated to Appendix II. A sketch of the proof is probably sufficient to convey the essential points.

It follows directly from the definition that $\geq^*$ is transitive, since if $x \geq^* y$ and $y \geq^* z$, then the defining sequences, $x \geq ' x_2, \ldots, x_{n-1} \geq ' y$;

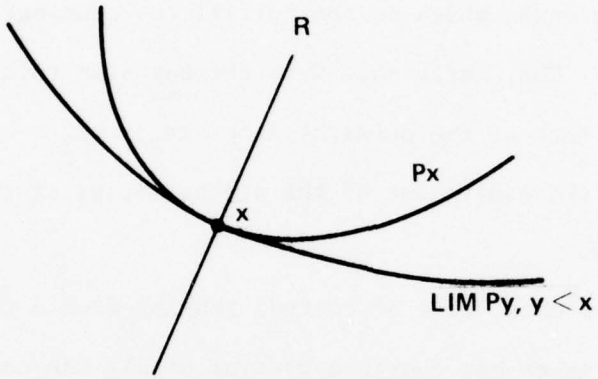
$y \geq' y_2, \ldots, y_{n-1} \geq' z$, form a single sequence connecting x and z. The transitivity of $\geq^*$ and dominance imply that along a positive ray, the sets P_x are nested.

Continuity implies that the P_x's along a positive ray are "tight," that is that if x_i approaches x along the ray, then the boundary of P_{x_i} approaches the boundary of P_x. The Archimedean condition assures that the P_x's are distinct, i.e., if x dominates y, then the boundary of P_x does not intersect the boundary of P_y. Figs. 40 a and b illustrate the kinds of pathologies ruled out by these two consequences.

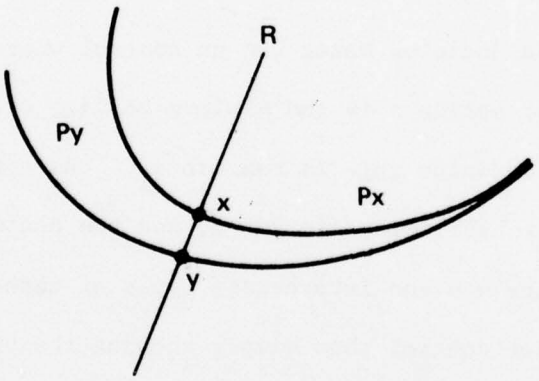
Finally, acyclicity implies that the sets of P_x's determined by different positive rays all fit together to form a single system of sets. The boundaries of this system of sets form a set of equivalence sets. The choice set C(S) of a convex set S are all the points which lie on the highest equivalence set that intersects S.

In summary, if it is assumed that an individual can make a choice out of every convex, closed, bounded from above set, where the choice fulfills the axioms of dominance, acyclicity, and continuity, then the choice can be formulated as an (ordinal) utility function on the potential actions, and the choice is made by selecting the action (or actions) with the highest utility.

This theorem complements one derived by Shapley and Shubik, in which they show that (with a similar underlying model) the assumptions of connexity, dominance, assymetry, and continuity imply a transitive preference function on X.[20] In their case, the need for an archimedean axiom is obviated by assuming transitivity for equivalence. Roughly speaking, I have shown that dominance, continuity, and acyclicity (a sort of weak form of transitivity and assymetry) imply connexity.



a) NON-TIGHT Px ON RAY R



b) NON-DISTINCT Px ON RAY R

Figure 40. Pathologies Ruled Out by Continuity and Archimedean Assumptions

The application of this result to information and control is suggested by Fig. 41. The diagram has been simplified to display only two contingencies, x_1 and x_2. Five potential decision rules have been drawn on the same diagram for comparison — in practice, of course, each would fill the entire space. A and E are limiting cases which do not fulfill the continuity axiom, H4, but do fulfill axiom H5. The middle rule C is the Bayesian rule, maximize expected value. The coefficients of the straight lines $ax_1 + bx_2 = c$, normalized so that $a + b = 1$, are the equivalent of the probabilities of the contingencies 1 and 2 respectively.

We can conceive of a scale of control ranging from A to E. In the case of A, the decision maker has complete control of all the contingencies -- he can, in effect determine which of the contingencies will occur. Thus, he can use as a utility function for an action x the maximum utility x can achieve over all the contingencies. The decision rule is thus $\max_x \max_i x_i$. At the other extreme, the decision maker has no control whatsoever. The most he can guarantee with any action x is the minimum utility over all the contingencies. Hence, the decision rule is $\max_x \min_i x_i$. As noted, the middle case is that where the probabilities are known, and the decision rule is $\max_x \sum_i p_i x_i$. B and D are new and interesting types of cases. In B, the decision maker has greater control than simply knowing the probabilities, but not complete control. Furthermore, there may be no way to resolve the decision problem by determining which of the contingencies are under control. They may all be equally "incompletely controlled." An illustrative set of equivalence curves might be $U = ax_1^2 + bx_2^2$. Under what circumstances might a decision maker choose to behave as in B? A simple case might be one where

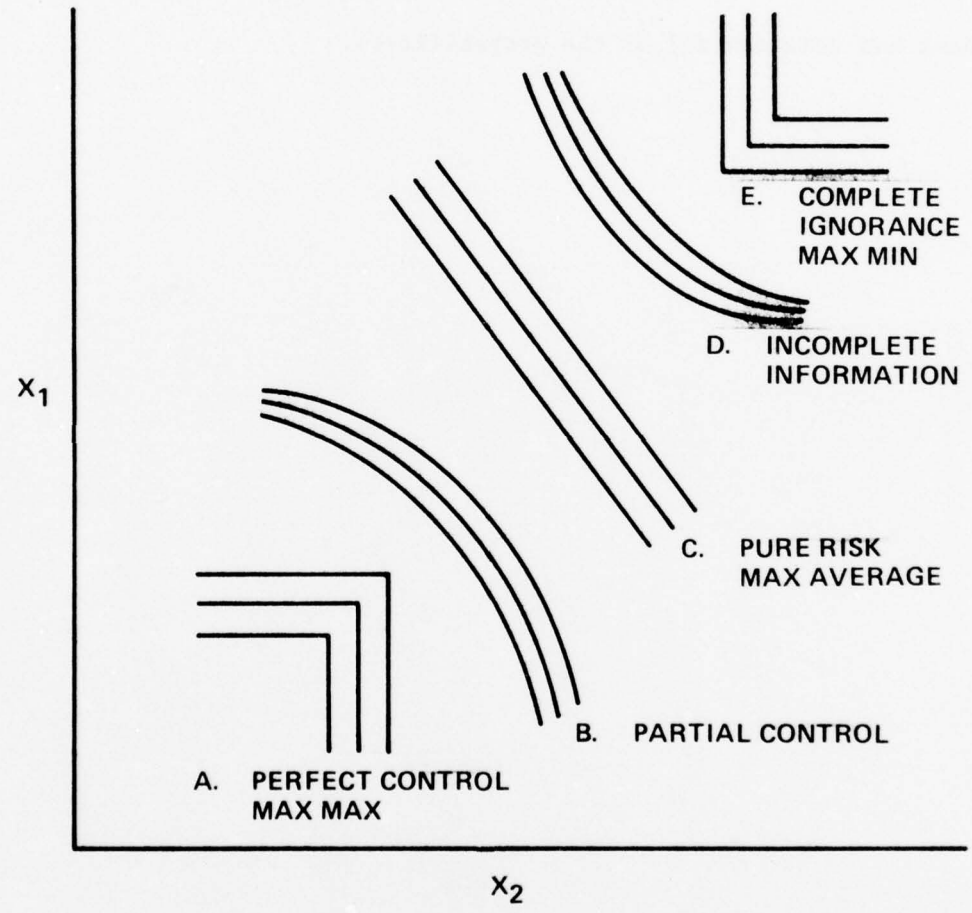


Figure 41. The Spectrum of Control

the decision maker can influence the probabilities. If he selects an x with $x_i > x_j$, then the probability of i is increased.

For D, an illustrative utility function might be $U = x_1 x_2$. In analogy with case B, D might be a case in which increasing x_i decreases the probability of contingency i. For example, D might be a case where a hostile opponent has some control of the probabilities.

CHAPTER V. AGGREGATION

1. Collective Judgment

Up to now we have been examining individual judgment from various points of view. In a sense, this has all been introductory to the present chapter, where we will investigate procedures for combining individual judgments into group judgments.

In the most general sense, we can think of the aggregation process as a way of combining the information in the heads of the members of a group and using the pooled information as a basis for a new estimate. In theory, the group process could consists of each individual stating everything that he can recall that is relevant to the question under consideration, and then applying some inductive procedure to the combined list of recalled items. For most interesting questions, such an exercise however, is totally impractical. As we saw in Chapter II, even relatively simple questions can generate a massive amount of miscellaneous "stuff," covering the gamut of relevance and solidity. Just how extensive this catalogue is for everyday decisions has never been explored, so far as I know. Some of it appears to be very difficult to articulate; and it may even be the case that some of it is inarticulable, either for lack of appropriate words, or because it does not reach the level of full consciousness. However, even assuming that all of the material can be elicited and spread out for full view, and assuming that the qualifications of relevance and solidity could be expressed in scales comparable across all members of the group, there would still remain the problem of taking that long list of items and formulating an answer based on it. At present, we do not have formal amalgamation techniques for such unstructured material.

In addition to recall of relevant material, it seems to be the case that there is something which could be called estimation skill; some individuals

can use unstructured inputs more effectively than others to generate estimates. Presumably, whatever procedure was designed to capitalize on the list of relevant material would also have to aggregate the skill components of the estimation process.

One very rough approximation to some of this is found in an aspect of current practice with group decisions, namely discussion. In a discussion, it is possible to share both information and "insights," i.e., ways of putting the information together. There is a fairly extensive literature that indicates that in practice the sharing is likely to be incomplete.[1] But we can imagine a process in which all relevant material is elicited, all "insights" are expressed, each individual separately aggregates the combined information and hints, and then some group process such as agreeing on a common answer is used to arrive at a group estimate.

Without extensive experiments, it is difficult to decide how effective procedures of this sort might be. Experiments to date do not give a clear picture of the relative effectiveness of various types of group interaction. There are a number of obscuring factors: differences in the type of estimation task (kind of question), difficulties in controlling group dynamics, variations in figures of merit, and, of course, variations in individual performance. Small groups are remarkably complex objects and the number of potential kinds of organizations that can be devised to carry out even so simple a task as estimating the answer to an uncertain question is practically infinite.

One methodological hypothesis that has guided a great deal of the present work is just this: to a first approximation, the most complete summary that an individual can (practically) furnish concerning what he knows about a question is just his estimate of the answer to that question, plus, perhaps,

an estimate of the solidity of his answer. In formulating his answer, the individual has taken into account the nuances of relevance and shadings of solidity that apply to his own information.

The hypothesis is not easy to verify, or, for that matter, to express in a manner leading to simple experiments. That's why I call it a working hypothesis.

The hypothesis suggests that most of what the group has to offer can be realized by starting with separate, in fact, independent, estimates from the members of the group and seeking the most effective formal ways to combine these independent estimates into a group estimate. We can call this simple procedure an elementary group estimate. A number of ancillary desirable features come along with this approach: (a) The definition of the group process can be made explicit and precise. (b) Application of figures of merit can be pursued by theoretical investigations, as well as by experiment. (c) A kind of "rock bottom" level of performance is defined which can act as a criterion for other group procedures. (d) The procedures are remarkably easy to implement (and replicate) in practice.

For elementary group estimates, then, there will be a set of individual responses $R = (R_1,\ldots,R_n)$ where R_i is the response of individual i, and n is the number of members of the group. These responses are relative to a universe of discourse U, and (usually) to a specific question concerning U. For most cases, the specific question will be represented by a particular partition of U into an event space $\{E_j\}$, in which case the individual responses will be of the form R_{ij} — i's estimate for the event E_j.

A group judgment is some function F, which generates a group response G based on the responses R. F can be a function of more than just the overt individual responses; it may depend on the specific group (e.g., in the form

of differential weights on the individuals), or on the form of the question. By and large, factors of this sort will be dealt with by specific notation when called for. For general discussion, we can write G = F(R).

With this simple definition of a group response, we can investigate a number of pertinent questions: (1) How does the group compare to the individual members of the group? This question divides into two subquestions: (a) How does the group performance compare to the average performance of the individuals? (b) How does the group compare to the best individual? (2) How does the accuracy of the group depend on the amount of disagreement (dispersion) among the members? (3) How does the group compare to the a priori knowledge available without the group? (4) How much is lost by employing various approximative techniques for aggregating the individual responses?

The generic notion of an n-heads rule will be used to refer to the demonstration that the group performance is superior to the individual performance in some specified way. Given the wide variety of estimation types, and the range of figures of merit discussed in Chapter III, a broad spectrum of n-heads rules can be explored.

2. Basic Rules

In this section a number of n-heads rules are examined that are of a paeticularly simple form. They apply, for the most part to any quantities that are determined at least to an interval scale. The rules assume only the existence of a set of individual estimates R, and an unknown true response T. They are "distribution free" — i.e., are independent of the shape of the distribution of the responses. In this respect, the rules are closer akin t arithmetic than to statistics.

Given R and T, there are three definitional items needed to formulate an n-heads rule: (1) an aggregation rule F(R), (2) a score rule S(R,T), and

(3) a criterion for comparing individual and group scores. A typical criterion is the difference between the average individual score and the group score. Generally, it is not possible to optimize such a criterion for a given score rule, since T is unknown. However, it is possible to establish useful inequalities.

The number of possible combinations of aggregation rules, score rules and criteria is essentially unlimited. In this section I have limited the investigation to a few aggregation rules resembling measures of central tendency, to various simple scaled distance scores, and to either the criterion comparing the group score with the average individual score, or the analogous criterion with the median substituted for the average.

Table I displays seven such rules showing the aggregation function, the score rule, and the n-heads statement. Number 1, for example, states that the error of the mean (average) is always less than or equal to the average individual error.

Table I. Elementary N-Heads Rules

	Aggregation Function	Score Rule	N-Heads Rule
1.	$\bar{R}$ (Average)	$\lvert R - T \rvert$	$\lvert \bar{R} - T \rvert \leq 1/n \sum \lvert R - T \rvert$
2.	$\bar{R}$ (Average)	$(R - T)^2$	$(\bar{R} - T)^2 \leq 1/n \sum (R - T)^2$
3.	$\bar{R}$ (Average)	$\left\lvert \frac{R - T}{T} \right\rvert$	$\left\lvert \frac{\bar{R} - T}{T} \right\rvert \leq 1/n \sum \left\lvert \frac{R - T}{T} \right\rvert$
4.	$\bar{R}$ (Average)	$\frac{(R - T)^2}{\lvert T \rvert}$	$\frac{(\bar{R} - T)^2}{\lvert T \rvert} \leq 1/n \sum \frac{(R - T)^2}{\lvert T \rvert}$
5.	Md (Median)	$\lvert R - T \rvert$	$\lvert Md - T \rvert \leq Md\ [R - T]$
6.	GM (Geometric Mean)	$\lvert \log R - \log T \rvert$	$\lvert \log GM - \log T \rvert \leq 1/n \sum \lvert \log R - \log T \rvert$
7.	HM (Harmonic Mean)	$\lvert 1/R - 1/T \rvert$	$\lvert 1/HM - 1/T \rvert \leq 1/n \sum \lvert 1/R - 1/T \rvert$

All Σ's over the set of individual responses.

Except for 5, the rules state that a given scaled distance score is smaller for some measure of central tendency than the average of that score for the individuals. 5 has the same form except that the median is substituted for the average. The analogues of 2, 3, 4 for the mean hold for the other three, but don't appear to have much relevance for present practice. 7 may also appear to be somewhat "academic," since the harmonic mean has not been used, in my experience, to aggregate individual estimates. I have included it in part to show that the general form of an elementary n-heads rule is not restricted to the better known types of scores or aggregation rules; but also it is possible that for some types of estimates — e.g., ratios — the harmonic mean may be an appropriate aggregation function.

The rules involving averages of absolute values all follow from a single principle, namely

$$\left| \sum_i x_i \right| \leq \sum_i \left| x_i \right| \tag{1}$$

which follows directly from

$$\left| \sum_i x_i \right| = \sum_+ x_j - \sum_- \left| x_k \right|$$

where $\sum_+$ is the sum over the positive x's and $\sum_-$ is the sum over the negative x's. This is clearly less than or equal to $\sum_+ x_j + \sum_- \left| x_k \right| = \sum_i \left| x_i \right|$.

The rules involving squared differences follow from

$$1/n \sum_i x_i^2 \geq (1/n \sum_i x_i)^2 \tag{2}$$

set $1/n \sum x_i = m$, then

$$1/n \sum_i (x_i - m)^2 \geq 0$$

$$1/n \sum_i (x_i^2 - 2xm + m^2) \geq 0$$

$$1/n \sum_i x_i^2 - m^2 \geq 0$$

whence (2) follows.

The various rules follow from (1) and (2) by substituting the appropriate expression for x; e.g., x = (R - T)/n for 1. The scaled rules, 3 and 4, follow from 1 and 2 via the fact that dividing each side of an inequality by a positive constant does not affect the inequality. The rules involving scaled values are not significant if T = 0, whence most of the scaled rules are appropriate only for ratio scales with T greater than zero.

From (2) we can formulate a more illuminating form of 2, namely,

$$(\bar{R} - T)^2 = 1/n \sum (R - T)^2 - \text{Var } (R) \qquad (3)$$

(3) is the same as (1) in Chapter III, now applied to a set of individual estimates, rather than to a sequence of estimates by a single individual. (3) states that the squared error of the mean is equal to the average squared error of the individuals minus the variance of the individual responses. Thus, the advantage of the mean over the individual increases with the amount of disagreement among the individuals.

These elementary n-heads rules provide a justification for using a group estimate where there is little or no basis for invoking one of the theories of estimation from Chapter II. Their meaningfulness for group estimation may seem a little mysterious since the rules themselves are true for any set of numbers R, and any other number T. The link is provided by the tacit assumption in practice that the group furnishing the estimates R have some pertinent information concerning the number T.

We can derive a statement which contains the size of the group as an explicit factor. Roughly speaking, the error of the group declines with increasing n, the number of members of the group. This will be stated only for the most elementary formulation of this relationship, primarily to illustrate that even for the "rock bottom" n-heads rules, n is a significant parameter. More diagnostic formulae relating the size of the group to accuracy can be derived using the various theories of estimation.

Suppose we have a group of n individuals. We can ask how the accuracy of this group compares with the average accuracy of the n subgroups that can be formed by leaving out one member at a time. Let x^j designate the average of the n-1 responses omitting response R_j, i.e., $x^j = \frac{1}{n-1} \sum_{i \neq j} x_i$. We have, from 1, Table I.

$$\left|1/n \sum_j x^j - T\right| \leq 1/n \sum_j \left| x^j - T\right|$$

$$1/n \sum_j x^j = 1/n \sum_j \frac{1}{n-1} \sum_{i \neq j} x_i = \frac{1}{n} \sum_i x_i,$$

whence

$$\left|1/n \sum_i x_i - T\right| \leq \frac{1}{n} \sum_j \left|x^j - T\right| \qquad (4)$$

(4) asserts that the average error of the subgroups with n-1 members is greater than (or equal to) the error of the total group with n members. Since this is true for any n, the average error monotonically decreases with n (providing we average over all available respondents for each potential group of n.)

Another way of viewing the elementary n-heads rules in Table I that may illuminate their applicability to group estimation is the following. Suppose we assume that each individual has an equal probability of being correct.

We would like to find a group estimate G that minimizes the expected error. In the case of the squared distance as the figure of merit, we would like to minimize the expectation of $(G - T)^2$. In case individual i is correct, the error would be $(G - R_i)^2$ and the expected error is then $1/n \sum_i (G-R_i)^2$. If we differentiate this expression with respect to G, and set the result equal to 0, we obtain

$$1/n \sum_i 2(G-R_i) = 0$$

whence

$$G = 1/n \sum_i R_i = \bar{R} \qquad (5)$$

In the context of adjudicating disagreement within the group, each individual i "sees" the group as making the error $(G - R_i)^2$. (5) states that R minimizes the average perception of the group error.

3. Theory of Errors

The theory of errors, as expounded in Chapter II, assumes that each individual's response is a sum of the true answer, a bias term, and a random error; i.e., $R_i = T + B_i + \epsilon_i$, where T and B_i are constants and ϵ_i is distributed with zero mean and some standard deviation S_i.* Each individual response is thus a random variable, with a distribution $D_i(R_i)$, with standard deviation S_i and mean $M_i = T + B_i$.

By definition, the error distributions of different individuals are independent, since the errors are assumed to be random. The vector R thus has the joint distribution $D(R) = \prod_i D_i(R_i)$. The joint distribution D(R) determines a distribution for the average of the individual responses,

*The non-conventional notation S_i is used for the standard deviation in this section to allow a simple distinction between expressions referring to estimates and expressions referring to the logarithms of estimates.

$\bar{R} = 1/n \sum_i R_i$; we can call this derived distribution $D(\bar{R})$. The importance of $D(\bar{R})$ for group estimation lies in the presumption that $\bar{R}$ is a reasonable expression of the group response. Part of the basis for this presumption is simple carryover from standard statistics, where $\bar{R}$ is the most common representative statistic, or equivalently, the most common measure of central tendency. In standard statistics the role of a construct like $\bar{R}$ is to characterize a population. The role of a group response in group decisions is to obtain the most accurate — or highest scoring — estimate based on the individual responses. We will show below that there are some persuasive n-heads rules associated with $\bar{R}$. However, it should be emphasized that these rules are not simple extensions of the role of $\bar{R}$ in standard statistics. In particular, the notion of $\bar{R}$ as a representative statistic is probably misleading as an "explanation" for its usefulness as a group response.

By a well known result,[2] the mean M of $\bar{R}$ is

$$M = 1/n \sum_i M_i. \tag{6}$$

The variance S^2 of $\bar{R}$ is

$$S^2 = 1/n^2 \sum_i S_i^2 \tag{7}^*$$

$$S = 1/\sqrt{n} \sqrt{1/n \sum_i S_i^2} \tag{8}$$

The first n-heads rule to follow from the theory of errors, then, could be labelled the n-heads rule for the standard deviation. (8) asserts that the standard deviation of R is $1/\sqrt{n}$ times the square root

---

*(6) holds for any joint distribution. (7) contains the assumption that the individual responses are independent.

of the average of the individual variances. If the individual variances are roughly equal, then the standard deviation of the group response is less than the individual standard deviations by a factor of $1/\sqrt{n}$. In behavioral terms, the average random error of R will be smaller by a factor of $1/\sqrt{n}$ than the random error of the individuals. Thus the group response will be more stable than the individual responses, and the likelihood of a large random errors on the part of the group will be reduced.

(8) tells us nothing about the bias of $\bar{R}$. A second n-heads rule can be derived which deals with the bias. We have $M = 1/n \sum_i M_i = T + 1/n \sum_i B_i$. The bias of the mean B is thus $1/n \sum_i B_i$. From 1, Table I,

$$|1/n \sum_i B_i| \leq 1/n \sum_i |B_i| \tag{9}$$

In words, the bias of the mean is always less than or equal to the mean bias. Invoking 3 we can assert

$$\bar{B}^2 = \overline{B^2} - \mathrm{Var}(B) \tag{10}$$

In words, the squared bias of the mean is equal to the average of the individual squared biases minus the variance of the individual biases. If all the individual biases are the same, then the group offers no advantage as far as bias is concerned; if the individual biases differ, then the group advantage is measured directly by the variance of the individual biases.

For the total error, the two effects "add"; the expected squared error of the mean $E(R - T)^2 = B^2 + S^2$. The same expression holds for the individual expected square errors, i.e., $E(R_i - T)^2 = B_i^2 + S_i^2$. Thus, if we abbreviate the average expected squared error of the individuals, $1/n \sum_i E(\bar{R}_i - T)^2$, by ESE, we have

$$E(\bar{R} - T)^2 = 1/n \ \mathrm{ESE} + 1/n^2 \sum_{i \neq j} B_i B_j \ . \tag{11}$$

If the individual biases include both positive and negative instances, so that the second term in 11 is small, then abbreviating $E(\bar{R} - T)^2$ by EM (expected error of the mean), we have

$$EM \sim 1/\sqrt{n} \quad ESE \tag{12}$$

As an illustration, consider the values in Table II.

Table II

i	B_i	S_i	ESE_i
1	-1	1	2
2	2	1.5	6.25
3	3	2	13

$$ESE = 7.0833$$

$$EM = 4.1107$$

$$1/\sqrt{3}\ ESE = 4.090$$

The situation is somewhat more complex for the expectation of the absolute error, $E|\bar{R} - T|$. For any distribution $D(\bar{R})$ of $\bar{R}$, the expected absolute error can be computed as

$$E|\bar{R} - T| = \int_{-\infty}^{T} (T - R)D(R) + \int_{T}^{\infty} (\bar{R} - T)D(\bar{R}) \tag{13}$$

Rearranging the terms in (13) and adding and subtracting $T \int_{-\infty}^{T} D(\bar{R})$ and $\int_{-\infty}^{T} \bar{R}D(\bar{R})$ we obtain

$$E|\bar{R} - T| = M - T + 2T \int_{-\infty}^{T} D(\bar{R}) - 2 \int_{-\infty}^{T} \bar{R}D(\bar{R}) \tag{14}$$

(14) is not particularly informative without knowing the form of the distribution $D(\bar{R})$. If we introduce the psychonumeric hypothesis from Chapter II, i.e., assume that the individual responses are log normal, and in addition assume that they are independent, then we can derive that the distribution of the geometric mean $\left[\prod_i R_i\right]^{\frac{1}{n}}$ is log normal.[3]

Using the notational convention of Chapter II, where lower case letters refer to the logarithms of quantities expressed by upper-case letters, $r_i = \log R_i$, m_i = mean of individual's log response distributions, $\bar{r} = 1/n \sum_i r_i$ is the logarithm of the geometric mean of R, $m = 1/n \sum_i m_i$, is the mean of $D(\bar{r})$, the distribution of mean log responses. Corresponding to (7) we have $s^2 = 1/n^2 \sum_i s_i^2$, where s_i^2 is the variance of the individual log responses, and s^2 is the variance of $\bar{r}$. $D(\bar{r})$ can be formulated explicitly; it is

$$D(\bar{r}) = \frac{1}{\sqrt{2\pi}\, s} e^{-\frac{(r-m)^2}{2s^2}} \tag{15}$$

There is a corresponding expression for $\bar{R}$, but there is no particular point in writing it down here.

Introducing (15) in (14) with the appropriate transformation of variables, we have (ae means average error)

$$ae = m - t + 2t \int_{-\infty}^{t} D(\bar{r}) - 2 \int_{-\infty}^{t} \bar{r} D(\bar{r}) \tag{16}$$

If we set $b = (t-m)/s$, and perform the integration on the last term, and recombine, we obtain

$$ae = -bs + 2bs\Phi(b) - \sqrt{2/\pi}\, s e^{-\frac{b^2}{2}} \tag{17}$$

where $\Phi(b)$ is the cumulative normal distribution with zero mean and unit standard deviation evaluated at b.

Thus, we can write

$$ae = sf(b) \tag{18}$$

where $f(b) = -b + 2b\Phi(b) - \sqrt{2/\pi}\, e^{-\frac{b^2}{2}}$.

(18) is perhaps deceptively simple, since b involves s.

If we take the derivative of f(b) with respect to b, we obtain

$$\frac{df(b)}{db} = -1 + \Phi(b) \tag{19}$$

The cumulative normal is virtually a constant beyond b = 2; hence for $b \geq 2$, f(b) is essentially a straight line with slope 1 and ae is directly proportional to s.

In Figure 42 the observed log error is plotted against the observed standard deviation of log responses for roughly 300 almanac questions.[4] The subjects were upper-class and graduate college students. The number of subjects per question was about 14 (ranging from 11 to 15). The lower dashed line is computed from (17) assuming b = 0, and assuming that the observed standard deviation is an acceptable estimator of $\sqrt{1/n \sum_i s_i^2} = s$. The latter assumption is correct only if the variance of the bias is approximately equal to s, the variance of $\bar{r}$. There is reason to suspect that for this set of data, the variance of the bias is greater than s, in which case the observed standard deviation is an overestimate of s, and the dashed line should be lower; however, there is no way to determine the individual variances from the date, and Fig. 42 can be used only to set a lower bound on the bias.

From Fig. 42 $E/s \sim .65\sqrt{14} \sim 2.43$. Since E/s is the estimate of the bias, b, we see that for this data, $b > 2$, and hence, the relationship between E and observed standard deviation should be a simple proportion, which indeed the figure demonstrates.

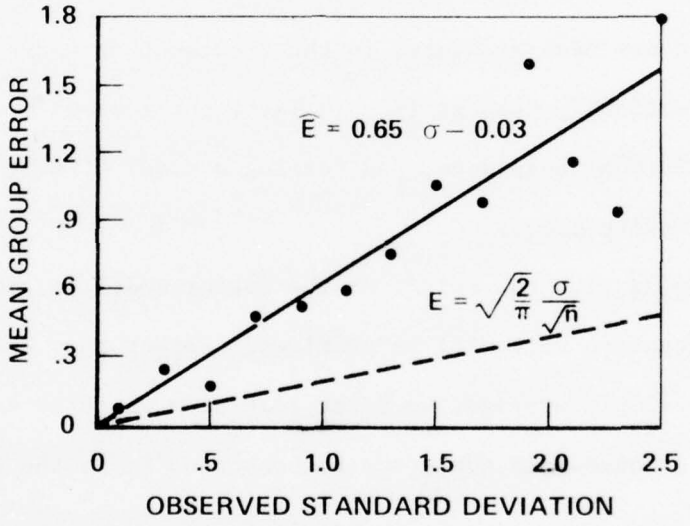


Figure 42. Relation Between Log Error and Observed Standard Deviation

For this particular set of data, then, the bias makes a much larger contribution to average error than the average random deviation. We would expect that the major contribution of the group in reducing the expected error. arises from reducing the bias via (9) then from reducing the standard deviation via (8), but additional analysis of the data would be required to establish that hypothesis.

4. Factor Model

Most of the formalism needed to discuss n-heads rules for factor models of estimation has been presented in the treatment of correlation and uniform weights in Section 7, Chapter IV. In fact, the analogy between aggregating a set of individual responses, and forming a model of multi-dimensional functions is quite close.

In keeping with the spirit of the factor model, a somewhat more general kind of aggregation rule will be examined. Rather than specializing immediately to the simple average, we first look at a weighted average, i.e., $R = \sum_i a_i R_i$, where R is the group response and R_i is the response of individual i. The notion of a weighted average for aggregating individuals has a certain amount of appeal, on the grounds that some individuals are more likely to give an accurate response than others, either because of greater information or greater skill in forming estimates or both. However, the issue of how to assign the weights does not appear to have a satisfactory resolution at present. I will use the results presented in the treatment of equal weights in Chapter IV to show that unless the individuals are very different in their capabilities, little is gained by non-uniform weighting.

Using correlation as the figure of merit for the estimates, we have $\rho(R,T) = E[(R-\bar{R})(T-\bar{T})/s_R s_T]$. Unpacking this expression in terms of R, and

rearranging, we obtain

$$\rho(R,T) = 1/s_R \sum_i a_i s_i \rho(R_i,T) \tag{20}$$

where s_i is the standard deviation of individual i's estimates and $\rho(R_i,T)$ is the correlation of individual i's estimates with the true values T. The correlation of the weighted average of the individual estimates with the true values is just the weighted average of the individual correlations with $a_i s_i / s_R$ as weights. Since $s_R^2 = E[(\bar{R} - R)^2]$, unpacking gives

$$s_R^2 = \sum_i a_i^2 s_i^2 + 2 \sum_{i<j} a_i a_j s_i s_j \rho(R_i,R_j) \tag{21}$$

$\rho(R_i,R_j)$ is the correlation of individual i's estimates with individual j's estimates.

There is no loss of generality in assuming that the individual estimates have been normalized by z-scores, so that $s_i = 1$ for all i, and (20) thus becomes

$$\rho(R,T) = 1/s_R \sum_i a_i \rho(R_i,T) \tag{22}$$

$$s_R^2 = \sum_i a_i^2 + \sum_{i<j} a_i a_j \rho(R_i,R_j) \tag{23}$$

It is clear from (23) that s_R is a maximum when $\rho(R_i,R_j) = 1$ for all i and j. In this case, $s_R^2 = \sum a_i^2 + 2 \sum_{i<j} a_i a_j = (\sum a_i)^2$; thus $s_R = \sum_i a_i$ and thus

$$\rho(R,T) \geq \sum_i \frac{a_i}{\sum_j a_j} \rho(R_i,T) \tag{24}$$

That is, the correlation of R (weighted average of the individual estimates) with T is greater than or equal to the weighted average of the individual correlations. In particular, if we have the case of equal weights,

$$\rho(R,T) \geq 1/n \sum_i \rho\ (R_i,T) \tag{25}$$

(25) is the most straightforward form of n-heads rule for factor models. It asserts that the correlation of the average of a set of estimates with the true answer is greater than the average of the individual correlations, equality occurring only in the uninteresting case that all the individuals give the same estimates.

If the responses of the individuals are independent — $\rho(R_i,R_j) = 0$ for all i and j — then setting $\bar{\rho} = 1/n \sum_i \rho\ (R_i,T)$,

$$\rho(R,T) = \sqrt{n}\ \bar{\rho} \tag{26}$$

The correlation of the average response with the true answer is precisely $\sqrt{n}$ times the average correlation of the individual responses with the true answer. The assumption of independence does not appear very plausible if the set of questions all refer to the same quantity, i.e., each individual is making estimates within the same model. However, the formal apparatus developed above applies equally well to the case of a list of separate questions. It is not clear that the correlation is a useful figure of merit in the case of a miscellaneous string of questions, but to the extent that covariance of an individual's answers with the true answers indicates some knowledge, and to the extent that the individual's answers are independent, (26) indicates a strong advantage of the average answer over the individual answers.

In general, there will be a set of optimal weights for the individuals which maximizes the correlation of the weighted average with the true answer. Such weights are difficult to come by in practice.* However, there is a simple analogy between estimating a quantity with a linear combination

*It is necessary to know both the individual correlations $\rho(R_i,T)$ and the intercorrelations $\rho(R_i,R_j)$.

of variables, and estimating a quantity by a linear combination of separate individual estimates. If the optimal weights are completely unknown, then (14) Chapter IV can be invoked to demonstrate that the maximum expected correlation of an assumed linear combination of individual estimates with the optimally weighted combination is obtained with equal weights. As is clear from the numerical results in Chapter IV, a great deal must be known about the individual estimates before something better than uniform weights can be devised.

5. The Impossibility Theorem

The aggregation of probability estimates differs from magnitudes estimates in that the theory of probability imposes a relatively rigid set of constraints on the resultant group estimates. In Chapter II, Section 5, the three axioms

A1. $0 \leq P(E)$

A2. $P(U) = 1$

A3. $P(E \vee F) = P(E) + P(F)$, providing $E.F = 0$,

were listed as basic postulates for numerical probability. Assuming that the individual members of the group are consistent probability estimators, their estimates will follow A1 - A3. Similarly, if we have a group function, F(R), it must also fulfill A1-A3.

If we let L stand for the unit-vector (all components = 1), we have

G1. $0 \leq F(R)$

G2. $F(L) = 1$

G3. $F(R+S) = F(R) + F(S)$, providing $R_i + S_i \leq 1$.

Where $R+S = (R_1+S_1,\ldots,R_n+S_n)$. G3 may seem a little strong, since A3 is asserted only for those cases in which the appropriate events are exclusive.

However, given any R and S which fulfill $R_i + S_i \leq 1$, there is a potential set of estimates for some E and F where E and F are exclusive and R and S are the individual estimates for E and F respectively, so there is no loss of generality in omitting the exclusivity condition.

G4. F is a function solely of the numerical vector R.

Other functions, where F depends on additional features of the decision situation can be devised. For example, F could involve various kinds of dependencies among the R_i. Functions of this sort will be treated in Section 7 below. In the present section, attention is limited to functions which depend only on the set of individual estimates.

One additional assumption completes the set,

G5. F is a continuous function of R.

Theorem 1: A1-A3, G1-G5 imply that $F = \sum_i a_i R_i$,

where $\sum a_i = 1$.

The theorem states that the only function fulfilling A1-A3, G1-G5, is the linear function $\sum_i a_i R_i$ with constant coefficients summing to 1. In other words, the group estimate is a weighted average of the individual estimates.

Lemma 1: $F(L-R) = 1-F(R)$

Proof: $F(L) = F(L-R + R) = F(L-R) + F(R)$, from G3.

From G2, $F(L) = 1$, whence the result follows.

Lemma 2: $F(aR) = aF(R)$, where a is any positive real number, $aR_i \leq 1$.

Proof: From G3, $F(nR) = nF(R)$, $nR_i \leq 1$, where n is an integer. Similarly, $F(R) = mF\left(\frac{1}{m} R\right)$. Putting these two together, we get $F\left(\frac{n}{m} R\right) = \frac{n}{m} F(R)$. Since F is continuous in R, the result follows.

Lemma 3: If $f(x)$ is a function of a single variable, and $f(x+y) = f(x) + f(y)$, then $f(x) = ax$, with constant a.

Proof: By an argument similar to that in Lemma 2, we obtain $f(ax) = af(x)$. Since f is a function of a single variable, we can set $x = 1x$, whence $f(1x) = xf(1)$, and setting $a = f(1)$, the result follows.

Lemma 4: Let R^i denote the vector where $R^i_i = R_i$, and $R^i_j = 0$, $j \neq i$. Then $F(R) = \sum_i F(R^i)$.

Proof: From G3 and $R = \sum_i R^i$.

Lemma 5: Let $F_i(R) = F(R^i)$. Then $F_i(R) = a_i R_i$.

Proof: Since $F_i(R)$ is a function of R_i alone, Lemma 3 applies.

Putting Lemma 4 and Lemma 5 together, we obtain $F(R) = \sum_i a_i R_i$.

Lemma 6. $\sum_i a_i = 1$.

Proof: $L = \sum_i 1^i$, whence $F(L) = \sum_i a_i F(1^i) = \sum_i a_i 1^i = 1$.

But $\sum_i a_i 1^i = \sum_i a_i$.

This completes the proof of the theorem.

In a previous publication, I announced an impossibility theorem for aggregation of probability estimates.[5] The impossibility arises from adding one further condition to G1-G5, namely:

G6. $F(R \cdot S) = F(R)F(S)$, where $R \cdot S = (R_1 S_1, \ldots, R_n S_n)$.

G6 embodies the product rule, $P(E.F) = P(E)P(F|E)$. This rule is sometimes taken as a postulate in probability theories, and sometimes taken as a consequence of the definition of the conditional probability $P(F|E) = P(E.F)/P(E)$. As in the case of exclusivity for G3, the analogue of the product rule for groups, G6, must be expressed without the restriction that

the relevant events are independent since for any R and S, there may be a pair of events E and F where R is the set of individual estimates of P(E) and S is the set of individual estimates of P(E|F).

Theorem 2: A1-A3 and G1-G6 are incompatible.

Proof: $\sum_i a_i R_i S_i \neq (\sum_i a_i R_i)(\sum_i a_i S_i)$. The non-equality is clear, but to give a simple example: if $a_i = 1/n$ for every i, then $\sum_i a_i R_i S_i = 1/n \sum_i R_i S_i$ whereas $(\sum_i a_i R_i)(\sum_i a_i S_i) = 1/n^2 (\sum_i R_i S_i + \sum_{i \neq j} R_i S_j)$. Equality would require $(n-1) \sum_i R_i S_i = \sum_{i \neq j} R_i S_j$. If n = 2, this implies $\sum_i R_i S_i = \sum_{i \neq j} R_i S_j$, which holds only if $R_1 = R_2$ or $S_1 = S_2$.

G6, with the other conditions except G3, implies that $F(R) = \prod_i R_i^{a_i}$, with $\sum_i a_i = 1$. This follows directly from Theorem 1 by setting $r_i = \log R_i$, $r = (r_1, \ldots, r_n)$, and rewriting G6 as G'6, $F'(r+s) = F'(r) + F'(s)$. Then Theorem 1 states $F'(r) = \sum_i a_i r_i$. Taking the antilogarithm gives the result. From the standpoint of the additivity of probabilities for exclusive events, the only consistent aggregation function is the weighted average. From the standpoint of the product rule for joint probabilities, the only consistent aggregation function is the weighted product.

Whether Theorem 2 is to be considered a strong impossibility theorem for probability aggregation is not completely clear cut. It certainly rejects normal practice in applying the probability calculus. It states that, even for independent events, it is not legitimate to obtain group estimates for two events separately and then multiply these to arrive at the group estimate for the joint occurrence. On the other hand, it could be contended that there is

nothing in A1-A3 which implies that this is the way in which probabilities for joint occurrences are to be obtained, even for individual estimates. Thus, Theorem 1 allows the procedure of first obtaining group estimates for all absolute (non-relative) probabilities on an event space U, and then defining all relative probabilities in the usual way. The multiplication rule would then hold for all these derived probabilities.

Although this procedure is logically impeccable, it has the awkward feature that pairs of events which every member of the group consider independent, may not be independent in the group probability distribution on U. And, of course, the derived relative probabilities will not be equal to the aggregates of the individual relative probabilities.

This appears to be a case where the Emerson principle may override elementary logic. It certainly is desirable that each member of the group make consistent probability estimates — otherwise we are somewhat at sea in evaluating the individual estimates. It is perhaps even more desirable that the group estimates form a consistent set, since computations will be made with them, and if they are not consistent, large errors can arise by "compounding" the inconsistencies.

Some light is shed on the issue here by reverting to the aggregation of non-probabilistic magnitudes. Strictly speaking, the analogue of Theorem 1 holds for any additive quantity, such as length, weight, etc. For example, if we wish to obtain the combined weight of a given object, e.g., the weight of an envisaged spaceship, by estimating the weights of the components, then the analogue of Theorem 1 would hold that the only consistent form of aggregation for individual estimates must be a weighted average. So far, there is no difficulty, since the magnitude weight does not involve anything comparable

to the multiplication rule for probabilities.* However, if we want to consider a multiplicative aggregate of two linear quantities, such as a performance criterion consisting of the product of speed and payload, then an analogous difficulty will arise. The aggregate of the product will not be the product of the aggregates. In fact, this difficulty will hold for any nonlinear combination of the two linear quantities.

It seems clear that choosing an aggregation procedure for quantities which are subject to mathematical operations outside the scope of their definitions requires criteria that will be incompatible with simple consistency.

6. Probabilistic Aggregation

In the preceding section we saw that there is no aggregation function for probabilities that is consistent with a set of individual probabilities. Armed with the Emerson principle, we do not have to remain content with this result — i.e., we can still ask whether there is some way to aggregate a set of probability estimates which is not consistent with the individual estimates, but which performs well. This is a difficult topic to deal with on a general level, since many of the more interesting results depend on special features of particular score rules. We first examine some results with averages, which shows that the simple average is not too bad. To discuss these results, it is useful to characterize the group in somewhat more detail than we have done up to now. For most of this section, it is sufficient to think of the group as an enterprise; that is, the group is characterized by a common decision matrix U_{ij}, the utility to the group if action A_i is taken and event E_j occurs, and it is taken for granted that the group will select some common action A_g.

---

*There is no problem involved with multiplication by a scalar (non-dimensional number).

Concave scoring rules. Consider the case where the only group task is to estimate a probability distribution on an event space, and where the group utility can be represented by a concave score rule $S(R,j)$. A concave score rule is one for which $S(aR + (1-a)R') \geq aS(R,j) + (1-a)S(R',j)$. The quadratic score, $2R_j - \sum_k R_k^2$, and the logarithmic score, $a \log R_j + b$, are examples of concave rules. In this case, if we examine the objective expected payoff OES_i that would be realized by following the advice of individual i, we have

$$OES_i = \sum_j P_j S(R_i,j) \tag{27}$$

where, as usual, P is unknown. The weighted average of these expected payoffs is

$$\sum_i a_i \, OES_i = \sum_i a_i \sum_j P_j S(R_i,j) \tag{28}$$

$$= \sum_i P_j \sum_i a_i \, S(R_i,j)$$

If $S(R,j)$ is concave, we have

$$\sum_i a_i \, OES_i \leq \sum_j P_j S(R_g,j) \tag{29}$$

where

$$R_g = \sum_i a_i R_i$$

(29) asserts that for concave score rules, the average expected score is always less than or at most equal to the expected score of the average estimate. This is a fairly strong result, in that it does not depend on the actual probability, and is true for any concave score rule, and any set of weights a_i. Thus, for informational scores like the logarithmic and the quadratic rules, the average of the individual estimates will always produce a higher expected score than the average expected score of the individuals.

The results, of course, specializes immediately to the non-weighted average

$$\frac{1}{n}\sum_i OES_i = \sum_j P_j(\bar{R}, j) \tag{30}$$

For groups whose primary output is a set of estimates (e.g., consulting firms) and for which the informational scores are a reasonable measure of performance, (29) or (30) are basic aggregation formulae.

To illustrate this result, in Figure 43 the group realism curve for the average of the individual probability estimates derived from the date of Capen is presented. The lower solid curve is the individual realism curve from Figure 9 Chapter II. The upper dashed curve is the relative frequency of correct responses plotted against $\bar{R}$, the average individual estimate. The difference between the two curves is dramatic. Whereas for the most part, the relative frequency correct is lower than the estimated probability for the individuals, for the group, if $R \geq .7$, the group is "always right."

If we take the conventional interpretation of the individual realism curve, namely that individuals "overestimate" their information, then Figure 43 indicates that the group, defined as the average of the individual responses, drastically "underestimates" its information.

The average individual quadratic score for the Capen data is .47. The quadratic score for the "complete ignorance" estimate $R_i = .5$ is .5. Hence the average score for the individuals is worse than if each individual had answered every question by saying "I don't know." The best average individual score was .643. The average score for the group response was .67. For this data, the group score is better than the best individual score.

Non-concave scoring rules. For enterprises whose score rule is not concave, the n-heads rule must be weakened somewhat. We assume that

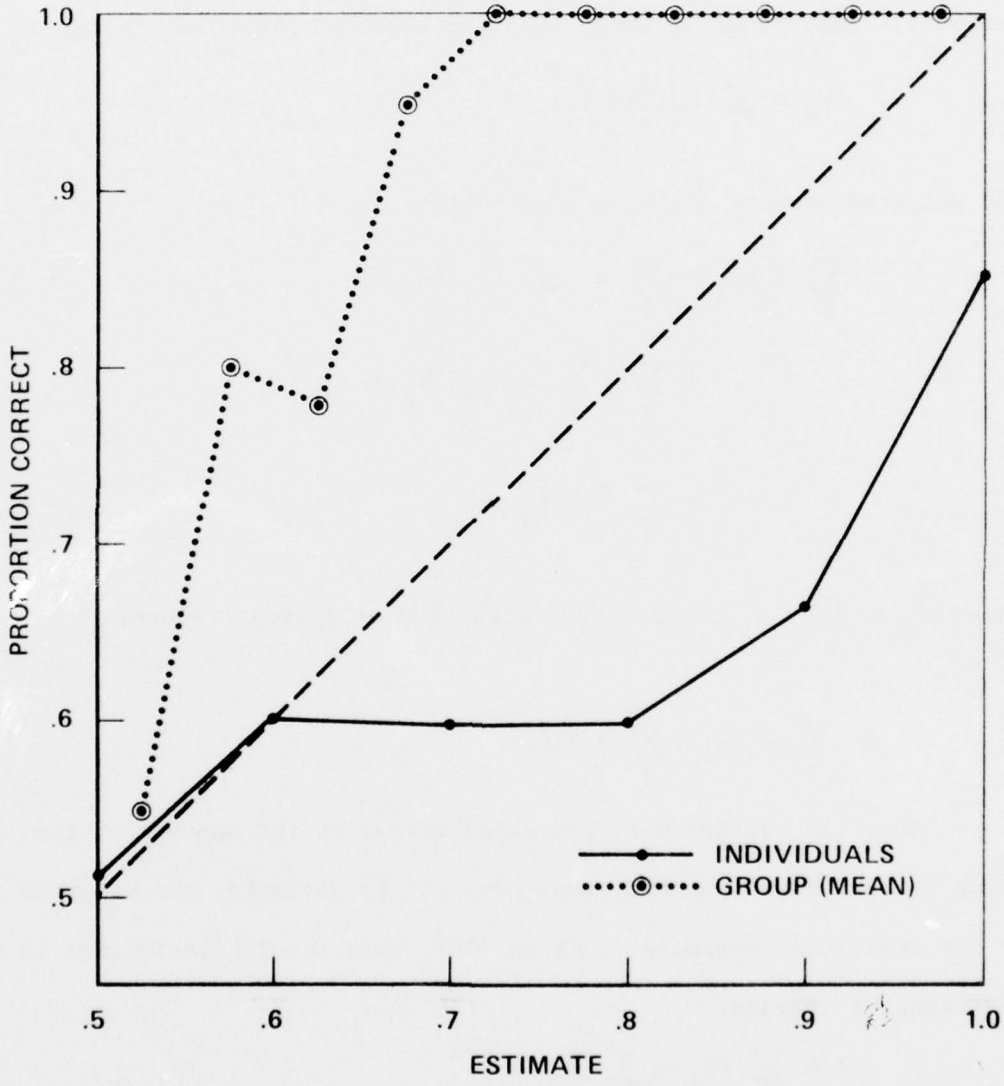


Figure 43. Individual and Group Calibration 43 Subjects, 120 Questions (Data from Capen)

the group decides to perform action A_g. There is no loss of generality in assuming that there is some estimate R_g for the probabilities for which A_g is optimal. Thus, selecting A_g is equivalent to selecting some R_g as the group estimate. Individual i, then, sees the expected group return as

$$EG_i = \sum_j R_{ij} S(R_g, j) \tag{31}$$

The weighted average of these expectations is

$$\sum_i a_i EG_i = \sum_i a_i \sum_j R_{ij} S(R_g, i) \tag{32}$$

$$= \sum_j \sum_i a_i R_{ij} S(R_g, j)$$

$$= \sum_j \bar{R}_j S(R_g, j) \tag{33}$$

where $\bar{R}_j = \sum_i a_i R_{ij}$. From the definition of a proper score,

$$\sum_i a_i EG_i \leq \sum_j \bar{R}_j S(\bar{R}, j) \tag{34}$$

Thus, the average weighted expectations of the members of the group is maximized when the group uses as its estimate, the weighted average of the individual estimates. As in (30), this result specializes to the non-weighted average.

$$1/n \sum_i EG_i \leq \sum_j \bar{R}_j S(\bar{R}, j) \tag{35}$$

where in this case

$$R_j = 1/n \sum_i R_{ij}.$$

Although (34) does not guarantee that the objective expectation of the group is maximized when the group acts in accordance with the average estimate,

it is still a fairly strong result. It states that independently of the form of the decision matrix, and the type of payoff involved (providing the payoff is linear in probabilities), the average estimate maximizes the average expectations of the individuals. Thus, if the enterprise is an economic unit, where the payoff is in money, and the weights represent a proportionate share of the return of the enterprise going to each individual, then average estimate maximizes the average proportionate expectation.

There are several other ways to express what is essentially the same result that clarify the import of (34). Suppose we examine the Monday morning quarterbacking situation where each individual is paid according to how the enterprise would have performed if it had followed his advice. The individual then has an expectation of

$$E_i = a_i \sum_j R_{ij} S(R_i, j)$$

Similarly, he has an expectation of the return for individual k of

$$E_{ik} = a_k \sum_j R_{ij} S(R_k, j)$$

Thus i's expectation of the return to the entire group is

$$\sum_k E_{ik} = \sum_k a_k \sum_j R_{ij} S(R_k, j)$$

The weighted average of these expectations is

$$\sum_i a_i \sum_k E_{ik} = \sum_i a_i \sum_k a_k \sum_j R_{ij} S(S_k, j) \qquad (36)$$

$$= \sum_k a_k \sum_j \bar{R}_j S(R_k, j)$$

Which again, by definition of a proper score

$$\leq \sum_k a_k \sum_j R_j S(\bar{R}, j) = \sum_j \bar{R}_j S(\bar{R}, j) \qquad (37)$$

In this disaggregated case, the average expectation of the total group return is maximized if each individual adopts the average estimate. For example, if the group consists of a loose confederation of "independent" operators, but each deal with the same basic decision situation, and, say, they agree to pool their earnings and redivide (e.g., a group of individuals betting separately on the same set of sports events, but pooling their earnings), then the average expectation of the group return is maximized if all use the same set of estimates — namely the weighted average — to make their "individual" decisions.

A straightforward corollary of (37) obtains if we reformulate the payoff in terms of regret. Define the regret of individual i as the difference between what he thinks the group can obtain using his estimate, and what he thinks will obtain using the group estimate. Then the weighted average estimate will minimize the average of the individuals' regrets.

This series of decisional n-heads rules shows that weighted or unweighted averages of the individual estimates do well compared to the average expected performance of the individuals. To do much better than this, it is necessary to take into account some additional properties of the group.

7. The Group as an Information System

In the opening section of this chapter, it was pointed out that, in theory at least, each member of a group can be conceived as possessing a certain stock of information, I_i, and a group estimation procedure can be thought of as a method of pooling that information to arrive at a collective answer to a question. A simple formal representation of this theory is to assume there is a probability function $P(E_j|I)$ which generates a probability distribution

on the event set $\{E_j\}$ given the vector of individual information sets $I = (I_1,\ldots,I_n)$. There is no difficulty in assuming that the events sets I_i are themselves partitions of the general universe of discourse U, and the group information set I is just the logical product of the individual information sets.

Although this model is formally well defined, it suffers from the fact that the I_i are not observable.*

One potential approach is to treat the information sets as "intervening variables;" i.e., to posit the existence of additional probability functions $P(R_i|I_i)$ which relate an individual's information to his report, and to formulate the group judgment, expressed say as $P(E_j|R.I)$, in terms of these probabilities. If carried out rigorously, this approach becomes quite complex.

It turns out that a theory can be generated which bypasses most of this complexity, and which is isomorphic to the theory that would ensue starting with the notion of information. The theory is generated by substituting the observable item, the individual report R_i, for I_i, and the group report R for I.**

It might be worth pointing out that this duality between information and reports (or more generally, between individual estimates and items of information) is more widely applicable than the use made of it in this section.

---

*There are other difficulties in practice — mainly in trying to characterize a universe of discourse that establishes a coherent structure for the miscellaneous material evoked by asking an uncertain question.

**The resultant formalism is similar in many respects to signal theory, where the individual reports R_i are treated as messages, or signals, and the group judgment is treated, as in signal theory, in terms of combining data from several "channels."

For example, the duality can be used to explore the value of augmenting individual estimates with additional information "fed in" during a group estimation process.

Most of the results of this section are of theoretical, rather than practical interest. However, they have some implications for practice. In particular, they establish certain "ideal" results which can be used to evaluate the effectiveness of practical aggregation techniques. For example, it will be shown that in theory the group is more accurate than the most accurate member. In order to achieve this desirable result, it is generally necessary to know more about the group than is possible in practice. However, knowing this result, there is reason to be discontented with procedures where the group does much more poorly than the most accurate member.

Suppose an individual announces a report R, i.e., he says "The probability of event E is R." We can treat R as a probability judgment, or we can treat it more cavalierly as a simple datum, and ask, "If individual i says R, what is the probability of event E?" The question assumes there is a probability function $P(E|R)$ which relates the report with the occurrences of E. Conceptually, R is not necessarily an assertion. It could consist of a nod of the head or a wave of the hand. However, since we want to apply the theory to the aggregation of probability statements, we will assume that the reports are probability statements.

The chief freedom allowed by treating reports as signals rather than as estimates is that the reports do not necessarily have to fulfill the probability postulates. Thus, initially at least, we do not run into the problems associated with calibration. Nor do we have to worry about consistency. Of course, if $P(E|R)$ were not roughly monotonic in R, we would

think that our estimator was a bit strange. But as we have seen in the discussion of counterprediction, for some questions and some individuals, $P(E|R)$ is, as a matter of fact, not monotonic in R.

In addition to the probability function $P(E|R)$, we assume there is a prior probability that the individual will report R. This prior probability is relative to the question being asked, which we will assume is simply to estimate the probability distribution on U, with a given partition E_j. Thus, there is a response space — all permissible probability distributions on U — and a probability distribution P(R) over these distributions. We can interpret P(R) as in the theory of errors as arising from "random error;" that is to say, there is some average R which the individual "aims at," but chance influences lead him to say something else. For the time being, we will assume that $P(E|R)$ is a function of the total report. To be more precise, in the general case the individual report consists of the probability assignment R_{ij}, where i denotes the individual and j denotes the event E_j. Let R_i stand for the set $\{R_{i1},\ldots,R_{im}\}$. Then, in general we allow the possibility that the probability for a given event, $P(E_j|R_i)$, depends on the entire report R_i. Otherwise, we would run into the problem of calibration, i.e., if $P(E_j|R_i) = F(R_{ij})$ then $P(E_j|R_j) = R_{ij}$, if the individual is consistent in his estimates.

In addition to the response space R_i and probability function $P(E_j|R_i)$ for individuals, we assume there is a joint response space $R = (R_1,\ldots,R_n)$ for a set of n individuals (the group), and a joint probability function $P(E_j|R)$, which expresses the probability that the event E_j will occur if the group says R. There is also a joint prior probability distribution P(R) on the group report. In the spirit of the theory of errors, we might assume that the joint distribution P(R) is simply the product of the individual distributions

$P(R_i)$; i.e., $P(R) = \prod_i P(R_i)$. If the individual distributions are assumed to be the result of "purely random" variations on the part of the individuals, this assumption would be reasonable. However, it will turn out that this assumption is not required for some of the most interesting consequences of the theory, so it will be postponed.

In some general sense, the function $P(E_j|R)$ is an aggregation function for the set of reports R. $P(E_j|R)$ is not necessarily a function of the reports R alone. In particular, it may reflect interactive effects among the reports, a topic that will be dealt with later in this section.

To complete the analysis, we assume that the aggregation problem arises within some context which will be labeled A (for a-priori), in which the responses R are generated. Based on whatever is known prior to the responses, there is an a-priori distribution on the events E_j which could be denoted by $P(E_j|A)$. However, since the context A is part of the "total scene," and the term A would appear as an antecedent in all probability expressions, we will suppress it. Thus for $P(E_j|A)$ we will write simply $P(E_j)$. Rather than $P(E_j|R.A)$ we write $P(E_j|R)$, etc.

Notice that the situation is entirely "objective." Some stimulus, e.g., asking a question, generates the responses R. The a-priori probabilities $P(R_i)$, $P(R)$ and $P(E_j)$, as well as the a-postereori probabilities $P(E_j|R_i)$ and $P(E_j|R)$ are assumed to be properties of the situation, and are not estimates. Of course, in order to apply the analysis, it is necessary to know these probabilities — but that is a different topic, to be pursued below.

Within this framework we can ask a number of pertinent questions: (1) How do the individual expected scores compare with the a-priori score? (2) How does the group score compare with the a-priori score? (3) How does the group score compare with any individual score? (4) What is the effect on the group score of adding new members to the group?

The expected score, based on the situation prior to the reports, is just $\sum_j P(E_j)S(P(E),j)$, where $P(E)$ represents the distribution $\{P(E_j)\}$. We can call this the a-priori score, AS. It is the score that would be expected if the probabilities $P(E_j)$ were taken as estimates.

To evaluate the individual scores, we consider

$$\sum_{R_i} P(R_i) \sum_j P(E_j|R_i)S(R_i,j) \tag{38}$$

Here we compute the total expected score, summing over all the possible responses of individual i. This could be called the before the fact expectation, i.e., it is the expectation before a specific R_i has been announced. Each term $\sum_j P(E_j|R_i)S(R_i,j)$ could be called an after the fact expected score since it is the score computed after the individual has announced R_i.

From the rule of elimination (F3, Chapter II), we have

$$P(E_j) = \sum_{R_i} P(R_i)P(E_j|R_i)$$

Substituting this in the expression for the a-priori score, we obtain

$$AS = \sum_{R_i} \sum_j P(R_i)P(E_j|R_i)\ S(P(E),j)$$

$$= \sum_{R_i} P(R_i) \sum_j P(E_j|R_i)\ S(P(E),j)$$

From the definition of a proper score, then,

$$AS \leq \sum_{R_i} P(R_i) \sum_j P(E_j|R_i)\ S(P_i,j) \tag{39}$$

where P_i is shorthand for the distribution $\{P(E_j|R_i)\}$.

(39) states that the total expected score for the estimate $\{P(E_j|R_i)\}$ is greater than the expected score a-priori. This is true before the fact — i.e., before the report R_i is announced. At first glance it might seem peculiar that the expected individual score is greater than the a-priori score, since, on one interpretation, the various reports R_i arise at random. The significant feature of the model, however, is that $P(E_j|R_i)$ is a function of R_i. Thus, once R_i has been announced, the probability of E_j changes.

(39) answers one of the questions raised earlier. The expected score for each individual based on the probability distributions $P(E_j|R_i)$ is greater than the a-priori score. Notice this is not the same thing as the expected score based directly on the reports. In general it will not be the case that $\sum_{R_i} P(R_i) \sum_j P(E_j|R_i)\ S(P_i,j) = \sum_{R_i} P(R_i) \sum_j P(E_j|R_i)\ S(R_i,j)$. And of course it is quite different from any subjective expectations the individual might have. The equality would occur only when the individual is completely realistic, i.e., when $P(E_j|R_i) = R_{ij}$.

Exactly the same line of reasoning that led to (39) can be used to demonstrate that

$$AS \leq \sum_R P(R) \sum_j P(E_j|R)S(P,j) \tag{40}$$

where P is shorthand for the distribution $\{P(E_j|R)\}$. That is, the average expected score for the group is always greater than or equal to the a-priori score, when the probability distribution $P(E_j|R)$ is assumed to be the report. Of course, the same comments concerning before the fact and after the fact hold for the group report as were made for the individual reports.

The third question how does the group score compare with any individual score? — can be answered by a similar line of reasoning, but with some additional notation. Let R_{-i} represent the vector of n-1 reports omitting R_i. From the rule of elimination we have

$$P(E_j.R_i) = \sum_{R_{-i}} P(E_j.R) \text{ with } R_i \text{ fixed}$$

$$= \sum_{R_{-i}} P(E_j|R)P(R)$$

and dividing both sides by $P(R_i)$

$$P(E_j|R_i) = \sum_{R_{-i}} P(E_j|R)P(R_{-i}|R_i) \tag{41}$$

Substituting from (41) in the right hand side of (40), we have

$$\sum_R P(R_i) \sum_j \sum_{R_{-i}} P(R_{-i}|R_i)\ P(E_j|R)\ S(P_i,j) \tag{42}$$

Rearranging, and noting that $P(R_i)P(R_{-i}|R) \approx P(R)$, we have

$$\sum_R P(R) \sum_j P(E_j|R)S(P_i,j) \leq \sum_R P(R) \sum_j P(E_j|R)S(P,j) \tag{43}$$

In short, the average expected score for the group is always greater than or equal to the average expected score of any member of the group.

The logic of this demonstration is actually the same as used to show that either the individual or the group has a higher averaged expected score than the a-priori score. The estimates of the additional members of the group act as a refinement of the estimate of any member of the group, and hence the average expected score of the total group is greater than that of any member.

The same sequence of steps can be inverted to show that the additional of a new member to a group never reduces the average expected score.

In a previous publication I stated that probabilistic aggregation is risky, in the sense that the expected score of the group can be either greater than or less than the a-priori score.[6] That statement was based on an analysis of after the fact scores, and was correct for that case. However, as the above derivation shows, the average expected score before the fact is not risky, in the sense that it is always greater than or equal to the a-priori score. Similar comments hold with regard to the comparison between the group and individuals.

This sequence of results puts in precise form a number of "obvious" features of group judgment. Since the group encompasses at least as much information as any member, theoretically, it should do at least as well as the best member. Similarly, if any new member is added, he cannot detract from the information already available to the group, and hence should not be counterproductive.

However, these statements hold only for the objective probabilities $P(E_j|R_i)$ and $P(E_j|R)$. They do not hold for the estimates R_i and any particular aggregation of R. In order to capitalize on (43), for example, it is necessary to know the function $P(E_j|R)$. By and large, it is not possible to compute $P(E_j|R)$, even if the $P(E_j|R_i)$ are known. It is even less feasible to compute $P(E_j|R)$ if only the R_i are known.

To explore this a little further, we can "unpack" $P(E_j|R)$ in terms of the $P(E_j|R_i)$ and some related probabilities.

For the moment, we will drop the subscript on E_j to streamline the notation. It will reappear when we reevaluate the expected group score.

We define two auxilliary notations

$$D_R = P(R) / \prod_i P(R_i) \tag{44}$$

D_R measures the degree of dependency among the individual reports R_i. $D_R = 1$ means that the reports are independent; the probability of a given conjunction is the product of the individual probabilities

$$D_R^E = P(R|E) / \prod_i P(R_i|E) \tag{45}$$

D_R^E measures the <u>event related dependency</u> among the reports. $D_R^E = 1$ means that, assuming the event E occurs, the probability of the joint report R is just the product of the probabilities of the individual reports.

Starting with the rule of the product

$$P(E|R) = P(R|E)P(E)/P(R)$$

and substituting for P(R) from (44) and for P(R|E) from (45)

$$P(E|R) = \frac{D_R^E \prod_i P(R_i|E)P(E)}{\prod_i P(R_i)\ D_R} \tag{46}$$

finally, invoking the rule of the product for the individual reports, we have

$$P(E|R) = \frac{D_R^E \prod_i P(E|R_i)}{P(E)^{n-1}\ D_R} \tag{47}$$

(47) displays the probability of interest, namely P(E|R), as a function of the individual probabilities $P(E|R_i)$, the a-priori probability P(E), and the two dependency terms.

From our previous discussions of the distribution P(R) it seems to be a reasonable assumption that $D_R = 1$, especially for a group process in which the individual reports are collected separately and anonymously. There is no simple way that I know to assess the event related dependency D_R^E.

Since (47) contains the product $\prod_i P(E|R_i)$, it is convenient to use the log score to assess the group performance.

Reinstating the subscripts for the events, we can compute the average expected log score AES_g for (47). Setting $D_R = 1$ for all R we have

$$AES_g = \sum_R P(R) \sum_j P(E_j|R) \log \frac{D_R^{E_j} \prod_i (E_j|R_i)}{P(E_j)^{n-1}} \tag{48}$$

Since the $P(E_j)$ terms do not involve R, expansion of (48) gives

$$AES_g = -(n-1)\ AS + \sum_R P(R) \sum_j P(E_j|R) \sum_i \log P(E_j|R_i) + D. \tag{49}$$

AS is the a-priori score defined earlier. D is the average expectation of the event related dependency $D_R^{E_j}$. Applying the same expansion as (42) to the central expression in (49), we arrive at

$$AES_g = -(n-1)AS + \sum_i \sum_{R_i} P(R_i) \sum_j P(E_j|R_i) \log P(E_j|R_i) + D. \tag{50}$$

Calling the average expectation of individual i, AES_i,

$$AES_g = -(n-1)AS + \sum_i AES_i + D. \tag{51}$$

Finally, it is convenient to introduce the notion of <u>net score</u>, namely the difference between the expected score of an individual or the group, and the a-priori score. The net score measures the improvement (or loss) due to employing the estimate rather than simply asserting the a-priori probabilities. Thus the net score of the group $NAES_g = AES_g - AS$, and the net score of an individual i is $NAES_i = AES_i - AS$.

From (51),

$$NAES_g = \sum_i NAES_i + D \tag{52}$$

The net score of the group is precisely the sum of the net scores of the individuals plus the expected dependency term.

The net expected score $NAES_i$ of each individual is positive. Hence, $\sum_i NAES_i$ is larger than the net expected score of any individual. However, D is not necessarily positive. We know from (40) that $NAES_g$ is positive, but it need not be as large as $\sum_i NAES_i$.

8. Approximations

In practice, it is rare that enough is known to apply formulas like (43) or (52). In particular, the event-related dependence D^E_{Rj} is difficult to express in terms of data that is likely to be available, and is "non-intuitive" when it comes to making a judgmental estimate. But in addition, the a-priori probabilities $P(E_j)$ are usually poorly known, as are the individual probabilities $P(E_j|R_i)$. Often, about all that can be said concerning the individual probabilities is something like "i is a good man in his field;" which is a long way from determining the probability that a given event will occur if i says R_i.

As we saw in the preceding chapter, a common approach given such a dearth of information, is to rely on some "plausible" or nominal assumptions. The assumption $D_R = 1$ is plausible, based on the assumption that much of the variation in an individual's report is due to "random" influences. The assumption is reinforced if the responses of the individuals are anonymous, and hence presumably independent. This particular assumption can be sidestepped; it is possible to reformulate (47) in a way that does not involve D_R. Since $\sum_j P(E_j|R) = 1$, we can divide (47) by $\sum_j P(E_j|R)$ to obtain

$$P(E_j|R) = \frac{\prod_i P(E_j|R_i)}{\sum_k D_{jk} \prod_i P(E_k|R_i)} \qquad (53)$$

where

$$D_{jk} = \frac{D_R^{E_k} \, P(E_j)^{n-1}}{D_R^{E_j} \, P(E_k)^{n-1}}$$

D_{jk} is a kind of relative dependency term, no easier to estimate than $D_R^{E_j}$. At all events, (53) does not contain the D_R term.

The traditional assumption of equal a-priori probabilities would seem to have as much justification in group estimation as it does in more conventional statistical inference. An additional tempting assumption is $D_R^{E_j} = 1$. Coupled with the assumption of equal a-priori probabilities, $D_R^{E_j} = 1$ implies

$$P(E_j|R) = \frac{\prod_i P(E_j|R_i)}{\sum_k \prod_i P(E_k|R_i)} \tag{54}$$

Finally, if we assume that all the individuals are realistic, i.e., $P(E_j|R_i) = R_{ij}$, (54) becomes

$$P(E_j|R) = \frac{\prod_i R_{ij}}{\sum_k \prod_i R_{ik}} \tag{55}$$

(54) and (55) are pleasantly simple formulae. If the assumptions leading to them could be justified, the aggregation problem would be well in hand. There are reasons for thinking (54) and (55) are oversimplified.

For the two event case — i.e., the case where the event space consists of an event E and its complement $\bar{E}$ — the joint assumption $D_R = D_R^{E_j} = 1$ leads to the consequence that $P(E|R_i) = P(E)$ for all R_i but one.* In short, for all but one of the respondents, their estimates add no new information beyond the a priori information.

A related difficulty is that for large groups, assuming (55) implies that almost all group estimates will be essentially 0 or 1. Thus for a group with thirty members, if the average response is .55 or greater for one

*The demonstration is given in Appendix III.

alternative, $P(E|R) \geq .998$; if the average response is greater than .6, $P(E|R) \geq .99999$.* Since it seems unlikely that for questions of the sort where group estimation is appropriate, the group knows enough to justify estimates of 0 or 1 for almost all questions, the independence assumptions are probably too optimistic for large groups.

A potential compromise is the geometric mean. The geometric mean retains the multiplicative character of (48), but is less "extreme" than the normalized product. It is also, in a way, a compromise with respect to the impossibility result derived in Section 5. There it was shown that the only function of a set of probabilities which is multiplicative is a weighted product with weights adding up to 1. However, a weighted product would not add up to 1 for exclusive events. Normalizing a weighted product to make it sum to 1 produces a generalization of the normalized geometric mean; the normalized geometric mean is, in fact, the normalized weighted product with uniform weights. Specifically, the assumption is

$$P(E_j|R) = \frac{\left[\prod_i R_{ij}\right]^{1/n}}{\sum_k \left[\prod_i R_{ij}\right]^{1/n}} \tag{56}$$

Some weak additional justification for trying the geometric mean comes from the likelihood that the prior distribution of responses $P(R_i)$ is skewed due to the constraint that R_i is between 0 and 1. A glance at Fig. 18

*The product formulation has a related "edge effect". If one of the responses is zero for a given alternative, then the group response is zero, independently of the other responses. If one individual reports zero for one alternative, and some other individual reports zero for its complement, then the product approximation is completely degenerate; both probabilities are zero. This edge effect can be dealt with by a suitable truncation; however, the results will be highly sensitive to the nature of the truncation, especially for large groups.

Chapter II illustrates this point. If the prior distributions are roughly log normal, then the theory of errors would suggest that the geometric mean is an appropriate aggregation function.

The geometric mean has a particularly straight forward n-heads rule using the logarithmic score as a figure of merit. We have

$$OES_g = \sum_j P_j \log \frac{\left[\prod_i R_{ij}\right]^{1/n}}{\sum_k \left[\prod_i R_{ik}\right]^{1/n}}$$

$$= \sum_j P_j \; 1/n \sum_i \log R_{ij} + C \qquad (57)$$

where $C = -\log \sum_k \left[\prod_i R_{ik}\right]^{1/n}$. Since C is not a function of j, it is a constant, depending only on the R_{ik}. Rearranging (57), we have

$$OES_g = 1/n \sum_i OES_i + C \qquad (58)$$

In words, the objective expected score of the geometric mean is the average of the individual objective scores plus a constant term. Since $\sum_k \prod_i R_{ik} \leq 1$, its logarithm is negative, and C is positive. Thus the advantage of the group score over the average of the individual scores is independent of the objective probability and depends only on the amount of disagreement within the group.* It can be computed immediately knowing the R_{ij}. C = 0 if and only if all the individual reports are the same.

In Figure 44 a subset of the Capen data is plotted, comparing the performance of the mean and the performance of the geometric mean for 18

*This feature is manifested even more clearly for the comparable n-heads rule for the quadratic score and the mean as the aggregation function. In this case $OES_g = 1/n \sum_i OES_i + \sum_j S_j^2$, where S_j^2 is the variance of the individual reports for event j.

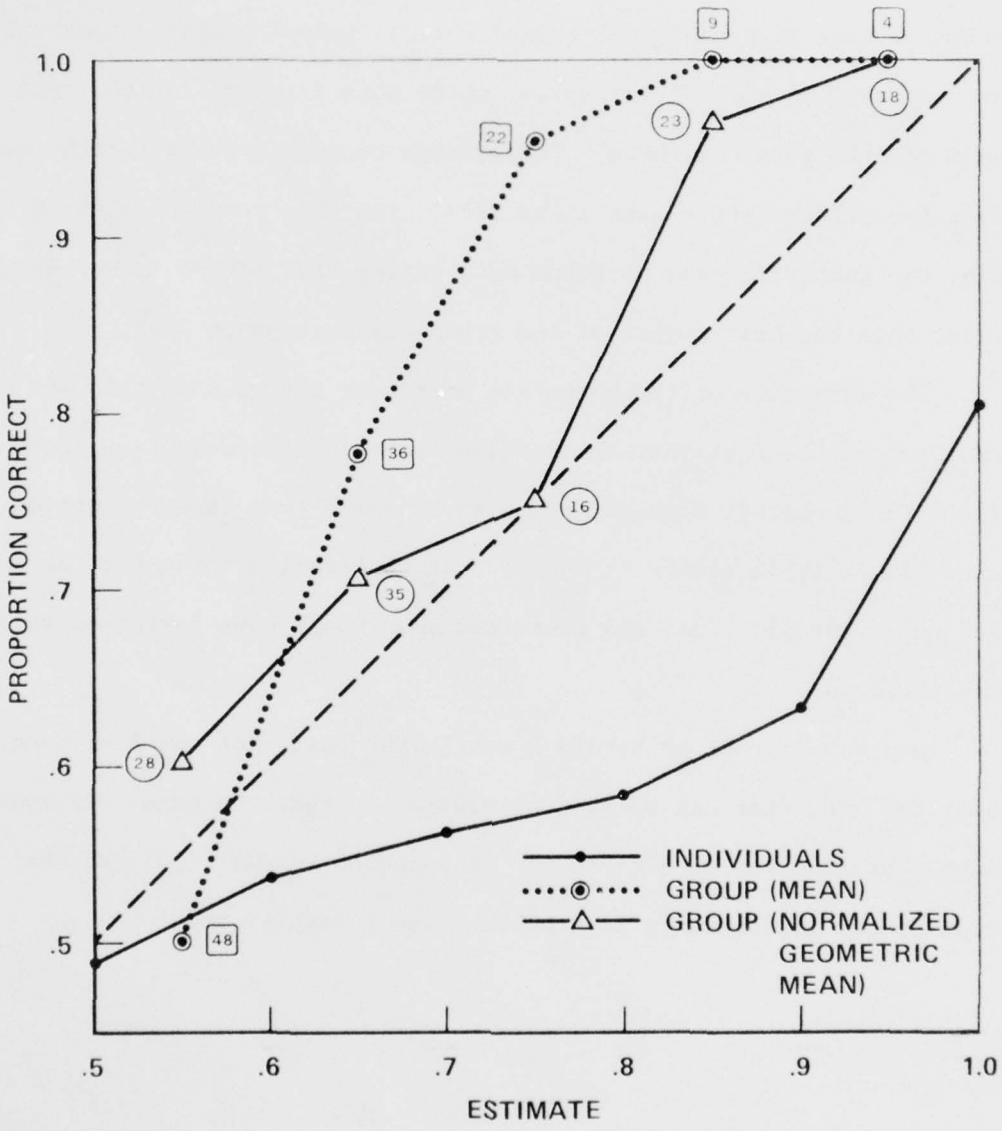


Figure 44. Individual and Group Calibration n = 18, 120 Questions
(Data from Capen)

subjects on the 120 questions.* Two features of the realism curve for the geometric mean are noteworthy: a) The curve is closer to the 45° (fully realistic) line than the corresponding curve for the mean. b) The group estimates have been displaced upward — i.e., toward higher estimates. Notice that for the .85 and .95 estimates, there were 13 cases for the mean and 41 cases for the geometric mean. The average quadratic score for the mean is .646; for the geometric mean it is .704. For this particular set of data, then, the geometric mean performs much better than the mean, and decidely better than the best member of the group (average score .63).

The advantage of the geometric mean over the mean in this set of data results from the fact that the realism curve for the mean lies above the 45° line. The geometric mean generates an estimate that is more extreme than the mean — i.e., it is closer to 0 or 1. If the realism curve for the mean had been below the 45° line, the geometric mean would have performed more poorly than the mean.

One set of data is hardly a sufficient basis for any firm conclusions. About the most that can be said at present is that the geometric mean is a rough approximation to the "ideal" aggregation formula (53) and that it gives surprisingly good results for the one case investigated.

---

*The selection of a subset of 18 subjects for this analysis was accidental. The 18 subjects were members of a UCLA Executive program for mid-career engineers.

CHAPTER VI. GROUP VALUES

1. Individual Utilities

The aggregation of value judgments presents an entirely different conceptual problem from the aggregation of factual judgments. For factual judgments we have the simplifying feature that figures of merit are the same for individual estimates as they are for group estimates. Thus, even in the face of logical difficulties, such as possible inconsistencies between individual and group estimates, it seems reasonable to prefer a group judgment over an individual judgment if the former is likely to be more accurate. In the present state of the art there is no comparable criterion for group value judgements. The question whether it is meaningful to speak of figures of merit for individual value judgements is still somewhat controversial. But even if the notion of figure of merit for individual value judgments was sharply defined, the same figure of merit would not apply to group judgments, except for some specialized cases such as the fixed share partnership.

The major emphasis in this chapter is on group values; but some attention must be paid to individual values as inputs to group decisions. Most of the conceptual apparatus needed has already been presented in Chapter II in the theory of probability estimates. In fact, it is only a small step from postulates P1-P8 of that theory to the theory of individual numerical value, usually called the theory of utility.

A major stumbling block in the theory of individual values is the lack of a well-defined figure of merit. Many decision theorists either implicitly or explicitly adopt the view that a figure of merit can be based on the tie between estimates of value and choice behavior. Thus, an estimate of the form "A is better than B" is considered correct (for a given individual) if, when presented with a free choice between A and B, the individual selects A

rather than B. This approach has led many economists to insist that the only "true" meaning to value statements is the correlative choice behavior. Hence, individuals should not be asked what their preferences are, but rather, the preferences should be deduced from their choices. One variant of this attitude is the doctrine of revealed preferences--individuals express their value judgments most directly in their market behavior.

Several things are scrambled in the revealed preference approach. As with all estimates, assuming there is a figure of merit, judgments concerning preferences are subject to error. Hence, statements of preference should be treated with the same caution as any other kind of estimate. On the other hand, choices such as market behavior are complex phenomena with cognitive elements playing a role. If a choice reflects not only "pure preference" but other types of estimates, such as estimates of probabilities, then a choice can be "mistaken" if the ancillary estimates are incorrect. Hence, it is not clear that choice behavior, especially market behavior, is always a reliable source of figures of merit for value judgments.

Despite these caveats, the notion that choice behavior is the "proper" objective correlate of value judgments has a great deal to recommend it. Other attempts to define a correlate--e.g., internal states of the individual or feeling tones--have not reached a level of precision that would allow useful figures of merit.* In the following, then, it will be assumed that choice is the most useful concommitant of preferences for decision theory.

---

*It is possible that a quite different mechanism, namely the phenomenon of reinforcement, could furnish a more diagnostic approach for objectifying value judgments. A characteristic of a situation which reinforces behavior (increases the probability of an associated act) might be considered a value in that situation. However, so far as I know, reinforcement has not as yet been used as the basis for a theory of decision.

To recapitulate some of the material in Chapter II, we assume there is a set X of situations, among which are contingencies of the form $(x_i|E_i)$, and the individual has a complete preference relation over X. The preference relation obeys the principles of dominance, stability, and sure-thing for contingencies. There is at least one event with probability 1/2, independent on repetition, and the set of events generated by repetitions of this event are archimedean. These assumptions lead to the consequence that there is a numerical scale of probabilities on the events that is additive for exclusive events.

For the purpose of introducing numerical utilities, it is convenient to modify the archimedian axiom, P8, to

P8'. If $x > y > z$ then there is an event E such that

$$y \sim (x,z|E)^*$$

P8' states that if y is intermediate in preference between x and z, then there is some contingency involving only x and z which is equivalent to it. This axiom is usually stated in the form: given the hypothesis, there is some probability p, such that the contingency x with probability p, z with probability 1-p is equivalent to y. What is usually left unstated in the axiom in this form is that the probability p is generated by a random device, independent on repitition. Armed with Theorem 7, Chapter II, it is not necessary to deal with contingencies of this restricted form.

P8' generates a mapping U(x) of the set X onto the real numbers. This is accomplished as follows: Choose any pair of situations $\underline{x}$ and $\underline{y}$, where $\underline{x} > \underline{y}$. Set $U(\underline{x}) = 1$, $U(\underline{y}) = 0$. For any z, if $\underline{x} > z > \underline{y}$, $U(z) = P(E)$, where

*With a slight modification of the combining axioms P3, P8' would do as well as P8 for completing the numerical theory of probability; however, the definition of numerical probabilities would be somewhat more complicated.

$z \sim (\underline{x},\underline{y}|E)$. If $z > \underline{x}$, then $U(z) = 1/P(E)$ where $\underline{x} = (z,\underline{y}|E)$. Finally, if $\underline{x} > \underline{y} > z$, $U(z) = P(E)/\left(P(E)-1\right)$, where $\underline{y} = (\underline{x},z|E)$.

It is straightforward, but tedious, to prove the following theorem:

THEOREM 1. (Von Neumann-Morgenstern):[1] The mapping $U(x)$ has the properties:

1. $x \geq y$ if and only if $U(x) \geq U(y)$
2. $U((x,y|E)) = P(E)U(x) + (1-P(E))\ U(y)$
3. If $z > w$, then the mapping $U'(x)$ based on z and w is related to the mapping based on $\underline{x}$, $\underline{y}$ by $U'(x) = aU(x) + b$, where a is a positive constant.

The utility scale characterized by Theorem 1 completes an intellectually satisfying theory of individual decisions for contingencies. Furthermore, it defines an implementable procedure for establishing utility measurements, namely, the mapping procedure outlined above. The process can be tedious if the probability is determined by a sequence of successive approximations. However, the process is relatively rapid if direct estimation of the probabilities is employed.

Probabilistic scaling is not the only way to establish a numerical scale for preferences. As Suppes and others have shown, if the individual can compare in a consistent fashion the differences in value between pairs of objects, scales can be generated that are also determined up to a linear transformation, and that need not be the same as the scales established by Theorem 1.[2] However, if the scales established by comparing differences are not the same as those found by comparing contingencies, then the former will not be linear in probabilities, i.e., property 2 in Theorem 1 will not hold. These comments have no bearing on which type of scale is "right." Nevertheless, the usefulness of scales which are linear in probabilities is so overwhelming that it would require a dramatic solution to some basic problem to make the pursuit of other forms of scaling of more than academic interest.

One scale which is certainly of more than academic interest is money. Numerous theoretical discussions, and some empirical investigation, have made it plausible that the utility of money is not linear in probabilities. Assuming that the utility of money is concave "explains" many types of risk aversive choices, as well as the purchase of insurance at a premium that is actuarily excessive (i.e., the expected value of the insurance contract in money is smaller than the cost.) Other assumptions about the shape of the utility for money curve can "explain" gambling behavior,[3] and, as we saw in Chapter IV, can resolve certain paradoxical choice phenomena such as the Allais puzzle. Some decision analysts appear to interpret these results as implying that the value of money is not "really" linear, i.e., that the value of money in some undefined absolute sense exhibits "decreasing returns to scale." "$1,000 is worth less to a millionaire than it is to a pauper." Statements of this sort have little or no meaning until the measure of worth is specified. The statement about the millionaire and the pauper appears to be true, if worth is measured by utilities established by probability scaling (i.e., by choices among contingencies.) However, the statement is false if worth is measured by what the money will buy. A pauper can buy no more shares of a given stock with $1,000 than the millionaire. Money is not linear in probabilities, but it is linear in many significant commodities, using exchange as the measuring process.

There is another point about money that is directly relevant to the issue of group values, namely money has a kind of intersubjectivity that individual utility does not possess. Over a wide range of transactions, money has the same exchange value for all members of society, and for groups as well as individuals. Thus, we have a model for a value scale which is equally valid for groups and individuals.

The hypothetical individual for which the decision postulates P1-P8' hold is not identified by the postulates. The consequences--the existence of probability and utility scales--hold for any entity that fulfills P1-P8'. Thus, if a group of individuals collectively can be said to have a complete order of perference over some class of situations X, and the other postulates hold, then that group has a collective probability scale and a collective utility scale. Since it is clear that groups do make decisions (i.e., exhibit choice behavior), there is no a priori reason why P1-P8' should not characterize these choices. The problem in formulating group value scales is not P1-P8', but rather the relationship between group and individual choice. This is the subject for the next section.

2. The Arrow Impossibility Theorem

In the Introduction I described a class of paradoxes which can arise when an attempt is made to define an aggregation function for a set of individual judgments, where that aggregation is intended to be consistent in some way with the individual judgments. Probably the most significant instance of this type of difficulty is the theorem due to Kenneth Arrow that there is no group preference function that fulfills a set of desirable and apparently innocuous conditions.

This theorem has played a major role in recent investigations in welfare economics and decision theory. On the one hand, it has motivated a large activity concerned with "resolving" the problem, and on the other hand it has acted as a restraint on the development of techniques for generating group preference functions in many areas such as voting methods, social value scales, and group decision procedures.

In the following sections I present a resolution of the Arrow theorem in what appears to be a reasonable sense of that term. Since the theorem is

correct, there is no resolution in a strict sense; however, if it can be shown that the conditions assumed by the theorem are, in fact, more severe than one would want to accept, and if it can be shown that a minor relaxation of the conditions leads to a set that is consistent, this appears to be a justification for the term "resolution."

The formal elements of a group preference function are: a set $I = \{i, j, k, \ldots\}$ of individual members of a group; a set $X = \{x, y, z, \ldots\}$ of objects; a set $K = \{R, \underline{R}, R', \ldots\}$ of vectors of individual preference relations (that is, each $R = (R_1, R_2, \ldots, R_n)$ in K is an indexed set of preference relations over X, where the indices correspond to the members of I); a function F(R) which maps each member of K onto a relation (group preference relation) over X.* A superfix arrow indicates strict preference, i.e., $x\vec{R}y$ means xRy and not yRx.

If F is intended to define a general social welfare function, then X would be interpreted as a set of potential states of society. However, the formal treatment of the problem is not concerned with the nature of the elements of X, and for simplicity they will be referred to as objects. Similarly, the fact that the individual relations R_i and the group relation F(R) are preference relations is not part of the formalism. The analysis is concerned with the existence or not of an aggregation function F which takes a vector of relations as its arguments, and which fulfills a set of conditions that look reasonable for a group preference function, but might equally be appropriate for a wide variety of aggregation procedures--e.g., the R_i might be a set of

*In Arrow's formalism, the set I is expressed by the indices on the individual preference relations, X is expressed implicitly as the field of the individual preference relations, and the set K is characterized as a set of admissable preference relations, where admissable is taken to mean a set "for which the social welfare function defines a corresponding social ordering."[4] The implicit nature of these entities leads to some minor ambiguities;[5] however, since these do not appear to affect the central possibility theorem, they will not be pursued here.

individual rank orderings of a set of objects on some psychological magnitude, and F(R) a "representative ordering" for the group. The only issue is the reasonableness of the stated conditions for the intended application. By and large, the conditions proposed by Arrow appear to be reasonable for a wide class of "representative" aggregation functions.

In this section, the conditions proposed by Arrow and the impossibility theorem are stated for reference purposes. In the following sections the resolution of the theorem is taken up. Except for A1 and A2, the numbering is kept consistent with Arrow's. Some minor notational differences are introduced, mainly to simplify the translation of the conditions to the corresponding ones for scales in the following sections.

It is convenient to have an additional piece of notation. Let T^B, where B is a class of objects and T a relation, designate the relation T restricted to the class B; that is, B is in X, and xT^By if and only if x and y are in B and xTy. T^{-x} will be used to designate the relation T restricted to the set $X - \{x\}$.

A1. For every R in K and every i, R_i is a weak ordering on X.

A2. For every R in K, F(R) is a weak ordering on X.

A1 and A2 simply assert that the individual relations and the group relation are weak orders on X. The next set of conditions define additional properties of F(R).

C1. There is a set S in X, such that S contains three members, and for any possible vector of orderings T of S, there is an R in K such that $T = R^S$. This condition is intended to assure that whatever the nature of K, it is possible to find at least three objects for which all possible orderings for n individuals are exemplified by some members of K. As Arrow remarks, the basic theorem is essentially demonstrated for this set of three objects.

The next condition will be introduced by a definition: $\underline{R}$ will be said to be a forward shift of x, with respect to R, $FS(\underline{R},x,R)$, when, $\underline{R}^{-x}=R^{-x}$, and for every i and y, if xR_iy then $x\underline{R}_iy$ and for every i and y if $x\overrightarrow{R}_iy$ then $x\overrightarrow{\underline{R}}_iy$. That is, R and $\underline{R}$ are identical except for x, and whatever location x has in R_i, it is at least as "high" in $\underline{R}_i$.

C2. If $FS(\underline{R},x,R)$ then $x\overrightarrow{F(R)}y$ implies $x\overrightarrow{F(\underline{R})}y$. Arrow calls this condition "positive association of social and individual values." For the next condition, a further notion is needed. Let C(S,T) designate the set of x in S, such that for every y in S, xTy. That is, C(S,T) is the set of maximal elements in S with respect to the relation T. If S has no maximal elements with respect to T (e.g., if S is the set of all real numbers less than 1, and T is "greater than") then C(S,T) is the null set. Arrow makes no provision for this case, but there is no loss, since he is concerned primarily with the finite special set S which does have maximal elements, for every T.

C3. For every R and $\underline{R}$, $R^B=\underline{R}^B$ implies $C(B,F(R)) = C(B,F(\underline{R}))$. This axiom, called the independence of irrelevant alternatives, is perhaps the key condition in the derivation of the impossibility theorem. It essentially has the effect that the social preference between any pair x and y will depend only on the individual preferences for that pair.

C4. For every x and y in X, there is an R in K, such that not $xF(R)y$. This condition, called the condition of Citizen's Sovereignty, (or also, the condition that the group preference is not imposed) is also a key condition for the impossibility theorem. It posits a decoupling between the admissable set K and the set of objects X. For example, it rules out the possibility that there exists in X one object which is preferred by everybody to some other object for all admissable orderings R. Although intended

only to make sure that the individual preference relations in fact determine the group preference relation, it asserts something stronger.

C5. For every i there exists x, y, and R such that xF(R)y and not $xR_i y$.

This condition, non-dictatorship, asserts that for any individual there is at least one pair of objects and a set of orderings for the other individuals which generates a group preference contrary to the given individual.

Given the preceeding conditions Arrow proves the theorem.

THEOREM 2 (Arrow): There is no function F with the listed properties.*

The proof is somewhat extensive and will not be reproduced here, since the primary purpose of listing the conditions is to show that a small modification of them will enable the demonstration of a "possibility theorem."

3. Digression on Measurement Theory

In discussions of measurement theory as applied to psychological and social magnitudes, a great deal of attention has been paid to the degree to which a scale is prescribed by a set of measurement rules. A classification of scales based on such rules was presented in Chapter II, section 2.

Although this transformational approach to measurement is extremely useful for many theoretical investigations, it tends to obscure some of the basic features of measuring scales as applied in practice. In particular, it deemphasizes the role of reference objects for practical scales, and, in fact, often subtly downgrades these by referring to them as "arbitrary constants." It is true that within elementary thermometrics, the zero-point and 100-point determinations are "arbitrary" but that does not mean they are

*Arrow states the theorem in what appears to be a more restricted form, namely a social welfare function satisfying conditions 1-3 and A1 and A2, is either dictatorial or conventional, but any combination of less than all the conditions could be selected and the assertion made that any welfare function satisfying them must violate the remaining ones.

dispensable. If a doctor is told that the thermometer reading of a patient is 46, he knows nothing about the temperature of the patient until he knows the specific scale on which it is measured, i.e., until he knows two physical conditions like the freezing point and boiling point of water which are associated with two locations on the given scale.

A similar requirement holds for ordinal scales if these are to be used for indirect comparison of objects. This feature of ordinal scales seems to have been overlooked by measurement theorists in the social sciences. Thus, it is common to find "x-point scales" (e.g., a 5-point scale like: 5-very pleasant, 4-somewhat pleasant, 3-neutral, 2-somewhat unpleasant, 1-very unpleasant,) with no definite reference objects at all. The question whether such scales "measure anything" is more profound than the question whether it is "legitimate" to perform arithmetical operations with the assigned numbers. More basically, the issue is whether the estimates by subjects have the requisite stability to assert, e.g., that the class of "very pleasant" objects is defined at all.

In the physical sciences, ordinal scales have been used in many different fields. A well-known physical ordinal scale is the Moh's hardness scale. This scale is based on the relation "scratches." One substance x is considered harder than another y if x scratches y. (This is the basis of the familiar test whether a gem is real or "paste" by seeing if it will scratch glass.) Although the relation "scratches" is well defined, it is of little utility to engineers by itself. It becomes useful when it is augmented to a scale by the addition of a set of reference objects. In the case of the Moh's hardness scale, there is a set of 10 reference substances--Diamond, sapphire, topaz, quartz, feldspar, apatite, flourite, calcite, gypsum, and talc. (Window glass has a hardness of 5.5 on this

is scratched by topaz, it has a hardness between 7 and 8.

The point is that if you know the location of two substances in the scale, you know their relative hardness without a direct comparison.

In essence, the Arrow requirements on a group preference function demand that the function be created without the stability of reference points. This is done, ostensibly, to rule out the necessity for interpersonal comparison of utilities, and to maintain the ordinal nature of the group preference. The difficulty lies, then, not in any of the specific conditions which Arrow requires, but in the informal contextual framework, in which the group preference function is required to be a function of the individual preference relations only. As we shall see below, if this stringent requirement is relaxed slightly to allow the group preferences to be a function of both individual preferences and individual reference objects, the difficulty disappears.

It could be asked whether including reference objects in the group function does not bring interpersonal comparisons of utility in via the back door. For myself, I have no strong objections to interpersonal comparisons of utility. One of the strongest objections is that it cannot be done, that the "strength of preferences" is a subjective, non-observable quantity. That particular objection disappears if each individual has a set of reference objects which are, so to speak, intersubjective. Interpersonal comparisons on those reference objects are clearly feasible, and logically unobjectionable. There is no requirement that such comparisons be couched in terms of the strength of feelings about object x to individual i, versus the strength of feelings of individual j.

Some of this discussion is proleptic, since the way in which reference objects will be used to construct a consistent group preference function has not yet been introduced. However, for the time being, no metaphysical entities are involved in establishing individual preference scales with reference objects, and similarly, none with group scales based on the individual reference objects, provided the latter are observable to the total group.

4. Group Anchored Scales

We turn to the construction of group preference functions, based on the notion of anchored scales.

Def. 1: An individual anchored scale S_i consists of a set of objects X, a weak ordering relation R_i on X, and a designated subset A_i of X. The scale value, $S_i(x)$ of an object in X is defined by the rule:

$S_i(x) = a$ means a is in A_i, a $R_i x$, and for every $b \neq a$ in A_i, aR_ib implies xR_ib.

This definition can probably be most conveniently explicated by a diagram, where $A_i = \{a, b, c, d\}$

$S_i(x) = a$ means x is in the interval between b and a. It is convenient, but not necessary, that the members of A_i be strongly ordered by R_i, i.e., if a, b are in A_i and $a \neq b$, then either $a\vec{R}_ib$ or $b\vec{R}_ia$. It is also convenient but not necessary that A_i contain the maximal element in X with respect to R_i, if there is one. For the case that A_i does not contain the maximal element of X (as illustrated by y in the diagram) an additional value must be defined for objects above the maximal object in A_i, say $\underline{m}$. Thus $S_i(y) = m$ means yR_ia. (If there is no maximal element, but a minimal element, then

for some purposes, it may be convenient to switch the definition of the interval, and call $S_i(x)$ the next lowest member of A_i. However, this leads to unnecessary complications with constructing a group scale if not all of the individual ordering relations have minimal elements.)

In the examples used to illustrate constructions, the A_i's are treated as if they were finite. This, again, is not necessary, but simplifies much of the discussion. It does not seem meaningful to speak of an infinite set of reference objects unless they are generated by some mathematical rule, and in a sense, this negates the basic notion of reference object. However, there is no logical difficulty in allowing infinite sets.

The scale S_i can be interpreted as inducing a new ordering relation on X which can be designated R_i^*, where xR_i^*y means $S_i(x)\ R_i\ S_i(y)$.

Def. 2: A group anchored scale F(S) consists of a set X, a set A, and a weak ordering relation G on A. S for this definition, is a vector of individual anchored scales, i.e., $S = (S_1, S_2, \ldots, S_n)$. A is the Cartesian product of the anchor sets of the individual scales, i.e., $A = A_1 X A_2, \ldots X A_n$ (The Cartesian product is the set of all n-tuples that can be formed by picking one object from each of the n individual anchor sets. This is illustrated in Fig. 45.) Since A is the Cartesian product of individual anchor sets, A is not a member of X. There are several different ways in which the relation G can be extended to X to produce an anchored scale on X. The simplest would appear to be to extend it to X with the definition

$$x\ G\ y \text{ means } S(x)\ G\ S(y)$$

Strictly speaking, G in the expression x G y is a different relation than in a G b, where a and b are in A; however, the distinction is sufficiently expressed by the difference in notation for objects in A and other objects.

The following diagram illustrates what is going on. Suppose there are two individuals, $A_1 = \{a,b,c,d\}$ and $A_2 = \{e,d,c\}$ (There is no logical relationship between the anchor sets of different individuals. They can be identical, overlap somewhat, or be entirely distinct. In practice, of course, there would be many advantages to having a common set of reference objects.)

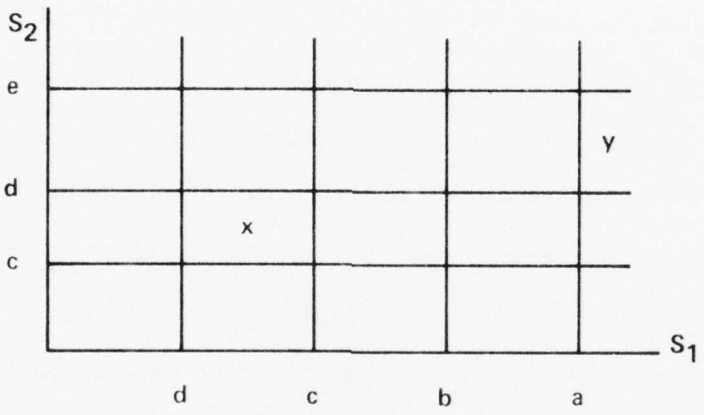


Figure 45. Group Ordinal Scale

The n-tuples of individual reference objects determine a set of n-dimensional "boxes." The group scale value S(x) is determined by the box in which x lies. Illustrated is the case S(x) = (c,d), that is, $S_1(x) = c$, $S_2(x) = d$.

G (not illustrated for reasons that will be clear shortly) orders the n-tuples. In the diagram, this means G orders the boxes. In this respect, the "edge" boxes like the one containing y in the diagram are treated like all the others.

A key step in constructing a group anchored scale consists in postulating the set of admissable individual anchored scales as follows: The set K = (S, S, S',...) of admissable anchored scales, for a group scale F, consists of all n-vectors of individual anchored scales $S = (S_1, S_2, \ldots S_n)$

such that A_i is identical for all S_i. That is, for all members of K and all individuals, the set of reference objects remains fixed, and a R_i b if and only if a $\underset{-i}{R}$b for all a,b in A_i, and all R_i, $\underset{-i}{R}$ in K.

Intuitively, what is involved should be fairly clear. Each individual selects a set of reference objects. These he orders according to his own preferences, and it is assumed that for the duration of the given group preference function, he does not change his preferences for this special set. The group function first addresses the various combinations of the individual reference objects and imposes an order on them. For any objects not in the reference sets of the individuals, the group preference relation is extended in the obvious way, by ordering them in the same way as the "regions" in which they fit as determined by the individual scales.

This leads to a slightly anomalous situation with respect to objects that belong to some but not all of the individual anchor sets. These cases have been dealt with below by excluding from the conditions imposed on the group function any **x**'s which belong to any of the A_i's. This is perhaps a little heavy handed; however, it saves expressing the complex of exceptions that would be needed if the partial members of reference sets should be included in the conditions. This approach does not appear to violate the spirit of a general welfare function, since we are primarily concerned with removing the difficulties with those objects for which all possible preference orderings are allowed. In the special case that the individual reference sets are identical, the blanket exclusion appears just right.

It perhaps should be pointed out that the definition of a group anchored scale contains an implicit assumption, namely that the group is indifferent

between all objects x, y such that $S(x) = S(y)$. This can be softened somewhat if we assume that the selection of a scale by an individual is equivalent to the assumption that he considers--for the purposes of group aggregation--x equivalent to y if $S(x) = S(y)$, and we add the assumption that the group is indifferent between x and y if all members of the group are indifferent between x and y. Strictly speaking, these comments are not part of the formal possibility theorem that is being sought. However, it is clear that for any judgment as to whether the approach is reasonable or not, the question whether an individual can generate a fine enough scale to make him accept a group aggregation is relevant.

Nothing in the preceeding determines in any way the number of members of the A_i's. Clearly, if any A_i is empty, then individual i is a dummy, and has no influence on the group preference. However, if there is only one member of A_i, the individual can discriminate to the extent of saying whether a given object is better or worse than a. Condition 1 below requires that each A_i have at least two members.

We now proceed to transliterate the Arrow conditions into corresponding ones appropriate for anchored scales. The primary change is excluding the anchor sets from the conditions. However, C2 is expressed in such a way that it applies to the anchor sets in a derivative fashion. It is clear that some conditions would want to be imposed on the group function G as it applies to the anchor sets.

Condition 1. There is a set of three objects B, such that B does not overlap with any A_i, and for every T, where T is a possible ordering of three objects by n individuals, there is an S in K such that R^{*B} is isomorphic to T.

Recall that R_i^* is the imposed relationship defined by R_i "restricted" to $S_i(x)$.

Condition 1 states that there are at least three objects which can take on any possible ordering (by n individuals) of the scale values of the objects. The condition is expressed in terms of the R*'s being isomorphic to T rather than identical with it on B, to save a certain amount of circumlocution concerning the relation of T to A. As noted above, Condition 1 requires that every A_i have at least two members (at least three intervals for each individual scale.)

Def. 3: $FS(x,S,\underline{S})$ means $FS(x,R,\underline{R})$, where R and $\underline{R}$ are the ordering relations corresponding to S and $\underline{S}$ respectively. This definition simply transfers the meaning of FS from relations to scales.

Condition 2: If x is not in any A_i and $FS(x,S,\underline{S})$, then if $x\vec{G}y$, $x\vec{\underline{G}}y$.

This is a fairly straightforward translation of Arrow's condition 2 to anchored scales. It is, of course, weaker than the Arrow condition, since we exclude the anchor sets.

Condition 3: If $S^B = \underline{S}^B$, then $F(S)^B = F(\underline{S})^B$. Here we must include in the operation S^B that, in restricting a scale to the set B, we also retain A.

Condition 4: For every x and y, x, y not in any A_i, there is an S such that $x\vec{G}y$, where G is the group relation associated with F(S).

Again, this is the direct analogue of Arrow's condition, with the exception of the reference sets. In a sense, F(S) is imposed for A. The

group selects a G, and this remains fixed, as far as A is concerned, for every acceptable S.*

Condition 5: For every i, there exists x not in A and S, y such that $S_i(x)\ R_i\ S_i(y)$, and $y\ \vec{G}\ x$.

It clearly would be unfair to ask only $x\ R_i\ y$, since this might mean $S_i(x) = S_i(y)$, which would be a poor negation of our dictator.

Given the above transcription of the Arrow conditions we can state the possibility theorem:

THEOREM 3. There is a function F(S) which satisfies conditions 1-5.

Actually, there are an infinite number of such functions. However, for the assertion of the compatibility of conditions 1-5, it is necessary to exhibit only one. A simple function which fulfills the conditions is one that could be called modified sum of ranks. To each of the individual *reference objects*, a rank-order number is assigned say in the ascending order of preference. Call this rank order number $S_i^o(a)$ and derivatively, $S_i^o(x)$ is the rank-order number of $S_i(x)$.

We then define

$$S^o(x) = \sum_i S_i^o(x)$$

$x\ G\ y$ if and only if $S^o(x) \geq S^o(y)$

It is clear that G is a weak order over X, since every x has a number assigned to it by S^o and $\geq$ is a weak order. Thus the definition assures that this F is indeed an anchored scale.

Taking up the conditions in turn, Condition 1 can be satisfied, as previously mentioned, if each A_i has at least two members.

---

*In a wider sense, this is not necessary, as the proof of the possibility theorem below shows. Thus, F(S) could be required to "work" for every A, as well as for every set of x not including A, providing condition 3 is restricted to the situation after an A has been selected. The specific preference scale used in the possibility theorem--modified sum of ranks--does in fact work for every finite A. But to be quite frank, I haven't found a sufficiently general notation within which this more powerful kind of condition can be expressed.

Condition 2: If we have $FS(x,S,\underline{S})$, then $\underline{S}^{o}(x) \geq S^{o}(x)$ and $\underline{S}^{o}(y) = S^{o}(y)$ for $y \neq x$. If $S^{o}(x) > S^{o}(y)$, $\underline{S}^{o}(x) \geq S^{o}(y) = \underline{S}^{o}(y)$. From which the conclusion $\underline{S}^{o}(x) > \underline{S}^{o}(y)$ follows.

Condition 3 is immediate. Restricting the individual scales to the set B (plus A) does not change $S^{o}(x)$ where x is in B.

Condition 4 is easily satisfied. We simply have to posit that there is an S such that $S^{o}(x) > S^{o}(y)$, which obtains if, e.g., $S_{i}^{o}(x) > S_{i}^{o}(y)$ for every i (unanimity.)

Condition 5 is equally easily satisfied. We only need find a case where $S_{j}(x) \vec{R}_{j} S_{j}(y)$ for every $j \neq i$, and $S_{i}(x) = S_{i}(y)-1$ (i.e., everybody except i prefers x to y, and i ranks x one level lower than y).

This completes the possibility theorem for group anchored scales. It should be clear that the specific group scale used above, namely $S^{o}(x) = \sum_{i} S_{i}^{o}(x)$, is only one of an infinite set which would fulfill the conditions.

The modified sum of ranks group scale was selected mainly because it made the demonstration easy. There was one other secondary reason. Ordinary sum of ranks does not fulfill Arrow's Condition 3. The modified sum of ranks does not get into trouble because the anchor set is retained in going to a subset. Thus, the modified sum of ranks is a good elementary example of how it is feasible to have an ordinal group scale which does not run into the difficulties attendant on aggregation of "pure" relations. It might be objected that group anchored scales defined by some function of ordinal numbers on the individual anchors are not "purely ordinal." In general such functions will not be invariant over separate monotonic transformations on the individual scales (e.g., if the order numbers of one of the individual scales was multiplied by a factor of 1000, the others remaining the same, the group scale would be dictatorial.) That does not mean,

however, that there are hidden cardinal assumptions in the notion of group anchored scale. A group scale can be fully specified with no reference to numbers whatsoever.

Although the modified sum of ranks group scale was selected primarily for its simplicity in showing that the conditions can be fulfilled, it should be of use in any situation where the ordinary sum of ranks appears relevant. A slightly casual suggestion along these lines is contained in the following section.

Someone might want to object that the modified sum of ranks, and similar group scales, falls down if the anchor sets are augmented or reduced. In some sense that is correct, but for the present purposes, augmenting or reducing the anchor sets is equivalent to defining a new group function F, and it is not expected that Condition 3, e.g., will hold across different F's.

Since there are many different group scales which will fulfill conditions 1-5, we suddenly have an embarrassment of riches. I would guess that one of the reasons Arrow dealt with such a spare and stringent set of basic notions was the hope that a few conditions would essentially determine the form of a rational group function, or perhaps limit the possibilities to a well-defined subclass. That, of course, would have been a highly significant result. Unfortunately, that hope is not fulfilled by group anchored scales. This, of course, leads to the important practical question -- how, in any given decision situation, one might go about selecting a particular group scale.

At the present time, there does not appear to be a way to determine the "right" group scale, based on some additional requirements. This is true only so long as we stay in the context of ordinal scales. If the notion of numerical utilities is introduced, then as demonstrated in Section 6

below, a few additional assumptions sharply restrict the form of a group scale. But within the ordinal context, about all one can say is that a group can adopt a particular group scale in about the same spirit in which it might adopt a constitution, or Robert's Rules of Order, or any other formal mechanism for systematizing decision-making.

More generally, it is my impression that we have very little knowledge concerning anchored scales in social measurement. This is particularly true for those cases where the objects being scaled are very complicated entities such as "states of society," or political or social systems. For this type of social scaling, it seems intuitively clear that criteria (even for an individual) are highly multi-dimensional, and the problem of defining reference objects must be tackled on the level of more elementary sub-scales. This consideration is one of the driving forces behind the social indicators movement. Society may be an object which is just too complicated to view "in toto."

5. Example: Electing a President

A relatively straightforward example of a possible application of the notion of anchored scales can be seen in voting schemes for public officials. One of the most serious consequences of the type of inconsistency formalized in the Arrow Theorem is the fact that elections need not result in the selection of the candidate most favored by the electorate. This fact is obscured in presidential elections in the United States by the two-party system and the large proportion of cases in which there are only two major candidates. It is a triviality that majority vote produces a "consistent" group preference relation when there are only two alternatives. However, for the case of more than two alternatives, several undesirable types of outcome are possible. In the French style election where there is a run-off between the two candidates with the

highest initial votes, there is a good chance that the candidate with the most overall support in the electorate will be eliminated on the first round. For example, consider the case where there are three candidates, A, B, C, and three voting blocs, X, Y, and Z, with the preferences (where 1 means most preferred, 3 means least)

	A	B	C	% of vote
X	1	2	3	25
Y	2	1	3	37.5
Z	2	3	1	37.5

Since B and C are the two winners on the first round, A is eliminated, and the winner of the run-off is B. However, A would win a two-candidate majority vote against either B or C, receiving 62.5% of the vote in either case.

It is easy to construct cases in which the "worst" candidate wins. Consider the following case

	A	B	C	% of vote
X	1	2	3	26
Y	3	1	2	28
Z	2	3	1	46

As before, A is eliminated on the first round, pitting B against C. Since B is preferred by 54% of the electorate, he wins. However, if we use a simple scaling procedure, namely the average rank of each candidate in the voters' preference relations, the average ranks are 2.02, 2.18, 1.80. B, with the lowest average rank, takes office! It seems highly likely that in the course of French elections situations at least as anomalous as those illustrated have occurred.

Even in the U.S., we occasionally have dubious cases. Take the Roosevelt, Taft, Wilson election of 1912.

	Roosevelt	Taft	Wilson	% of vote
R	1	2	3	.30
T	2	1	3	.25
W	2	3	1	.45

The tally of votes in the three-way election does not tell the whole story. A plausible assumption is that those who voted for Taft or Roosevelt would have preferred either to Wilson, giving the preference pattern in the table. With this preference pattern, Roosevelt would have won a two candidate majority vote against either Taft or Wilson, and Taft would have won in a straight majority contest with Wilson. Thus, by majority vote, the preference order is Roosevelt, Taft, Wilson. The average preference rank for Roosevelt is definitely better than the other two. Wilson barely nudges out Taft.

These anomalies can be straightened out by using a form of anchored voting. In the case of presidential elections, there is a natural set of anchors, namely the list of past presidents. In the crudest application, each individual voter at his leisure would rank order the past presidents. On election day, the voter reports the scale position of each of the present candidates in his scale of past presidents. Tallying would consist of adding up the scale values of each candidate, and the one with the highest scale sum wins. There is no requirement that the individual voters rank the past presidents in the same order—each can have his own personal ordering.

Anchored voting would not only guard against electing a non-preferred candidate, it would give a relatively unambiguous rating to all of the current candidates. There is one drawback to the elementary election procedure just described. There might be a tendency for voters to give an artificially high rating to their favorite candidate and an artificially low rating to all the

others. This would, of course, negate the procedure. There is a form of proper scoring procedure that is applicable to anchored ratings which would keep the voters honest. It is similar to the bidding rule described in Chapter III. In the case of voting, a large set of initial candidates is chosen, say 100 as a round figure. The electorate ranks all of these against their presidential scale. The final slate is a much smaller number, say 10, selected out of the initial list at random. The candidate with the highest rating in this sub-list is declared the winner. In this case, if the voter inflates his rating for his favorite candidate, he runs the risk of having that favorite candidate ruled out by the random selection, in which case he would have penalized his lower ranking choices.

As it stands, the form of election procedure just described is somewhat more cumbersome than present procedures. It is not without compensating virtues. As a television show, the drawing for the final slate of candidates could be a highly dramatic event. A certain amount of streamlining would, of course, be feasible. The set of anchors need not be all past presidents, but some smaller subset. The number of initial candidates, and the size of the final slate could be pared by judicious statistical design.

6. Cardinal Group Utility

Having arrived at the point where a group preference function is at least logically feasible, there doesn't appear to be a strong reason why the simplifications possible through numerical utilities shouldn't be taken advantage of. There seems to have been fairly general acceptance of the possibility of ascertaining utility functions for individuals using probability scaling as outlined in Section 1. Actually, the acceptance has been perhaps more enthusiastic than experimental attempts to determine such scales warrants. The question just how consistent individuals are in

rating contingencies remains unclear. Most of the practical applications have been more in the spirit of using utility theory as a prescriptive set of rules (a sort of axiological logic) rather than as a straightforward measurement technique.

For the time being we shall forego the question whether a prescriptive or descriptive role for individual utilities is involved, and examine the consequences of a simple additional postulate concerning the nature of group preference functions.

The group functions we want to examine, then, are of the general form $F(U,A,X)$ where as before X is the set of objects to be evaluated by the group, $U = (U_1,U_2,\ldots,U_n)$ is a vector of utility scales, and $A = (A_1,A_2,\ldots,A_n)$ is a vector of sets of reference objects. In this case, each A_i is simply a pair of objects, (a_i,b_i), where, to avoid triviality, we will assume $U_i(a_i) \neq U_i(b_i)$.

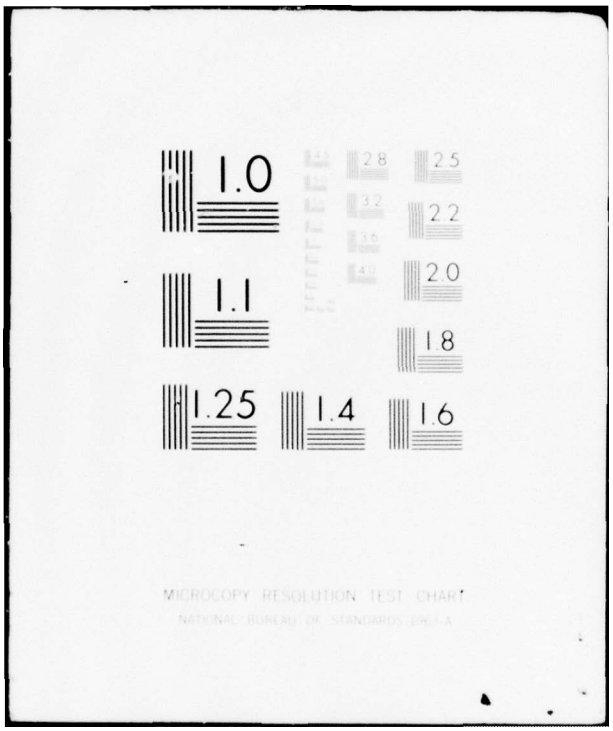
There are two levels at which a function F can be sought: (1) F maps the vectors U onto an ordinal scale--that is onto numbers fixed only up to a monotonic transformation, (2) F maps the vectors U onto a scale which is itself a utility scale (fixed up to a liner transformation.) In the first case, F is simply a device to generate a weak ordering on the utility vectors. In the second case it is a device to generate a cardinal utility scale for the group. The second might appear to be a much stronger requirement than the first. As it turns out, the step from (1) to (2) is rather small.

The utility vectors $U = (U_1,\ldots,U_n)$ form a space which, for simplicity, will be assumed to be Euclidean n-space. The set X of objects (outcomes, consequences) is mapped onto the utility space by the individual utility functions

AD-A042 852 CALIFORNIA UNIV LOS ANGELES SCHOOL OF ENGINEERING A--ETC F/G 5/10
GROUP DECISION THEORY.(U)
JUL 77 N C DALKEY
UNCLASSIFIED UCLA-ENG-7749 N00014-69-A-0200-4056
NL

4 OF 4
AD
A042852

END
DATE
FILMED
9 -77
DDC


1.0
1.1
1.25
1.4
1.6
2.8
2.5
3.2
2.2
3.6
2.0
4.0
1.8
MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS 1963-A

Note that if two objects x and y are considered equivalent by all members of the group, then x and y are mapped onto the same point in U.

Given the feasibility of group preference functions, a reasonable place to start would be to assume that there is a complete order of group preference on U. However, an appropriate theory has already been developed which is more informative, namely the theory generated in Chapter IV to deal with incomplete information. If the vectors $x = (x_1,\ldots,x_n)$ of that theory are reinterpreted as vectors of individual utilities, and the choice sets C(S) are reinterpreted as resulting from group choice, then postulates H1-H5 appear as reasonable for group as for individual decisions.

A specific group decision problem is represented by a set S of attainable utility vectors in U. The cooperative nature of the decision is expressed by including the outcomes of all potential coordinated actions of the members of the group. In addition, it is assumed that all probability combinations of the actions (equivalent to all probability combinations of the outcomes) are included in S. In this way the convexity of the set S is assured.

A key assumption is that the group choice C(S) is solely a function of the set S; i.e., it is a function solely of the individual utility vectors. This assumption has been questioned when contingencies are included in the outcomes.[6] Consider the utility space for a group of two, illustrated in Fig. 46. Individual utility theory implies $U_i(z) = U_i((x,y|\frac{1}{2})) = U_i((u,v|\frac{1}{2}))$ for both i = 1 and 2. Hence, they are all mapped on the same point and treated identically by the group choice function. It has been objected that z can be a sure equal allocation of utility between the two individuals, whereas $(u,v|\frac{1}{2})$ allocates all the utility to one individual and none to the other.

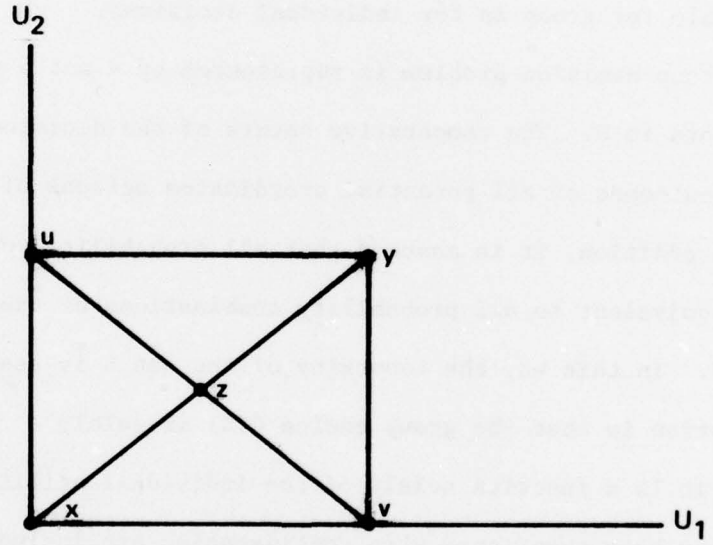


Figure 46. Illustration of Unanimity Condition for Contingencies

In a similar vein, $(x,y|\frac{1}{2})$ assures that in either case the two individuals are treated equally whereas $(u,v|\frac{1}{2})$ involves unequal allocations.

An appeal to unfairness toward individuals in this situation seems inappropriate. By definition, if a given individual finds two outcomes equivalent, they are equally acceptable.* The question whether there are group interests which override the interests of the individuals is more profound. One consideration relating to fair division is discussed below after introducing the equivalence condition. Another consideration is whether the group is more or less risk averse than its members. For example, in Fig. 46, if x is disaster for both, and y is utopia, each individual separately might find the gamble equivalent to the intermediate state z, whereas the group strongly prefers z to the gamble. This is not an easy issue to handle with generalities. The not fully conclusive data on the "risky shift" phenomenon appears to indicate that, if anything, groups are less conservative than individuals.[7] The fact that highly risky group behavior such as wars and revolutions are rather common suggests that groups are not risk-averse, at least under some circumstances. Perhaps the strongest support for assuming that the group is neither more not less risk averse than its members comes from the unanimity principle applied to contingencies. If all the members find a given contingency as desirable as a given "certainty equivalent" than by unanimity the group would make the same judgment.

The foregoing is not an overwhelming justification of the assumption that group choice is a function solely of the individual utilities. It

*Since the "objects" being evaluated by each individual are states of the group, e.g., allocations of rewards, feelings concerning the desirability of equal allocations can be absorbed in the individual utility scales. It is quite possible, contrary to the unfairness argument that many individuals would prefer a gamble in which they at least had a chance to "come out ahead" to a bland equal distribution.

appears solid enough to expect that the assumption is appropriate for a useful range of group decisions.

To recap conditions H1-H5: For any subset S of U which is closed, bounded from above, and convex, there exists a choice set C(S) which, on the present interpretation, represents the points in S which the group would select if S were the alternatives available in a decision. The function C(S) defines a partial preference relation on U, where $x \succeq' y$ means x is in C(S) and y is in S. The ancestral relation $\succeq^*$ of $\succeq'$ is transitive but not necessarily connected. $\succeq^*$ is acyclic, continuous, Archimedean, and fulfills the dominance condition. On these assumptions, there is a complete order of group preference $\circledcirc$ on U which is an extension of $\succeq^*$.

Although the role of the reference set A is not explicit in the above formalism, it cannot be overlooked. Since individual utility functions are fixed only up to a linear transformation, the reference set is required to insure the stability of the choice function. If a group choice function C(S) is defined for a particular assignment of numbers to the reference objects, then if one or more of the individual scales is changed, the choice function must be rescaled accordingly.

Some additional comment on the conditions H1-H5 as they apply to group decisions is probably in order. H1, as already noted, has the effect of assuring that the group choice will be a function of the individual utilities. H2, acyclicity, appears to be a fairly stalwart postulate. The evils that can arise with cyclic preferences have been extensively explored in the literature.* Acyclicity plays the same role in the present approach

*E.g., the rather persuasive notion of a "money pump."[8]

as independence of irrelevant alternatives plays in the Arrow postulate set, and guards against the same kind of instability.

Acyclicity has one rather stern consequence; it rules out a basic tenet of cooperative game theory, namely the postulate of individual rationality. The postulate of individual rationality holds that an individual will not accept an outcome that is less preferred (by him) than an outcome he can guarantee with his own efforts. For example, a player in a multi-person game can guarantee himself the max min of his payoff where the minimization is taken over all coordinated strategies of his opponents. At first glance, there is no reason why the player should accept anything less than this guaranteed amount. However, in the present formalism, a given set S of alternatives can be generated by many different decision situations with different individual rationality points. If C(S) is made conditional on the individual rationality points (which it is for most cooperative game solution concepts) then within the same set S we can have $x >' y$ and $y >' x$, a violation of acyclicity.

This point is not an objection to cooperative bargaining models per se; however, it appears to be a definite objection to bargaining models as a basis for group decisions with continuing groups, i.e., groups that can be expected to conduct a sequence of interrelated decisions.

Given H1, the dominance condition H3 appears to involve little that is controversial. It includes, of course, the unanimity condition as a special case.

The continuity axiom H4 introduces a rough kind of comparison between the utilities of different individuals, in the sense that a small utility difference for one individual will not be construed as "infinitely greater"

than any utility difference for another. In this regard, it rules out an absolute dictator. It does not rule out, of course, a "dictator in fact"; one individual could still have an overwhelming influence on most decisions.

The Archimedean axiom H5 in its present form is difficult to interpret directly in terms of group behavior. It involves potentially unlimited sequences of decisions, something difficult to set up in the laboratory or in real life. It can be replaced by a less global assumption concerning limited variation of C(S) in the vicinity of a given point (so-called Lipchitz conditions.[9]) However, the local assumption is not much more perspicuous than the global one. Roughly speaking, the assumption holds that the difference between x and y, if x is "barely preferred" to y, is negligibly greater than the difference if y is "barely preferred" to x.*

H1-H5, as we have seen, have the consequence that there is a complete ordering on U, which is an extension of $\succeq^*$. We need only one more assumption to specify the <u>form</u> of the group utility function completely. The additional assumption is

H6. If $x \mathbin{\textcircled{\sim}} y$, then $x \mathbin{\textcircled{\sim}} (x,y|E)$

Here $\mathbin{\textcircled{\sim}}$ means equivalent according to the group preference relation.

The verbal justification of H6 is fairly obvious – if the group is indifferent between two situations x and y, then there appears to be no reason why it should prefer either one to selection of one by a chance

---

*Continuity and Archimedean axioms are difficult to justify on direct grounds. They are usually easier to justify in terms of the consequences they generate for observable phenomena. In the present case, there is a certain awkwardness in this fact. As will be amplified in Section 8, H1-H5, and H6 which will be introduced shortly, do not characterize most group decision processes now in use. They have the status more of recommendations for improved group decision procedures. Under these circumstances there are no "observable phenomena" to explain. In the absence of figures of merit, about all that can be appealed to is the "face validity" of the consequences.

mechanism. The verbal justification appears highly persuasive when applied to individual decisions. However, there are interesting new issues that arise for groups. Returning to Fig. 46, suppose the group is indifferent between u and v (e.g., the two individuals are co-equals, and the group cannot distinguish between one or the other getting a given reward.) It still might be the case that tossing a coin to determine who would obtain the reward is preferrable (to the group) to awarding it by fiat to one individual.

Tossing a coin is a traditional way of settling the problem of fair division in the case of indivisible objects. The fairness concept involved is closely related to the "Buridan's Ass" issue for groups. If the group is indifferent among a set of situations, how is it to select one? Even though the alternatives are equivalent as far as the group is concerned, a method of selection which is biassed can lead to manifest unfairness. As an obvious case in point, suppose the rule were: in the case of equivalent allocations to individual 1 or to individual 2, always make the favorable allocation to individual 1. A random assignment is clearly more desirable than the biassed one. If the requirement for avoiding bias in selecting among equivalent alternatives is included on the ground floor, so to speak, then H6 is apparently untenable.

The same kind of problem arises, but in a milder form, with the Buridan's Ass puzzle for individuals. Suppose a given individual has several actions which have equal expected utility. How should he pick the action to pursue? It is not difficult to design biassed methods of selection which over the course of several decisions could result in unfortunate circumstances. An equal probability mixture of the equivalent actions would be preferrable to the biassed rule. However, if this bit of

prudence were included in the formal definition of preferred action, the consequence would be that the individual would exhibit a positive preference for gambling — the standard utility axioms would not hold. For individuals, then, it seems desirable not to include bias reduction procedures in the elementary framework of utility theory. Random selection among equivalent actions is a useful rule to add after the basic definition of equivalence.

The desirability of avoiding bias in selection among equivalent outcomes for groups is perhaps stronger than for individuals. However, the same resolution appears to be applicable in both cases. Given a definition of equivalent cases, then, to avoid bias, the additional rule can be imposed to select one out of the equivalent set at random. In this regard, the fairness rule would be non-Archimedean. The random selection is preferred to any member of the equivalent set, but is not preferred to any alternative which is preferred in a primary way to any alternative in the set.

With this somewhat extended advocacy of H6, we can turn to the implications, which are rather far reaching. H1-H6, plus the assumption that U is a utility space lead to the theorem.

Theorem 4. There is a group utility function G on U, $G(x) \geq G(y)$ if and only if $x \textcircled{\geq} y$, and G consists of a weighted sum of the individual utilities, i.e., $G(x) = \sum_i a_i U_i(x)$, where the a_i are constants.

The full proof of Theorem 4 is given in Appendix IV. The essence is straightforward. H6 implies that the equivalence sets on U are convex, hence they are bounded both above and below by hyperplanes. Since the bounding hyperplanes cannot intersect, they must be parallel. The defining equations of these hyperplanes, $\sum_i a_i u_i = c$ also specify the equivalence sets.

$\sum_i a_i U_i$ is a utility function which fulfills the conditions: (a) $x \textcircled{\geq} y$ if and only if $G(x) \geq G(y)$, (b) $G((x,y|E)) = P(E)G(x) + (1-P(E))G(y)$.*

The constants a_i have been interpreted in the literature as weights representing the relative importance which the group places on the individual utilities. However, until some more determinate structure is placed on the individual utility scales, the most that can be said is that the constants a_i act as adjustments on the individual utility scales, where the adjustments may or may not include assessments of relative importance. This topic will be explored further in Section 8.

7. Minimizing Regret

One illuminating way to examine the import of the weighted sum group utiltty is in terms of regret. The regret of individual i, if the group selects point g in some set S is

$$R_{ig} = \max_{u_i} [u \text{ in } S] - U_i(g) \qquad (1)$$

That is, the regret of individual i if the group selects g in S is just the maximum utility the individual could receive from any point in S minus the utility he obtains from g.**

The weighted sum of the individual regrets, $\sum_i a_i R_{ig}$ is one measure of the degree to which the group decision satisfies the individual members. The weights in this expression have the same interpretation as the weights in the weighted sum utility function. From (1)

---

*Theorem appears to be somewhat more general than a related result due to Harsanyi.[10] He has shown that if a group utility function exists on a utility space, which fulfills unanimity for equivalence, then the group utility function must be of the form $\sum_i a_i U_i$.

**The notion of regret appears to have been promulgated by Savage, although he occasionally seems to prefer to saddle someone else with the idea.[11]

$$\sum_i a_i R_{ig} = \sum_i a_i \max_{u_i} [u \text{ in } S] - \sum_i a_i U_i(g) \tag{2}$$

$$= M - \bar{U}(g) \tag{3}$$

where M abbreviates the first term in (2) and $\bar{U}(g) = \sum_i a_i U_i(g)$. Since M does not involve g, the minimum of the weighted average regret is obtained if the second term in (3) is maximized, i.e.,

$$\min_g \sum_i a_i R_{ig} = M - \max_g \bar{U}(g) \tag{4}$$

(4) establishes the theorem:

Theorem 5. Minimizing the weighted sum regret is equivalent to maximizing the weighted sum of the individual utilities.

Although the theorem is not very deep, it can be considered as a second route for arriving at the weighted sum of individual utilities as a group utility.

The approach to group decisions via minimizing regret resembles an approach called the "theory of the displaced ideal" by Zeleny.[12] For any S, define u_m by

$$u_{mi} = \max_{u_i} [u \text{ in } S]$$

that is, u_m consists of the best possible outcome in S for each individual separately. Generally u_m will not be in S. If it is, unanimity legislates that u_m will be in C(S). If u_m is not in S, then C(S) is defined to be the set of points in S which are nearest to u_m; i.e.,

$$C(S) = \{u \mid u \text{ in } S \text{ and } d(u,u_m) = \min_v [d(v,u_m), v \text{ in } S]\} \tag{5}$$

u_m is considered to be the "ideal" point which the group would like to attain, but usually cannot. If u_m is not attainable, i.e., not in S, then the group selects the point that is as close as possible to it. Various

decision rules are obtained by using various definitions of distance. If d(u,v) is taken to be

$$d(u,v) = \left| \sum_i a_i (u_i - v_i) \right| \tag{6}$$

then (5) defines the same outcomes as the min regret rule (4). d(u,v) defined as in (6) is not a true distance. It fulfills all the conditions (D1-D3, Chapter II) for a distance except d(u,v) = 0 implies u = v. If u and v are on the hyperplane $\sum_i a_i u_i = c$, d(u,v) = 0, even though $u \neq v$. The d(u,v) of (6) can be thought of as a generalized form of distance. It is a member of the family of "distances" which can be generated by complete orderings on U. Specifically, if $\geq$ is a complete ordering on U which can be represented by a real-valued function f, i.e., $u \geq v$ if and only if $f(u) \geq f(v)$, then the generalized distance defined by f is just $d_f(u,v) = |f(u) - f(v)|$.

The displaced ideal approach with true distances will generally lead to violations of acyclicity. I have not proved this in complete generality, however Fig. 47 shows a typical case. For the large set S, x is the point closest (ordinary Euclidean distance) to the ideal point u_m for S. If a subset S' is selected where $u'_{m2} = x_2$, then C(S') = y. We thus have $x >' y$ and $y >' x$.

On the other hand, if the displaced ideal approach is taken, using generalized distances as defined above, then the group decision will not be cyclic. I haven't proved this with complete generality either, but it appears plausible if the class of complete orders is restricted to those generated by H1-H5.

Theorem 6. The complete order $\textcircled{\geq}$ implied by conditions H1-H5 can be represented by a real-valued function f, such that $u \textcircled{\geq} v$ if and only if $f(u) \geq f(v)$.

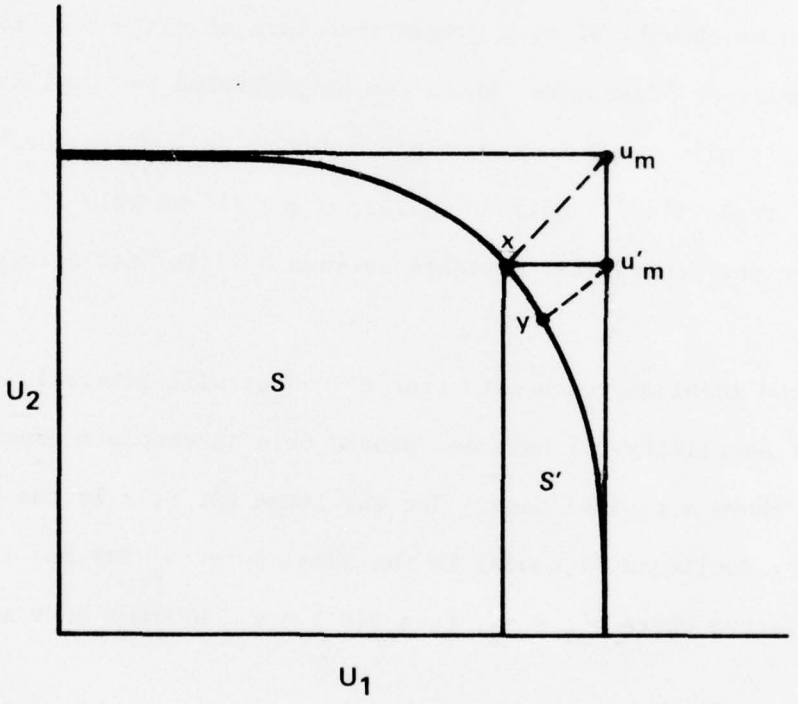


Figure 47. Illustration of Violation of Acyclicity by Displaced Ideal Decision

Proof: Select any positive ray R. Every equivalence set B intersects R at some point. These intercepts map the set of equivalence sets onto the continuum of points on R. Any real-valued function f which is monotonic on R (that is, given x and y on R, $f(x) \geq f(y)$ if and only if $x_i \geq y_i$ for all i) is an appropriate representation.

In the present context, H6 restricts d(u,v), to the form (6).

One happy feature of the min weighted regret criterion, then, is that it does not violate acyclicity (or analogously, does not violate independence of irrelevant alternatives). Other notions based on regret, such as min max regret do violate these conditions.

8. Note on Establishing Weights

The preceding discussion appears to give a fair amount of support to the assumption that the weighted sum of individual utilities is a reasonable form of group utility function for many group decisions. With that assumption, the only major issue arising in practice is the determination of the weights. In theory this is not a deep problem. Given fixed utility scales for the individuals — i.e., specific assignments of numbers to the individual reference objects — weights can be obtained empirically. The group makes an appropriately large number of choices among potential outcomes, each of which is representable by a vector $(u_1,\ldots,u_n)$ of individual utility ratings, and the optimal linear regression of the group choice against the individual utilities determines the (implicit) weights underlying the group choice.* In this respect, determining the individual weights is no more abstruse than

*If the group choice is expressed in a set of binary choices between pairs of outcomes, the computation is more intricate than elementary linear regression, but no new conceptual difficulties are introduced.

eliciting an individual's subjective probabilities, or determining an individual's utility function on a given set of objects.

The difficulty is that, in practice, most groups do not have a well-defined decision process; or if they do, no one would expect that the process would fit conditions H1-H6. As mentioned earlier, the most widely utilized group decision process is some variant on the dictatorial, which is unlikely to fulfill the continuity condition. Groups that rely heavily on non-anchored voting schemes are likely to violate the acyclicity condition.

If the weighted sum group utility is interpreted not as a good approximation to actual group practice, but as a recommendation for an improved group decision procedure, then the issue of determining weights requires additional assumptions beyond H1-H6.

There are two additional criteria which have been given some attention in the literature. These could be labelled the equity principle and the merit principle. Roughly, the equity principle states that (without some special reason to the contrary) the weights should be equal. The merit principle holds that individuals should be weighted in accord with some measure of their worth to the group, where worth may mean either contribution to the group utility, or some intrinsic value, or both.

Before either of these criteria can be stated precisely, there is a technical hurdle to overcome. The simplest statement of the hurdle can be made in reference to the equity principle; namely, what is meant by equal weights? Since the individual utility scales are determined only up to a linear transformation, attaching equal weights to any particular set of individual utility scales has no special significance.

If each individual's utility assignments over the set of outcomes in all decisions were just a linear transformation of any other individual's

assignments, then there would be no problem. The various individual scales could simply be rescaled to coincide, and equal weights would then have precise meaning. This is the kind of happy situation which makes the use of many different thermometric substances for the measurement of temperature feasible in physics.

Individual utilities are usually not simple linear transformations of each other over the outcomes in a decision situation. If they were, unanimity would be the only decision rule required for groups. In a typical decision we can expect both differential rewards across different outcomes and possibly different evaluations of comparable rewards — a mixture of disagreement on interests and values. These two can be disentangled by stepping outside the decision situation. As a gedanke experiment, suppose it were possible to formulate a description of all possible conditions which any individual in the group might experience in any potential outcome of any potential decision situation. In order to make the descriptions of the conditions complete, it would be necessary to include not only the explicit disposition of rewards which define the outcomes, but also any contextual factors which might influence the relative evaluations of the rewards. To use a cliche example, if one member of the group is a pauper, and another a prince, and in some decision $1,000 is awarded to some individual, then the list would include being a pauper and receiving $1,000, and being a prince and receiving $1,000.

We now ask each individual to formulate his utility function on this set of hypothetical conditions. Leaving aside the question whether anyone could actually finish this assignment (the axioms of individual utility theory say anyone can do it!) we ask how the various utility functions

compare. If by some stroke of luck, the utility scales are linear transformations of each other, then for that group, the assumption that they are identical (when rescaled to coincide), seems to be a reasonable basis for defining equal weights. On the rather scanty experimental evidence now available, there is no guarantee that the utility scales of different individuals will be linearly related. They could, for example, look like the two curves in Fig. 48.

The assumption of identical utilities if the scales are mutual linear transforms is a convention that seems innocuous since the group still has the option of establishing unequal weights for the group utility function. But there does not appear to be a "natural" convention for the case of non-linearly related utilities. If a pair of conditions out of the master list is selected as the group reference points, and the individual utility scales are normalized to coincide on those points, then different rescalings will result from different pairs of reference points. At the present time there does not appear to be any theoretical basis for distinguishing potential reference points.*

One straightforward convention that at least has the advantage of not requiring selecting an arbitrary pair of reference points is a minimal discrepancy rescaling. That is, constants r_i, s_i are computed for each individual i so that

$$\sum_i \sum_j d(U'_{ij} , \bar{U}'_j) \tag{7}$$

*If the doctrine of bounded utility were generally accepted, and there were a method of identifying the least upper bounds of the individual utilities, then at least one reference point could be obtained by the convention that the least upper bound of one individual represents the same amount of utility as the least upper bound of any other individual.

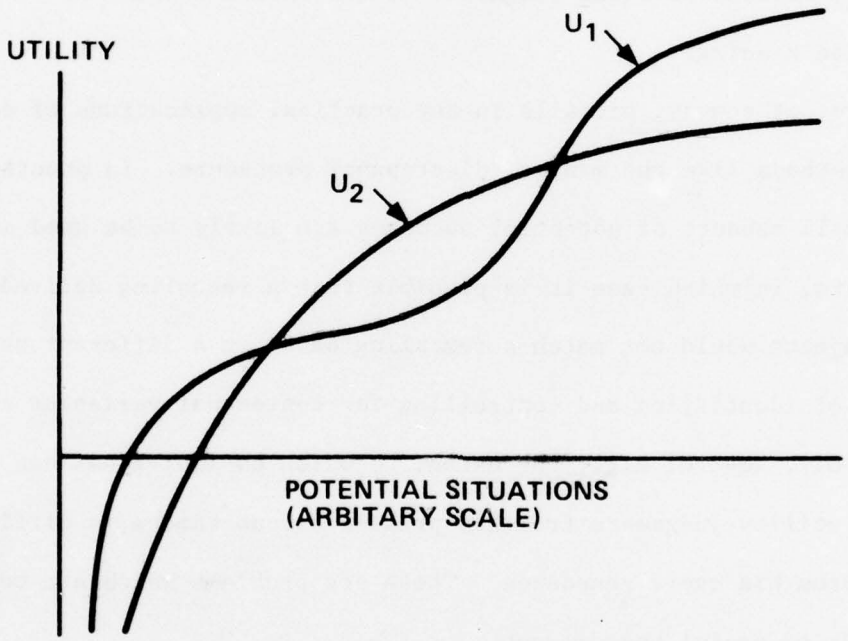


Figure 48. Non Linearly Related Utilities

is minimized, where $U'_{ij} = r_i + s_i U_{ij}$. U_{ij} is individual i's utility assessment of object j, and $\bar{U}'_j$ is an appropriate average of the U's for object j. Various distance functions might be appropriate, such as difference squared, absolute difference, absolute difference of logs, etc., depending on the shape of the utility functions.

The minimal discrepancy rescaling is, so to speak, a best approximation to equating the utility scales. Starting with the rescaled individual utility functions, the notions of equal weights or differential weights are at least given a precise meaning.

There are, of course, pitfalls in any practical applications of empirical scaling methods like the minimal discrepancy procedure. In practice, relatively small subsets of potential outcomes are likely to be used as comparison sets, in which case it is possible that a rescaling derived from one set of objects would not match a rescaling based on a different set. The problems of identifying and controlling for contextual variables are, to put it mildly, severe; e.g., the extent to which an individual has divorced his utility judgments from his present circumstances is difficult to evaluate from his overt responses. These are problems which are common in many fields of social measurement.

NOTES AND REFERENCES

Chapter I

1. Some applications of cluster analysis to point-of-view differences are discussed in Chap. III of Studies in the Quality of Life, N. Dalkey, et al., D. C. Heath, Lexington, Mass., 1972, and in N. Dalkey, "A Delphi Study of Factors Affecting the Quality of Life" in The Delphi Method, H. Linstone and M. Turoff, Eds., Addison Wesley, Reading, Mass., 1975.

2. Dalkey, N., "An Experimental Study of Group Opinion," Futures, Vol. 1, No. 5, Sept., 1969, pp. 408-426.

3. von Neumann, J., and O. Morgenstern, Theory of Games and Economic Behavior, Princeton University Press, Princeton, N.J., 2nd Ed., 1947.

4. Gardiner, P. C., and W. Edwards, "Public Values: Multi-Attribute Utility Measurement for Social Decision-Making," Social Science Research Institute Report, Univ. of Southern Calif., 1976. Also Dalkey et al., Studies in the Quality of Life, op. cit., Chap. IV.

5. Dalkey, N., and B. Brown, "Comparison of Group Judgment Techniques with Short-range Predictions and Almanac Questions." The Rand Corporation, R-678-ARPA, May, 1971.

Chapter II

1. Tversky, A., and D. Kahneman, "Judgment under Uncertainty," Science Vol. 185, No. 4157, 27 Sept., 1974, pp. 1124-1131.

2. An entertaining discussion of this paradox is given in Salmon, W. C., "Confirmation," Scientific American, Vol. 228, No. 5, May, 1973, pp. 75-83.

3. E.g., the comments on universe of discourse in A. Tarski, Introduction to Logic, Oxford Univ. Press, New York, 2nd Ed., 1946, Sec. 23.

4. Savage, L. J., The Foundations of Statistics, 2nd Ed., Dover Publications, New York, 1972, Chap. V, Sec. 5, Small Worlds.

5. There is an extensive literature on the subject. Discussion relevant to the present chapter can be found in Formal Representation of Human Judgment, B. Kleinmuntz, Ed., John Wiley and Sons, New York, 1968.

6. Bavelas, A., "A Mathematical Model for Group Structure," Appl. Anthrop. Vol. 7, 1948, pp. 16-30.

7. Stevens, S. S., "Ratio Scales of Opinion," in D. K. Whitla, Ed., Handbook of Measurement and Assessment in Behavioral Science, Addison-

8. Anderson, N. H., "Information Integration Theory: A Brief Survey," in R. C. Atkinson et. al., (Eds.), Contemporary Developments in Mathematical Psychology, Vol. II, Freeman, San Francisco, 1974.

9. Brunswick, E., The Conceptual Framework of Psychology, Univ. of Chicago Press, Chicago, Ill., 1952.

10. Keeney, R. L., and H. Raiffa, Decisions with Multiple Objectives: Preferences and Value Trade-offs, John Wiley & Sons, New York, 1976.

11. E.g., Hammand, K. R., C. J. Hursch, and F. J. Todd, "Analyzing the Components of Clinical Inference," Psychol. Rev., Vol. 71, 1964, pp. 438-456.

12. Goldberg, L. R., "Man Versus Model of Man," Psychol. Bull., Vol. 73, No. 6, 1976, pp. 422-432.

13. Ramsey, F. P., "Truth and Probability," in The Foundations of Mathematics and Other Logical Essays, Kegan Paul, London, 1931; L. J. Savage, Foundations, op. cit., B. de Finetti, Theory of Probability, Vol. I, 1974, Vol. II, 1975, John Wiley and Sons, New York.

14. von Mises, R., Probability, Statistics and Truth, McMillan, New York, 1957.

15. Capen, E. C., Atlantic Richfield Co., Los Angeles, personal communication. Related material in E. C. Capen "The Difficulty of Assessing Uncertainty," paper presented at 50th Annual Fall Meeting of the Society of Petroleum Engineers of AIME, Dallas, Texas, Sept., 1975.

16. Shuford, E. H., A. Albert, and H. E. Massengill, "Admissible Probability Measurement Procedures," Psychometrica, Vol. 31, No. 2, June, 1966, pp. 125-145.

17. Lichtenstein, S. C., B. Fischhoff, and L. Phillips, "Calibration of Probabilities: The State of the Art." Proceedings of the Fifth Conference on Subjective Probability, Utility, and Decision-Making, Darmstadt, Germany, in press.

18. Dalkey, N., "Toward a Theory of Group Estimation," in The Delphi Method, H. Linstone and M. Turoff, Eds., Addison and Wesley, New York, 1975.

19. Dalkey, N., "An Experimental Study..." 1969, op. cit.

20. Raimi, R. A., "The Peculiar Distribution of First Digits," Scientific American, Vol. 221, No. 6, Dec., 1969, pp. 109-120.

21. Anderson, N. H., 1974, op. cit.

22. Slovic, P., B. Fischhoff, and S. Lichtenstein, "The Certainty Illusion," Oregon Research Institute Bulletin, 16, 4, 1976.

Chapter III

1. Birnbaum, A., "Some Latent Trait Models and Their Use in Inferring an Examinees Ability," in F. M. Lord, and M. R. Novick, Statistical Theories of Mental Test Scores, Addison-Wesley, Reading, Mass., 1968.

2. Kemeny, J., and J. L. Snell, Mathematical Models in the Social Sciences, Blaisdell, New York, 1963.

3. Dalkey, N., "An Experiment Study...," 1969, op. cit.

4. The issue whether correlations are useful criteria appears to depend on the aim of the researcher. Correlations are clearly weak in distinguishing between several multi-dimensional models, where each of the independent variables is positively correlated with the dependent variable. This topic is explored in Chap. IV, Sec. 7. However, if the aim is not to select the "best" model, but simply to measure the accuracy of a set of judgments, correlations are simply a transform of the common average squared error criterion. If we assume the variables have been normalized in formula (5) so that $Var(R) = Var(T) = 1 = s_R = s_T$, and $\bar{R} = \bar{T} = 0$ then average squared error = 2 + correlation of R and T.

5. The theory of probability scoring has generated a fairly extensive literature. Some of the more pertinent: Savage, L. J., "Elicitation of Personal Probabilities and Expectations," Jour. Am. Stat. Assoc., Vol. 66, No. 336, Dec., 1971, pp. 783-801. Shuford, et al., 1966, op. cit. Brown, T., "Probabilistic Forecasts and Reproducing Scoring Systems." The Rand Corporation, RM-6299-ARPA, July, 1970. Brier, G. W., "Verification of Forecasts Expressed in Terms of Probability," Monthly Weather Review, Vol. 78, 1950, pp. 1-3. Winkler, R. L., "Scoring Rules and the Evaluation of Probability Assessors," Jour. Am. Stat. Assoc., Vol. 64, Sept., 1969, pp. 1073-1078.

6. Reichenbach, H., The Theory of Probability, Univ. of Calif. Press, Berkeley, 1949, Chap. 10.

7. Raiffa, H., "Assessments of Probabilities," unpublished working paper, Harvard Univ., 1969.

8. Marschak, J., and R. Radnor, Economic Theory of Teams, Yale Univ. Press, New Haven, 1972, Chap. 2, Sec. 6. The Marschak-Radnor result is more complete than Theorem 7; it includes the necessity as well as the sufficiency of the refinement condition. The more general result is similar to the author's generalization of perfect information for general games: Dalkey, N., "Equivalence of Information Patterns and Essentially Determinate Games," in Contributions to the Theory of Games, Vol. II, H. W. Kuhn and A. W. Tucker, Eds., Princeton Univ. Press, Princeton, N.J., 1953.

9. Marschak, J., "Information, Decision, and the Scientist," in Pragmatic Aspects of Human Communication, C. Cherry (Ed.) D. Reidel Publishing Co., Dordrecht, Holland, 1974.

10. Brier, G. W., 1950, op. cit.

11. Brown, T., 1970, op. cit.

12. Marchak, J., "Actual versus Consistent Decision Behavior," Behavioral Science, Vol. 9, April, 1964, pp. 103-111

13. Dalkey, N., and B. Brown, 1969, op. cit., Tables 5 and 6.

Chapter IV

1. Keynes, J. M., A Treatise on Probability, Macmillan and Co., New York, 1921.

2. Carnap, R., Logical Foundations of Probability, Univ. of Chicago Press, Chicago, 1950.

3. Slovic, P. B., et al., 1976, op. cit.

4. Knight, F., Risk, Uncertainty and Profit, Houghton Mifflin, Boston, 1921.

5. Reichenbach, H., 1949, op. cit., Sec. 68.

6. Guilford, J. P., Psychometric Methods, McGraw-Hill, New York, 1936, pp. 426ff.

7. Tribus, M., Rational Descriptions, Decisions and Designs, Pergamon Press, New York, 1969, Chap. V.

8. Fisher, R. A., "On the Mathematical Foundations of Theoretical Statistics," Philos. Trans. Royal Soc., London, Series A, Vol. 222, 1922.

9. Savage, L. J., Foundations, 1972, op. cit., Chap. 3, Sec. 6.

10. Goldman, S., Information Theory, Dover, New York, 1953, Chap. IV.

11. von Neumann, and Morgenstern, 1947, op. cit. Theorem (17:6).

12. Ellsberg, D., "Risk, Ambiguity and the Savage Axioms," Quarterly Journal of Economics, 1961, pp. 670-689.

13. Allais, M., "Le Comportement de L'Homme Rationnel Devant le Risque" Econometrica, Vol. 21, 1953, pp. 503-546.

14. Friedman, M., and L. J. Savage, "The Utility Analysis of Choices Involving Risk," Journal of Political Economy, Vol. 56, 1948, pp. 279-304.

15. MacCrimmon, K. R., and S. Larsson, "Utility Theory: Axioms Versus "Paradoxes," Univ. of British Columbia, Working Paper No. 131, May, 1975.

16. Summarized in MacCrimmon, 1975, op. cit.

17. Einhorn, H., and R. M. Hogarth, "Unit Weighting Schemes for Decision-Making," Organizational Behavior and Human Performance, Vol. 13, 1975, pp. 171-192. The authors explore some of the implications of a somewhat more general formula than (19), including an average correlation between the variables, derived by E. Ghiselli in Theory of Psychological Measurements, McGraw-Hill, New York, 1964.

18. Dawes, R. M., and B. Corrigan, "Linear Models in Decision-Making," Psychological Bulletin, Vol. 81, No. 2, Feb., 1974, pp. 95-106.

19. Goursat, E., A Course in Mathematical Analysis, Vol. II, Part Two, Dover Pubs. Inc., New York, 1945, Chap. II, Sec. 30.

20. Shapley, L., and M. Shubik, "Game Theory in Economics — Chap. IV: Preferences and Utility," The Rank Corporation, R-904/4-NSF, Dec., 1974.

Chapter V

1. Maier, N.R.F., "Assets and Liabilities in Group Problem Solving: The Need for an Integrative Function," Psychological Review, Vol. 74, No. 4, July, 1967, pp. 239-249.

2. Feller, W., An Introduction to Probability Theory and its Applications, Vol. I, John Wiley & Sons, New York, 2nd Ed., 1957, Chap. IX, Sec. 2.

3. Aitchison, J., and J.A.C. Brown, The Lognormal Distribution, Cambridge Univ. Press, Cambridge Eng., 1957, Theorem 2.3.

4. Dalkey, N., "An Emperimental Study...," 1969, op. cit.

5. Dalkey, N., "An Impossibility Theorem for Group Probability Functions," The Rand Corporation, P-4862, June, 1972.

6. Dalkey, N., in Linstone and Turoff, 1975, op. cit.

Chapter VI

1. von Neumann, J., and O. Morgenstern, 1947, op. cit., Chap. I, Sec. 3.

2. Suppes, P., "Behavioristic Foundations of Utility," Econometrica, Vol. 29, 1961, pp. 186-202. Also Shapley and Shubik, 1974, op. cit.

3. Friedman and Savage, 1948, op. cit.

4. Arrow, K., Social Choice and Individual Values, Yale Univ. Press, New Haven, 2nd Ed., 1963, p. 24.

5. Discussed in Keeney, R. L., and C. W. Kirkwood, "Group Decision-Making Using Cardinal Social Welfare Functions," Massachusetts Institute of Technology, Operations Research Center Technical Report No. 83, Oct. 1973.

6. C.f., The discussion of the Arrow postulates in Luce, R. D., and H. Raiffa, Games and Decisions, John Wiley & Sons, New York, 1957, Chap. 14, Sec. 4.

7. Wallach, M. A., and N. Kogan, "The Roles of Information, Discussion, and Consensus in Group Risk Taking," J. Exp. Soc. Psychol., 1965, pp. 1-19.

8. Raiffa, H., Decision Analysis, Addison-Wesley, Reading, Mass., 1968, p. 78.

9. Coursat, E., 1945, op. cit.

10. Harsanyi, J. C., "Cardinal Welfare, Individualistic Ethics, and Interpersonal Comparisons of Utility," Journal of Political Economy, Vol. 63, 1955, pp. 309-321.

11. Savage, L. J., "Foundations," 1972, op. cit., Chap. 9, Sec. 8.

12. Zeleny, M., "The Theory of the Displaced Ideal," in Multiple Criteria Decision-Making, M. Zeleny, Ed., Springer-Verlag, Berlin, Heidelberg, New York, 1976.

APPENDIX I. Proof that if $f(1/n) + f(1/m) = f(1/nm)$ then $f(x) = k \log x$.

One additional assumption is needed, namely that $f(x)$ is continuous. Set $f(1/n) = f'(n)$. We then have $f'(n) + f'(m) = f'(nm)$ for any positive integers n and m. Now set $f'(x) = g(z)$, where $z = \log x$. We then have $g(z) + g(w) = g(z+w)$, from which we obtain $g(nz) = ng(z)$, or $g(z) = ng(z/n)$. Interating, we have $g\left(\frac{n}{m} z\right) = \frac{n}{m} g(z)$. If f is continuous, then g is continuous, and we have $g(xz) = xg(z)$ where x is any real number. Set $z = 1$, and $g(1) = h$. Then $g(x) = hx$, and $f'(x) = h \log x$. Since $f'(1/x) = f(x) = - h \log x$, $f(x) = k \log x$, which was to be proved.

APPENDIX II. Proof that H1-H5 imply the existence of a complete order on x.

The proof follows a slightly different route from the sketch presented in the text. The basic point is to show that the relation $\geq^*$ can be extended to a relation $\textcircled{\geq}$ which is transitive and connected. The notation is the same as that of Section 8, Chapter IV.

Let $x|y$ mean that neither $x \geq^* y$ nor $y \geq^* x$ holds; that is, x is "disconnected" from y with respect to $\geq^*$. Define

Definition 1. $x \textcircled{\geq} y$ means either $x \geq^* y$ or $x|y$.

Lemma 1: $\textcircled{\geq}$ is connected.

Proof: Immediate, since either $x \geq^* y$ or $y \geq^* x$ or neither.

Since $\textcircled{\geq}$ is connected, all that needs demonstration is that is is transitive. This is equivalent to showing that $\textcircled{\geq}$ is acyclic, as Theorem 1 shows.

Theorem 1: If a relation $\geq$ is connected and acyclic, it is transitive.

Proof: Assume $x \geq y \geq z$. Since $\geq$ is connected, either $x \geq z$ or $z \geq x$. If $x \geq z$, the theorem holds. If not $x \geq z$, then $z > x$. But $z > x$ is rejected by acyclicity.

Thus, all that needs to be shown is that $\textcircled{\geq}$ is acyclic, i.e., $x \textcircled{\geq}^* y$ implies not $y >' x$. By definition $x \textcircled{\geq}^* y$ implies there is a chain x_i, $x = x_1$, $y = x_n$, such that either $x_i \geq^* x_{i+1}$ or $x_i|x_{i+1}$. If all the links in this chain are of the form $x_i \geq^* x_{i+1}$, then H2 applies and not $y >' x$. The only open case, then, is that in which one or more links are of the form $x_i|x_{i+1}$. The term "strictly dominates" will be used in a narrow sense; x strictly dominates y means $x_i > y_i$ for all i.

Lemma 2: If $x >^* y$, then there is a z which strictly dominates y and $x >^* z$.

Proof: Let w be any point which strictly dominates both x and y. Then $w >^* x >^* y$. By H4 there is a b, $0<b<1$, such that $x >^* bw + (1-b)y$. Set $z = bw + (1-b)y$. z is on the positive ray determined by w and y. Since $z >^* y$, z strictly dominates y.

Lemma 3: If $x \geq^* y$ and z strictly dominates x, then there is a w which strictly dominates y, and $z >^* w$.

Proof: Since $z >^* x$, acyclicity rejects $y \geq^* z$. But not $y|z$, since $z \geq^* y$. Hence $z >^* y$, and Lemma 2 applies.

Definition 2: $P_x = \{y|y \geq^* x\}$. $Q_x = \{y|x \geq^* y\}$.

The definition is simply a reminder of the meaning of P_x and Q_x introduced in the text.

Lemma 4: Let x be any point and R any positive ray not containing x. R intersects P_x and Q_x and g.l.b. of P_x on R and l.u.b. of Q_x on R both exist.

Proof: R can be specified by the condition $y = w + ts$ where w is any point on R, s is a vector of positive numbers $0<s_i<1$, and t is any real number, positive or negative. Each value of t specifies a point y on R. Given any x, there is some t such that $w_i + ts_i > x_i$ for all i, and a t' such that $w + t's_i < x_i$ for all i. t defines a point u which dominates x, and hence is in P_x, and t' defines a point v which is dominated by x, and hence is in Q_x. The points on R are completely ordered by dominance, and u is an upper bound for Q_x, and v is a lower bound for P_x. Thus there is a g.l.b. for P_x on R and a l.u.b. for Q_x.

H5 specifies that for any positive ray R, g.l.b. P_x = l.u.b. Q_x. Hence the boundaries of P_x and Q_x coincide, and can be referred to by one notation,

B_x, the boundary of P_x and Q_x. The next lemma determines that the boundary is unique.

Lemma 5: If y is the intersection of B_x and some positive ray R, then for any other positive ray R' passing through y, y is the intersection of B_x and R'.

Proof: Assume some u on R', $u \neq y$, is the intersection of B_x and R'. u is either above y or below it on R'. Suppose u is above y. Then u dominates some u' on R' which in turn dominates some v on R that dominates y. Hence u cannot be the g.l.b. of P_x on R'. A similar argument involving Q_x rejects u below y.

Lemma 6: If $x|y$, then x is in B_y.

Proof: By hypothesis, x is not in P_y nor in Q_y. Thus, x is in B_y.

Lemma 7: $x|y$ implies that for any z which strictly dominates x, there is a w that strictly dominates y, and $z >^* w$.

Proof: Since z strictly dominates x, and $x|y$, not $y \geq^* z$. But by Lemma 6, not $y|z$, otherwise both x and z are in B_y and on a common positive ray. Hence $z >^* y$, and Lemma 2 applies.

Lemma 8: $x \textcircled{\geq}^* y$ implies for every z which strictly dominates x, there is a w which strictly dominates y, and $z \geq^* w$.

Proof: The lemma follows from Lemmas 3 and 7, and induction on the number of links in the chain from x to y.

Lemma 9: $\textcircled{\geq}$ is acyclic.

Proof: If $x \textcircled{\geq}^* y$, and $y >' x$, then from Lemma 2, there is a z which strictly dominates x and $y >^* z$. But from Lemma 8, there is a w which strictly dominates y, and $z \geq^* w$, thus $y >^* z \geq^* w >^* y$, which violates H2.

Lemma 9 completes the proof that $\textcircled{\geq}$ is a complete order on X.

APPENDIX III. Proof that complete independence ($D_R = 1$ and $D_R^{Ej} = 1$) implies that for two events, $P(E|R_i) = P(E)$ for all i but one.

The proof proceeds by induction on the number of respondents. For two events, E and $\bar{E}$, and two responses, R_1 and R_2, set $P(E) = p$, $P(\bar{E}) = 1-p$ $P(R_1|E) = q_1$, $P(R_1|\bar{E}) = r_1$, $P(R_2|E) = q_2$, and $P(R_2|\bar{E}) = r_2$. From the rule of elimination, $P(R_1) = pq_1 + (1-p)r_1$ and $P(R_2) = pq_2 + (1-p)r_2$. $D_R = 1$ implies $P(R_1.R_2) = P(R_1)\,P(R_2) = (pq_1+(1-p)r_1)\,(pq_2+(1-p)r_2)$. $D_R^{Ej} = 1$ implies $P(R_1.R_2|E) = q_1q_2$, and $P(R_1.R_2|\bar{E}) = r_1r_2$; whence, from the rule of elimination again, $P(R_1.R_2) = pq_1q_2 + (1-p)r_1r_2$. Equating these two expressions for $P(R_1.R_2)$ and expanding, the terms involving p cancel, leaving

$$q_1q_2 + r_1r_2 = q_1r_2 + q_2r_1$$

which can be factored

$$(q_1 - r_1)\,(q_2 - r_2) = 0$$

Thus, either $q_1 = r_1$ or $q_2 = r_2$. Assume $q_1 = r_1$. Then, from the theorem of Bayes, $P(E|R_1) = \dfrac{pq_1}{pq_1 + (1-p)r_1} = p$.

Assume the theorem holds for n-1 respondents. If complete independence holds for n respondents, then it holds for n-1, as shown by the following, where $q_i = P(R_i|E)$ and $r_i = P(R_i|\bar{E})$. Complete independence implies (where $(1-q_n) = P(\bar{R}_n|E)$, $(1-r_n) = P(\bar{R}_n|\bar{E})$)

$$p\prod_{i\neq n} q_i(1-q_n) + (1-p)\prod_{i\neq n} r_i(1-r_n) =$$

$$\prod_{i\neq n} (pq_i + (1-p)r_i)\,(p(1-q_n) + (1-p)\,(1-r_n))$$

$$p\prod_{i\neq n} q_i + (1-p)\prod_{i\neq n} r_i - p\prod_{i=1}^{n} q_i - (1-p)\prod_{i=1}^{n} r_i =$$

$$\prod_{i\neq n} (pq_i + (1-p)r_i)\ (1 - pq_n - (1-p)r_n) =$$

$$\prod_{i\neq n} (pq_i + (1-p)r_i) - \prod_{i=1}^{n} (pq_i + (1-p)r_i)$$

The terms involving $\prod_{i=1}^{n}$ on each side of the equality are equal by the assumption of complete independence, and hence the terms involving $\prod_{i\neq n}$ are equal.

Abbreviate $\prod_{i\neq n} (pq_i + (1-p)r_i)$ by Π, $\prod_{i\neq n} q_i$ by q and $\prod_{i\neq n} r_i$ by r. We have, from complete independence,

$$pqq_n + (1-p)rr_n = \Pi\ (pq_n + (1-p)r_n)$$

$$pq_n(q-\Pi) + (1-p)r_n\ (r-\Pi) = 0$$

and from complete independence for n - 1 respondents

$$pq + (1-p)r = \Pi$$

$$q = \frac{\Pi - (1-p)r}{p}$$

whence

$$pq_n\left(\frac{\Pi - (1-p)r}{p} - \Pi\right) + (1-p)r_n\ (r-\Pi) = 0$$

$$(1-p)\ q_n(\Pi-r) + (1-p)r_n\ (r-\Pi) = 0$$

$$r_n - q_n = 0\ ,\ r_n = q_n$$

Since, by the inductive hypothesis, at most one of the n-1 pairs $q_i, r_i, i = 1,\ldots,n-1$ are different, at most one of the n pairs are different.

APPENDIX IV. Proof of Theorem 4, Chapter VI.

Lemma 4 in Appendix II shows that the equivalence sets of the group preference function are not single points; in fact they are unbounded; that is, given any x and any u_i there is a y in B_x $y_i > u_i$. H6 implies that the equivalence sets are convex. Thus for any equivalence set B_x and for any point y not in B_x, there is a hyperplane H(x,y) containing y which bounds B_x — i.e., B_x is either entirely above or entirely below the hyperplane. If y is on a positive ray through x, and x dominates y then B_x is entirely above H(x,y). On the other hand, if z dominates x, then H(x,z) lies entirely above B_x. These two hyperplanes cannot intersect, since within the plane defined by R and any point of intersection w, (Fig. 49) there would be a positive ray R', where the intersection of B_x and R' would be above H(x,y), say the point u, and below H(x,z), say the point v, — contrary to the bounding hyperplane condition. Therefore H(x,y) and H(x,z) are parallel. Since z is any point above x on R and y is any point below, B_x must be the hyperplane parallel to H(x,y) through x. By a similar argument B_y for any y on R must be the hyperplane through y parallel to the hyperplane B_x. There is thus a set of constants $\{a_i\}$ such that $B_x = \{y | \sum a_i y_i = \sum a_i x_i\}$.

The constants a_i are all positive, since the slope of the intersection of any hyperplane with any given coordinate plane is negative. To show that $\sum_i a_i x_i$ is a utility function, let $z = (x,y|E)$. $\sum_i a_i U_i(z) = \sum_i a_i (P(E)U_i(x) + (1-P(E))U_i(y)) = P(E)\sum_i a_i U_i(x) + (1-P(E))\sum_i a_i U_i(y)$.

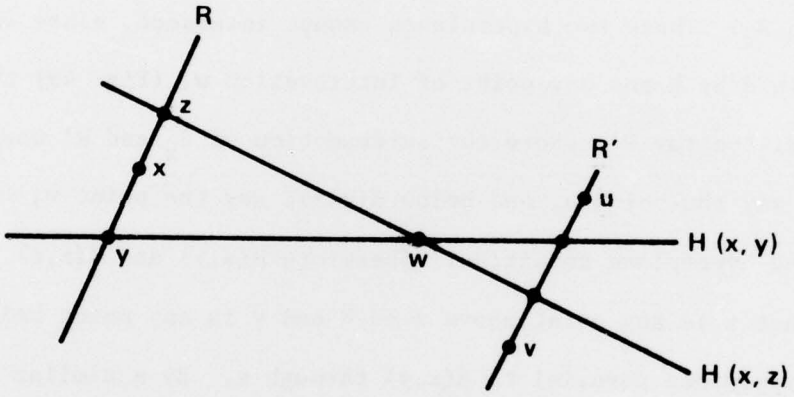


Figure 49. Intersecting Hyper-Planes Bounding B_x