Section I. PROBABILITIES AND THE NORMS OF BEHAVIOR

Recently, the U. N. forces in Korea faced the choice between two decisions: (I) to cross the 38th parallel or (II) not to cross it. Their commanders did not know whether, if the parallel were crossed, the Chinese adversary would or would not answer by counter-attacking in superior power, from across Korea's Northern border. In the language of current newspapers, our officers did not know whether the Chinese did or did not have "aggressive intentions." This is a vague term. The following table gives, however, an implicit definition that is precise enough. The intentions of the adversary are defined by their results, as follows:

<table>
<thead>
<tr>
<th>Chinese behavior</th>
<th>(I)</th>
<th>(II)</th>
</tr>
</thead>
<tbody>
<tr>
<td>behavior</td>
<td>Cross the</td>
<td>Don't cross the</td>
</tr>
<tr>
<td>(a) &quot;aggressive&quot;</td>
<td>parallel</td>
<td>parallel</td>
</tr>
<tr>
<td>(b) &quot;non-aggressive&quot;</td>
<td>U. N. Losses</td>
<td>U. N. Gains</td>
</tr>
<tr>
<td></td>
<td>Stalemate</td>
<td>Stalemate</td>
</tr>
</tbody>
</table>

We shall assume that, from the U. N. point of view, gains are preferable to stalemate, and stalemate is preferable to losses. We shall use the phrase "probability is equal to one" as equivalent to "it is certain." Now suppose the probability of the Chinese being "non-aggressive" is one. Then policy (I) produces a result (gains) preferable to the result (stalemate) of policy (II). The U. N. commander "should" choose (I). If the probability of (b) were not one but zero — i.e., if the probability of (a) were one — he

"should" choose (II). Thus if we know that our commander was sure (rightly or wrongly) of his opponent's intentions and has done what he "should" have done, then his opinion about those intentions – his assignment of probabilities to (a) and (b) – would be revealed by his action.

A similar inference from action to opinion, under the assumption that the acting person is doing what he "should" be doing, can be made also in a more general case: namely, in the case when our commander is not necessarily sure as to the adversary's intentions, and thus associates them with probabilities that are not necessarily 1 or 0 but may be intermediate between 1 and 0. Such fractional probabilities we have not yet defined. Yet you have no doubt acted on the basis of judgments such as "an assertion is almost certain to be true" (= "is very probable"), "almost certain to be untrue" (= has small probability), "as likely as not to be true" (= has probability 1/2).

Let us look into this more carefully. To avoid terminological confusion, let us call the probabilities assigned by our commander to the alternatives (a) and (b), his "subjective probabilities." We are not interested at present in how they originate – whether the commander derived them from some systematic observations, or based them on some vague experience, or even possibly on mere prejudice. We shall explore, instead, whether these subjective probabilities can be defined in terms of his action. The suspicion that this can be done is suggested by the following. Suppose our commander is "almost certain" (subjective probability "a little below one") that (b) is true; then, is it not right to say that he "should" cross the 38° parallel? If he is "almost certain" that (b) is not true (subjective probability "a little above zero"), he "should" not cross. One may therefore be tempted to say that there exists a certain critical level for the subjective probability of (b), with the following property: if the subjective probability of (b) is above this level, our commander chooses (I); if it is below, he chooses (II). "Man is the measure" – if not of all things then, at least, of his own subjective probabilities, as revealed by his action.

A little reflection shows, however, that the commander's action cannot have been determined by his subjective probabilities, unless some – equally subjective – "values" or "utilities" had been attached to the outcomes of his actions – to the "gains," "losses" and "stalemates" entered in our table. As a matter of fact, we had already made a statement about the utilities of these outcomes, when we said that gains are preferred to stalemate and stalemate is preferred to losses. If the three outcomes are thus "completely
ordered," we can attach to them "utility numbers" which would be
ordered in a corresponding fashion: for example, 7, 5, 1; or 6, 4,
2; or log 6, log 4, log 2; — in short, any decreasing sequence of
three numbers. If any such sequence is used to designate utility
numbers, the outcome with a higher number will be preferred to
one with a lower number. Yet such ordering of the outcomes is
not sufficient to determine choice of action, even if the subjective
probabilities of outcomes of a given action are fixed.

Suppose, for example, that two U. N. officers, an American
and a South-Korean, agree that the Chinese are as likely as not to
counter-attack if we cross 38° parallel. Yet the American may
disfavor the crossing, while the South-Korean may advocate it.
That is, the American may prefer a stalemate to a 50-50 chance
of gains or losses; while the South-Korean (possibly reflecting his
government's eagerness to re-unite the country) has an opposite
preference. Thus, choice is determined by the complete ordering
(the "order of preferences") of probability distributions of out-
comes, not by the complete ordering of the outcomes themselves.

The complete ordering of probability distributions is consis-
tent with (though not necessarily equivalent to) assigning "utility
numbers" to probability distributions of outcome, as well as to the
outcomes themselves, according to the following rule: the utility
number attached to a given probability distribution of outcomes
equals the average of the utility numbers of the outcomes, weighted
with the respective probabilities of those outcomes.²

In our example, both the American and the South-Korean may
assign the same utility numbers to the best and the worst outcome:
say 100 to "gains," 0 to "losses." If they both want to maximize
the expected utility, then their assumed disagreement about what
to do reveals that, for the American, the utility of a stalemate is
larger than 50 — namely, larger than \(\frac{1}{2} (100) + \frac{1}{2} (0)\) — while
for the South-Korean it is less than 50. One can, in fact, imagine
an experiment in which a commander has to make (hypothetical)
decisions, being faced not only with the choice "a stalemate or a
50-50 chance of gains or losses," but also with choices such as
"a stalemate or a 60-40 chance of gains or losses," "a stalemate
or a 70-30 chance of gains or losses," etc. Suppose one would
thus find a break-even point: for example, suppose that the com-
mander has decided to cross the parallel if he felt that the chance
of gains is larger than or equal to 65 percent; and not to cross if
he felt the chance of gains is smaller than 65 percent. Then we
can say that, for him, stalemate is equivalent to a 65-35 chance of
gains or losses. Consequently, on a scale of utilities in which 100
"should" choose (II). Thus if we know that our commander was sure (rightly or wrongly) of his opponent's intentions and has done what he "should" have done, then his opinion about those intentions -- his assignment of probabilities to (a) and (b) -- would be revealed by his action.

A similar inference from action to opinion, under the assumption that the acting person is doing what he "should" be doing, can be made also in a more general case: namely, in the case when our commander is not necessarily sure as to the adversary's intentions, and thus associates them with probabilities that are not necessarily 1 or 0 but may be intermediate between 1 and 0. Such fractional probabilities we have not yet defined. Yet you have no doubt acted on the basis of judgments such as "an assertion is almost certain to be true" (= "is very probable"), "almost certain to be untrue" (= has small probability), "as likely as not to be true" (= has probability 1/2).

Let us look into this more carefully. To avoid terminological confusion, let us call the probabilities assigned by our commander to the alternatives (a) and (b), his "subjective probabilities." We are not interested at present in how they originate -- whether the commander derived them from some systematic observations, or based them on some vague experience, or even possibly on mere prejudice. We shall explore, instead; whether these subjective probabilities can be defined in terms of his action. The suspicion that this can be done is suggested by the following. Suppose our commander is "almost certain" (subjective probability "a little below one") that (b) is true; then, is it not right to say that he "should" cross the 38° parallel? If he is "almost certain" that (b) is not true (subjective probability "a little above zero"), he "should" not cross. One may therefore be tempted to say that there exists a certain critical level for the subjective probability of (b), with the following property: if the subjective probability of (b) is above this level, our commander chooses (I); if it is below, he chooses (II). "Man is the measure" -- if not of all things then, at least, of his own subjective probabilities, as revealed by his action.

A little reflection shows, however, that the commander's action cannot have been determined by his subjective probabilities, unless some -- equally subjective -- "values" or "utilities" had been attached to the outcomes of his actions -- to the "gains," "losses" and "stalemates" entered in our table. As a matter of fact, we had already made a statement about the utilities of these outcomes, when we said that gains are preferred to stalemate and stalemate is preferred to losses. If the three outcomes are thus "completely
and 0 are, respectively, assigned to gains and losses, the utility of stalemate (for that commander) will be 65, because \( (100 \times 65\%) + (0 \times 35\%) = 65 \). Thus, if we fix the "0" and "100" points in any arbitrary way (this is similar to the "0" and "32" points of the Fahrenheit scale, or the "0" and "100" of the centigrade scale of temperature), and if we assume that the man is "doing what he should do" and that he "should" maximize expected utility, then his actual responses to given probabilities of outcomes of his actions can be interpreted as revealing the utility numbers attached to each of those outcomes.³

Conversely, if the utilities attached by the decision-maker to alternative outcomes are known, it is possible to infer from his decisions the subjective probabilities that he attaches to alternative outcomes — provided that he "is doing what he should do" and that he "should" always maximize expected utility.

It is historically interesting that Thomas Bayes, one of the founders of the theory of probability, used just this kind of inference in order to define probability: In his posthumous essay (famous for another reason) we read⁴

"The probability of any event is the ratio between the value at which an expectation depending on the happening of the event ought to be computed, and the value of the thing expected upon its happening" (Definition 5); and

"If a person has an expectation depending on the happening of an event, the probability of the event is to the probability of its failure as his loss if it fails to his gain if it happens" (Proposition 2).

Note Bayes' phrase "ought to be computed." His definition is a norm of behavior. To your gain \((a_i, \text{ dollars, say})\) that is contingent upon an uncertain event, there "ought to" correspond in your mind a smaller but sure gain \((a, \text{ dollars, say})\) such that you are indifferent between gaining \(a_i\) upon the happening of the uncertain event, or gaining \(a\) with certainty. Hence, there is (or rather ought to be) in your mind also a ratio of these two numbers, \(a/a_i = p_i\), say. It is called by Bayes, probability of the event; while he calls it the "expectation." Suppose further that, if the event fails to happen, you neither gain nor lose: Then clearly \(a = a_i \cdot p_i + 0 \cdot (1 - p_i)\) is the average of gain and loss, weighted with their respective probabilities. You "ought to compute" this weighted average before making a choice. Of two uncertain gains choose the one with the higher value of expectation.
In Bayes' second quoted statement the person must have chosen to conclude a fair bet – that is, the expectation value \( p_1 a_1 + p_2 a_2 = 0 \), where the loss (taken as a positive number) equals to \(- a_2\) and has probability \( p_2 (= 1 - p_1)\). Hence \( p_1 / p_2 = - a_2 / a_1\).

No problem of ascertaining utility arises for Bayes because his gains and losses are money amounts. We can reformulate his definition of probability as the following normative statement (which we shall call "Bayes' norm"): 

Let the possible decisions be numbered \( 1, \ldots, n\); let the possible states of the world be numbered \( 1, \ldots, m\). Denote by \( a_{ij} \) the gain that the person will obtain if he takes the \( j\)-th decision, and the world is in its \( i\)-th state. Then, if the person behaves "as he should," there exists a set of \( m \) non-negative numbers \( p_i \) which add up to unity

\[
\sum_{i=1}^{m} p_i = 1
\]

and have the following property: one decision (say the \( j\)-th) is not preferred to another (say the \( k\)-th) when and only when

\[
\sum_{i=1}^{m} p_i a_{ij} \leq \sum_{i=1}^{m} p_i a_{ik} \quad (j, k = 1, \ldots, n)
\]  \( (1.1) \)

A number \( p_i \) is called the (subjective) probability of the \( i\)-th state of the world; the two sums in (1.1) are called expected gains associated with, respectively, the \( j\)-th and the \( k\)-th decision.

As an example (with \( m = n = 2\)), the possible gains of a man who has decided between buying and not buying stocks are tabulated on the following page.

If the subject decides according to the norm there postulated, then there must exist subjective probabilities \( p_1, p_2 (= 1 - p_1)\), with the following property: \( p_1 a_{11} + p_2 a_{21} \leq p_1 a_{12} + p_2 a_{22} \) if the subject does not prefer buying to not buying; and such that \( p_1 a_{11} + p_2 a_{21} \geq p_1 a_{12} + p_2 a_{22} \) if he does not prefer not buying to buying. Consequently, if he is indifferent between buying and not buying, both these relations are true so that
\[ p_1 a_{11} + p_2 a_{21} = p_1 a_{12} + p_2 a_{22} \]

\[ \frac{p_1}{p_2} = \frac{a_{22} - a_{21}}{a_{11} - a_{12}}, \text{ where } p_1 + p_2 = 1. \] (1.2)

(Note that if we had put, for simplicity, \( a_{12} = a_{22} = 0 \), we would obtain Bayes' "Proposition 2"). Thus by confronting the subject with varying quadruplets of prospective gains one can estimate the subjective probabilities he assigns to the stocks' rise or fall, respectively. For example if (1.2) is satisfied for \( a_{11} = -100 \) dollars, \( a_{21} = +300 \) dollars, and \( a_{12} = a_{22} = 0 \), then "the odds are" \( p_1 : p_2 = 3:1 \), in favor of falling stocks; or \( p_1 = 3/4, \ p_2 = 1/4 \).

<table>
<thead>
<tr>
<th>Decisions</th>
<th>States of the world</th>
<th>1.</th>
<th>2.</th>
</tr>
</thead>
<tbody>
<tr>
<td>j</td>
<td></td>
<td>Buy</td>
<td>Don't buy</td>
</tr>
<tr>
<td>i</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1. Stocks will fall</td>
<td>a_{11}</td>
<td>a_{12}</td>
<td></td>
</tr>
<tr>
<td>2. Stocks will rise</td>
<td>a_{21}</td>
<td>a_{22}</td>
<td></td>
</tr>
</tbody>
</table>

It is empirically ascertainable whether a subject's behavior is consistent with this norm. If it is, the subjective probabilities themselves are ascertainable. Consider, for example, the following "payoff matrices" — each being simply the Table I.2, with captions omitted and letters \( a_{ij} \) replaced by dollar figures:

\[
\begin{pmatrix}
-100 & 0 \\
300 & 0 \\
\end{pmatrix}, \begin{pmatrix}
-1 & 0 \\
3 & 0 \\
\end{pmatrix}, \begin{pmatrix}
-2 & -1 \\
4 & 1 \\
\end{pmatrix}, \begin{pmatrix}
-1 & 0 \\
2 & 0 \\
\end{pmatrix}.
\]

Table I.3

Suppose that the subject is indifferent between buying and not buying when the gains are as in the first matrix. This is simply the case.
already studied; it has revealed subjective odds 3:1 in favor of falling stocks. Now, the norm implies that if he is indifferent between buying and not buying when faced with the first matrix, he should also be indifferent between buying and not buying when faced with the second and third matrices, but should not buy when faced with the fourth matrix of Table I.3: because not only

\[ (-100) \frac{3}{4} + (300) \frac{1}{4} = (0) \frac{3}{4} + (0) \frac{1}{4} \] but also

\[ (-1) \frac{3}{4} + (3) \frac{1}{4} = (0) \frac{3}{4} + (0) \frac{1}{4} \] and also

\[ (-2) \frac{3}{4} + (4) \frac{1}{4} = (-1) \frac{3}{4} + (1) \frac{1}{4} \] while

\[ (-1) \frac{3}{4} + (2) \frac{1}{4} < (0) \frac{3}{4} + (0) \frac{1}{4} \].

Such empirical tests of consistency of a man's behavior with some given norm have been occasionally described as meaningless; with the consequence drawn that the (approximate) measurement of subjective probabilities has no meaning. The argument is as follows. Suppose the subject is indifferent between buying and not buying when faced with the first matrix of Table I.3, and also when faced with the fourth matrix. This need not imply that his behavior is inconsistent with Bayes' norm. The subject may have merely changed, during the time required by the experiments, his subjective odds, from 3:1 to 2:1 [since \((-1) \frac{2}{3} + (2) \frac{1}{3} = 0\)]. Suppose, on the other hand, that the subject, when faced with different payoff matrices, almost always makes the same choices as he would make if he maintained unchanged subjective odds and obeyed the Bayes' norm. Yet the subject may have behaved in contradiction with Bayes' norm; he may have changed his opinion about the odds, from experiment to experiment, in exactly such a fashion that the contradiction was masked.

This objection makes it necessary to reformulate every behavior norm by requiring that it should be "approximately" valid during a specified period of time. But what is meant by "approximate," and how to determine the time interval over which the norm is valid? This will depend on the use that has to be made of the norm in practice. The next two lectures give several illustrations of how the merely "approximate" description of human behavior is used for prediction and policy.

What we called the Bayes' norm is quite innocent of the concept of utility. Herein lies its limitation. Most decisions do not result in monetary gains and losses. The military gain, or loss, in our Korean example (Table I.1) is not expressible, either as a money amount or as any other single quantity. It is, at best, a combination of quantities (a "vector"): so and so many guns, prisoners, villages
taken. To be sure (and this will help us later), even when the alternative outcomes differ only because of the presence or absence of some "qualitative" attributes ("to live in the corn-belt," "to die in New York") one can express them as a combination of quantities, each possibly having only a finite number of values: e.g., life or death = 1 or 2; corn-belt or New York or West Coast = 1 or 2 or 3; so that life in the corn-belt = vector (1,1). But this may be inconvenient. For our purposes, it is unnecessary. Our Korean example has shown, in a preliminary way, that we can interpret choices by assigning subjective "utility numbers" to alternative events, the events themselves not having been expressed either as values of single objective variables (such as the number of guns or of prisoners) nor as combinations of such values (the number of guns and of prisoners). We have seen that, given the probabilities of outcomes, one can infer from manifest decisions to underlying subjective utilities, provided the subject has chosen "as he ought to." To formalize this statement, let us replace Bayes' norm by the following one:

Let \( A \) be the set of all possible outcomes of the subject's decisions. Let \( a_{ij} \) \((i = 1, \ldots, m; j = 1, \ldots, n)\) be that element of \( A \) which he associates with the \( i \)-th state of the world and his \( j \)-th decision. If the probability assigned by the subject to the \( i \)-th state of the world is

\[
p_i \quad (p_i \geq 0; \Sigma p_i = 1; i = 1, \ldots, m)
\]

and if he behaves "as he should" then there exists a real valued function \( u(a_{ij}) = u_{ij} \) such that one decision (say, the \( j \)-th) is not preferred to another (say, the \( k \)-th) when and only when

\[
\Sigma_{i=1}^{m} p_i u_{ij} \leq \Sigma_{i=1}^{m} p_i u_{ik} \quad (j, k = 1, \ldots, n).
\]

(3)

The function \( u(\ ) \) is called utility function; its value \( u_{ij} \) is called the utility of the outcome \( a_{ij} \). The two sums in (3) are called expected utilities associated, respectively, with the \( j \)-th and the \( k \)-th decision.

The norm just stated presupposes that probabilities are given to the subject; and his actual decisions reveal his system of utilities.
This is, in a sense, opposite to Bayes' norm because, if a subject acts according to Bayes' norm his actual decisions reveal his subjective probabilities; while the utilities are given, and are simply identical with amounts of money. We may call the norm just stated, the "Daniel Bernoulli norm." His concept of "moral expectation" is, in fact, identical with the "expected utility" of the statement just formulated, at least for the special case when each \( a_{ij} \) is a number, viz., the monetary wealth of the subject. More generally, the \( a_{ij} \) may denote any kind of situation which the subject expects to be the result of a given decision of his.

To exemplify Bernoulli's norm we may use the captions of Table I.2 but replace monetary gains \( a_{ij} \) by utility numbers \( u_{ij} \). We can, similarly, interpret the numbers entered in Table I.3 as utility numbers. Note that, in the case when these numbers measure the utilities associated with the monetary gains, it is the utilities not the dollar amounts that have, for example, in the first column of the first matrix, the ratio - 1:3. If the person knows (or thinks he knows) the odds to be 3:1 in favor of stocks falling, he will be indifferent between buying and not buying if faced with the first three matrices; and will not buy if faced with the fourth.

If the probability estimates of the subject are known, and if he behaves according to the "D. Bernoulli norm," it is possible to ascertain the utility numbers he assigns to various events, although only (as with centigrades and Fahrenheit degrees) up to two arbitrary constants. For example, suppose the subject is indifferent between buying and not buying stock when faced with the payoff matrix in dollars, given on the left part of the following table:

<table>
<thead>
<tr>
<th>Payoff in dollars</th>
<th>Payoff in utilities</th>
</tr>
</thead>
<tbody>
<tr>
<td>( a_{11} = -100 ) ( a_{12} = 50 )</td>
<td>( u_{11} = u (-100) = 0 ) ( u_{12} = u (50) = ? )</td>
</tr>
<tr>
<td>( a_{21} = 200 ) ( a_{22} = 50 )</td>
<td>( u_{21} = u (200) = 1 ) ( u_{21} = u (50) = ? )</td>
</tr>
</tbody>
</table>

Suppose (as before) that the odds are 3:1 in favor of the first row (falling stock prices). On the right side of Table I.4, we have fixed arbitrarily the zero and unity of the utility scale: \( 0 = \) utility of \(-$100\); \( 1 = \) utility of \$200. What is, on this scale, the utility of \$50? From the subject's indifference between the two decisions (represented, respectively, by the first and second column) we
conclude, using D. Bernoulli's norm: \( (0 \cdot \frac{3}{4}) + (1 \cdot \frac{1}{4}) = u \cdot (50) = 0.25 \). Thus, by varying the odds on which the subject is supposed to base his decision, and finding dollar payoff matrices such that the subject is indifferent between two decisions, one obtains the subject's utility scale for various outcomes of his decisions having arbitrarily fixed two points on this scale. This has been essentially the procedure of Mosteller and Nogee, already mentioned. The experimental difficulty includes, in particular, the ascertaining that the subject has "learned" the odds that he is told are prevailing. Moreover, analogous to the discussion we had on the occasion of the Bayes' norm, one has to meet the objection that the person may change his behavior parameters (his utility numbers, in the present case) during the experiment.

In the example just used, the outcomes \( a_n \) were dollar amounts; it was shown that the utility number corresponding to any given dollar gain or loss — and thus the "utility function of money gains" — can be ascertained from the actual choices of a subject assumed to obey a certain norm. Instead, the \( a_n \) might equally well have been, not dollar gains or losses, nor the values of any other single variable (number of prisoners taken), but vectors of quantities, or "qualitative" outcomes. Our example illustrates, incidentally, that Bayes' norm is not serviceable, not only because it has to deal with outcomes that are values of a single variable only, but also because it disagrees with the experience that people who can be expected to behave "as they should" may have non-linear utility functions of money: the gain of $50 is half-way between - $100 and + $200; yet the choice has proved consistent with the utility of $50 being not half-way but one-quarter of the distance between - $100 and + $200. It seems that Bayes overlooked the reasons which had led his predecessors (Daniel Bernoulli, Cramer) to insist on the difference between "physical wealth" and "moral wealth": the latter not being, in general, proportional to the former. Present day economists call the latter, "utility" (of a given monetary wealth or, more generally and precisely, of a given combination of goods in one's present and future possession). Even for the directors of a corporate firm, who decide on the basis of monetary profits and losses and not of ultimate "pleasures and pains," it is not true that a dollar has the same weight in the decision regardless of the dollars already earned. A loss leading to bankruptcy subtracts more from the utility than a profit of equal size adds to it — the utility function is non-linear. For an attacker bent on destruction, the utility is not a linear function of the physical damage inflicted upon the
enemy; rather, destruction beyond a certain point is useless to
the attacker. And so on.

The D. Bernoulli norm, formulated in its essence around
1730, was accepted by modern economists like Alfred Marshall.
But the assumption that a person computes and compares expected
utilities associated with his alternative decisions may appear as
complicated and artificial (not more or less, though) than does the
computation of expected monetary gains which seemed so evident
to Bayes as to make him base on it his definition of probability!
In our days, von Neumann and Morgenstern, the authors of the
Theory of Games, improved on their predecessors by providing
a set of postulates, equivalent to, but in some sense simpler, more
appealing than the postulate of maximizing expected utility. More
recent writers have tried to propose postulates that are still more
"transparent."

It is a remarkable psychological fact that a proposition may
be equivalent to a set of intuitively appealing propositions and yet
be not itself intuitively appealing. There would be otherwise no
need for mathematical textbooks! This is a human limitation.
For God, of course, the simplest axioms and the most difficult
theorems are equally obvious, and their roles interchangeable.
Let us, then, look at behavior postulates that are logically equiva-
 lent to but "psychologically" simpler than the Bernoulli norm.

Let A, B, C, . . . denote probability distributions of outcomes;
(in a special case a "distribution" may assign probability 1 to one
event, and 0 to all others; it is, in this case, identical with that single
event). Denote by ABp the following probability distribution:
"probability p that the distribution is A; probability 1-p that the
distribution is B." Let A ≤ B mean "A is not preferred to (or
"not better than") B." "A = B" will be written instead of "A ≤ B
and B ≤ A" and will mean "the subject is indifferent between A
and B." "A < B" will be written instead of "A ≤ B and not B < A"
and will mean "A is worse than B." Then the subject "should" obey
the following rules:

I. The A, B, C, . . . are "completely ordered" by the
relation ≤ ; that is,

1) A ≤ B or B ≤ A ("comparability");

2) if A ≤ B and B ≤ C then A ≤ C ("transitivity").

II. If A = B then CAp = CBp.

III. If A < B and B < C then there exists a unique p such
that B = Acp.
Rule I needs no explanation. A verbal comment on Rule II can be provided as follows. If A is not preferred to B and B is not preferred to C, I am indifferent as between the two. I am now offered two lotteries – one promising either C or A, with odds p:(1 - p), and the other promising either C or B, with the same odds. Thus the same outcome, C, with probability p, will occur on both lotteries; and if C does not occur then the first lottery will give me A, while the second will give me B. But I am indifferent between A and B. Hence, there is no reason for me to prefer one lottery to another.

Rule III asserts, in a rather strong form, the "continuity" of the preference relation. Actually a milder formulation of Rule III, and also of Rule II would suffice to obtain the needed result, viz., the "D. Bernoulli norm." Our stronger formulation gives an easier, though more superficial, insight into the matter, although, even so, I shall have merely to hint at the proof. Note first that our rules permit us to "calibrate" a utility scale as one does a thermometer. Choose three prospects A, B, C such that I am not indifferent between any two of them. Then, by Rule I, they can be ordered in a unique fashion. Suppose this order is: B is preferred to C, and C to A. On a horizontal straight line, plot them in the order (from left to right): A, C, B (Fig. I.5). Call a number u(A) the utility of A; this will be also the utility of all prospects A', A'', which, for me, are neither better nor worse than A. Similarly call u(B) the utility of B and its equivalents. What number u(C) shall we then have to assign to the utility of C and of its equivalents? By Rule III there exists a unique probability p such that, to me, C is equivalent to BAp. Call u(C) the number equal to pu(B) + (1 - p) u(A). In particular if we arbitrarily fix u(A) = 0 and u(B) = 1 then u(C) is simply equal to p. We can now place C on our scale (Fig. I.5), making the distance AB to represent one utility unit, and the distance AC to represent p utility units (0 < p < 1). By Rule II, the number p will also measure the utility of all lotteries such as A'BP, AB'p, A'B'p, where A' is equivalent to A and B' is equivalent to B. We can thus "calibrate" the utility scale for all prospects which are not better than B and not worse than A: they will be expressed by appropriate numbers ranging from 0 to 1. If a prospect D is worse than A, then A lies on our scale between D and B. Suppose the odds (those required by the Rule III) which make A equivalent to a lottery promising B or D are q:(1 - q); that is, I am indifferent between A and BDq. Then call the utility of D, u(D), a number such that u(B)·q + u(D) (1 - q) = u(A). With u(A) = 0 and u(B) = 1 this makes u(D) = - q/(1 - q).
Fig. I.5

<table>
<thead>
<tr>
<th>Outcomes:</th>
<th>D</th>
<th>A</th>
<th>C</th>
<th>B</th>
<th>E</th>
</tr>
</thead>
<tbody>
<tr>
<td>Utilities:</td>
<td>$-q/(1-q)$</td>
<td>$0$</td>
<td>$p$</td>
<td>$1$</td>
<td>$1/r$</td>
</tr>
</tbody>
</table>

\[
\overline{AC} = p(1-p) \quad \overline{CB} = q(1-q) \\
\overline{BA} = u(A) = u(B)q + u(D)(1-q) \\
\overline{AB} = u(E)r + u(A)(1-r) \\
\]

By an analogous procedure utility numbers are assigned to prospects – such as E in our figure – that are better than B. We have thus calibrated the utilities of all possible prospects. If we now assign the numbers 0 and 1 to the utilities not of A and B, but of some other pair of (non-equivalent) prospects, the origin and the unit of measurement of utilities is changed (just like we can measure the altitude of a mountain from the sea level as well as from a lake level, in yards as well as in feet). But in any case, the Bernoulli norm will follow. For, to take up first the case when a prospect involves only two alternative outcomes, we have seen (see Figure I.5) that the utility of such a prospect equals the sum of the utilities of the outcomes, multiplied by respective probabilities; i.e., the utility of a prospect equals its "expected utility." And since higher utilities are preferred to smaller ones, a prospect will not be preferred to another one unless its expected utility is higher. To extend this to the case of prospects involving more than two alternative outcomes is a matter of easy algebra. We shall skip it.

Our Figure I.5 could be also used to explain the approach of Bayes, if we reverse the problem and consider the outcomes as represented by given distances from the zero-point, measured in money units: A = 0 dollars, B = 1 dollar. Then Bayes defines the probability of, say, C, by the ratio of the distances $\overline{AC} : \overline{AB}$. This "calibrates" probabilities.

We can say that in the Bayes norm the utilities were assumed objectively given (viz., identical with money amounts), and the probabilities were derived as parameters of the subject's behavior, as his "degrees of belief." Contrariwise, the Bernoulli norm assumes probabilities to be objectively given and derives subjective utilities. We had criticized Bayes' norm for its identifying the outcomes of decisions with the values of a single variable (such as the money gain) observable without reference to the subject. But a similar
criticism applies to the Bernoulli norm. The probabilities on
which the subject bases his action need not be identical with some
objective properties of chance devices (cards, dice) which the ex-
perimenter uses. This was observed by the English mathematician
and logician, F. P. Ramsey. He shows that manifest decisions can
be thought of as revealing both the subject's probabilities and util-
ities. We can call "Ramsey's norm" the following statement:

Let \( A \) be the set of all possible outcomes of the
subject's decisions. Let \( a_{ij} \) \((i = 1, \ldots, m; j = 1, \ldots, n)\)
be that element of \( A \) which he associates with the \( i \)-th state
of the world and with his \( j \)-th decision. If the subject be-
haves "as he should," then there exists a unique set of num-
bers \( p, \{p_i \geq 0; \sum p_i = 1; i = 1, \ldots, m\} \) and a real valued
function \( u(a_{ij}) = u_{ij} \) such that one decision (say, the \( j \)-th)
is not preferred to another (say, the \( k \)-th) when and only when

\[
\sum_{i=1}^{m} p_i u_{ij} \leq \sum_{i=1}^{m} p_i u_{ik} (j, k = 1, \ldots, n).
\]

The numbers \( p_i \) are called subjective probabilities (or de-
grees of belief). The function \( u(a_{ij}) \) is called utility func-
tion; its value \( u_{ij} \) is called the utility of the outcome \( a_{ij} \).
The two sums in (3) are called expected utilities associated,
respectively, with the \( j \)-th and the \( k \)-th decision.

A set of behavior rules, equivalent to Ramsey's norm but
"psychologically" more appealing was proposed by Ramsey himself
and, in our days, by de Finetti and, especially, by Savage.7 It would
be difficult to convey their proofs in this lecture. But it is possible
to indicate that subjective probabilities can be defined on the basis
of observable decisions of a subject, even if one rejects the identifi-
cation (practised by Bayes) between utilities and money gains; and
to indicate how this definition of probability is related to the norm
requiring maximization of expected utility. Consider the matrix in
Table I.6 of outcomes \( a_{ij} \), with \( m = n = 2 \). Suppose I am indifferent
between the two decisions. Then Ramsey will define the degrees of
belief in (or the subjective probability of) each of the two states of
the world as \( p_1 = p_2 = 1/2 \). For example, suppose an urn contains
white and red balls and I am given the choice between the following
two options: (1) if I draw white, I die; if red, I live; (2) if I draw red,
I die; if white, I live. If I am indifferent between the two options (but not between life and death) I have equal degrees of belief (defined as 1/2) that the ball drawn will be white or red. We postulate that the subject has consistent degree of belief. That is, he will remain indifferent between any two options represented by the matrix of Table I.6, regardless of what meaning is given to the outcomes \( \alpha \) and \( \beta \). Having thus defined subjective probability = 1/2 (but not yet any other probabilities) enables Ramsey to define a scale of measurable utilities, as follows. An outcome \( \gamma \) is said to be "half-way" between \( \alpha \) and \( \beta \) if the subject is indifferent between \( \gamma \) and the prospect of having \( \alpha \) or \( \beta \) with subjective odds 1/2 : 1/2.

It is again postulated that the subject is consistent, in the sense that \( \gamma \) will remain "half-way" between \( \alpha \) and \( \beta \) regardless of the particular pair of the states of the world to which the subjective probabilities 1/2, 1/2 are attached. Now Ramsey can define the utility of \( \alpha \) as zero, the utility of \( \gamma \) as 1, the utility of \( \beta \) as 2, and (with the help of some further axioms on order and continuity) he can construct the scale of utilities. This brings him to the same point at which Bayes had made the start, except that Bayes' money amounts are replaced by utility numbers. Like Bayes, Ramsey can now calibrate probabilities: these are ratios between certain differences between utilities – as on our Figure I.5. The norm that prospects with higher expected utilities should be preferred to those with lower ones follows immediately.

It is worth noting that the logic of connecting between "measurable," (and not merely "completely ordered") utilities as required by the Bernoulli and the Ramsey norm has been patterned, to a large extent, after Hilbert's "Foundations of Geometry" in which the relation was established between points ordered along a straight line.
and the real numbers measuring distances. In a more general sense, too, discussion of behavior norms is similar to that of foundations of geometry. Various geometries – various internally consistent sets of axioms and theorems – are possible. So are various sets of behavior rules. Such consistent systems do not assert anything about empirical reality – e.g., about physical or psychological phenomena. Yet they do help to order these phenomena, by comparing facts with ideals. The normative discussion of behavior shares this feature with mathematics and logic. In addition, it shares with logic another feature: certain modes of thinking and acting may be "preferable" to others. For example, it is preferable, when one is told that all men are mortals, not to conclude that all mortals are men; and it is preferable, when choosing A in preference to B, and B in preference to C, not to choose C in preference to A. We are warned against "bad logic" as well as against "inconsistent behavior" (the latter is even called sometimes "illogical behavior"). The warning is of a pragmatic nature: bad logicians won't survive. In the last lecture, we shall return to this practical aspect of normative social science, and of probabilities in social science.

The approach of Ramsey and of D. Bernoulli is more congenial to social scientists than that of Bayes because Bayes confined himself to outcomes which are money gains. But still more important, the common feature of the approach of both Ramsey and Bayes, is particularly congenial to social scientists. It permits us to deal with "probabilities of single events." While the XVIII century Continentals, in Paris or Petersburg, indulged in dice and cards, and watched the frequencies of repeated events, the English seem to have thought more of horses and fighting cocks and wrestlers. The betting books of Oxford colleges are full of guesses such as "Napoleon is dead" (dated November 1812) accompanied by the betting ratios – whether in guineas or in bottles of port: This is right after the social scientist's taste! He is often unable to observe repeated events.

But, then the following question arises: if I happen to be able to observe repeated events, how shall I use those observations in computing my "degrees of belief," which are my guide in decision-making? Shall I, for example, equate my degree of belief in drawing a white ball from an urn, to the relative frequency of such drawings in the past, at least as an approximation? In short: is there any relation between subjective probabilities and the statistical method? Is there any sense in the statement: "The chances that I shall survive the operation are the same as getting a royal flush"?
(These questions are important for our future discussion, because in the two remaining lectures, statistical methods will be applied to human and social phenomena and to policy decisions.)

These questions were, in essence, answered in the affirmative by Bayes himself, and again, in our time, by Ramsey. To justify the affirmative answer, one first shows that the "law of large numbers," properly interpreted, does follow from the properties of subjective probabilities. To do this, it will be convenient to use the concepts of "conditional probabilities" and of "independence." We shall define them as we proceed. (These concepts will also prove useful in Lectures 2 and 3.)

Let \( \bar{W} \) be the negation of the state of the world \( W \), and let \( \bar{V} \) be the negation of the state of the world \( V \); and suppose that the subject is indifferent between the following options:

\[
\begin{align*}
\alpha & \text{ if } \bar{W} \\
\beta & \text{ if } W
\end{align*}
\]

and

\[
\begin{align*}
\alpha & \text{ if } \bar{W} \\
\beta' & \text{ if } W \text{ and } V \\
\beta'' & \text{ if } W \text{ and } \bar{V}
\end{align*}
\]

For example: let \( \alpha = \) death, \( \beta = \) survival; \( \beta' = \) early arrival; \( \beta'' = \) late arrival; \( \bar{W} = \) plane has engine trouble; \( V = \) plane is fast.

Then the option on the left hand is: to die if the plane has engine trouble; to survive if it has not. The option on the right hand is:

to die if the plane has engine trouble; to survive and arrive early if the plane has no engine trouble and is a fast one; to survive and arrive late if the plane has no engine trouble and is a slow one.

Our assumption is, then, that I am indifferent between the option on the left hand and the option on the right hand. That is, by Ramsey's postulate, the expected utility of the left-hand option is equal to that of the right-hand option. But the expected utility of the left-hand option is clearly

\[
u(\alpha)p(\bar{W}) + u(\beta)p(W),
\]

where \( u(\alpha), u(\beta) \) are the utility numbers attached to the outcomes \( \alpha \) and \( \beta \), respectively, and \( p(\bar{W}) \) and \( p(W) \) are the probabilities of the alternative states of the world \( \bar{W} \) and \( W \) respectively. In a similar notation, the expected utility of the right-hand option is
\[ u(\alpha) p(\bar{W}) + u(\beta') p(W \text{ and } V) + u(\beta'') p(W \text{ and } \bar{V}). \]

Hence the indifference between the two options implies the equality of the two expressions we have written. This equality implies — since the term containing \( u(\alpha) \) cancels out —

\[ u(\beta) p(W) = u(\beta') p(W \text{ and } V) + u(\beta'') p(W \text{ and } \bar{V}). \quad (1.3) \]

Let us now give the two quantities

\[ \frac{p(W \text{ and } V)}{p(W)} \text{ and } \frac{p(W \text{ and } \bar{V})}{p(W)}, \]

the names "conditional probability of \( V \) given \( W \)" and "conditional probability of \( \bar{V} \) given \( W \);" and denote these two quantities, respectively, by

\[ p(V \mid W) \text{ and } p(\bar{V} \mid W). \]

Then our last equation becomes — dividing by \( p(W) \) —

\[ u(\beta) = u(\beta') \cdot p(V \mid W) + u(\beta'') \cdot p(\bar{V} \mid W). \quad (1.4) \]

Now define "independence" as follows. Suppose the state \( x_2 \) is "irrelevant" to the state \( x_1 \) in the sense that the conditional probability of \( x_2 \) given \( x_1 \) is the same as the probability of \( x_1 \):

\[ p(x_1 \mid x_2) = p(x_1). \]

By the definition of conditional probability this means

\[ \frac{p(x_1 \text{ and } x_2)}{p(x_2)} = p(x_1), \]

\[ p(x_1 \text{ and } x_2) = p(x_1) \cdot p(x_2). \quad (1.5) \]
Thus the "irrelevance" of \( x_a \) for \( x_i \) is equivalent to the relation (1.3), which will define "independence" between \( x_i \) and \( x_a \). Since (1.3) is symmetrical in \( x_i \) and \( x_a \) the following are equivalent: independence between \( x_i \) and \( x_a \); irrelevance of \( x_a \) for \( x_i \); and irrelevance of \( x_i \) for \( x_a \). One can extend the concept of independence to, say, \( N \) events and define it by

\[
p(x_1 \text{ and } x_2 \text{ and } x_3 \ldots \text{ and } x_N) = p(x_1)p(x_2)p(x_3)\ldots p(x_N). \tag{1.6}
\]

Thus if one tosses coins, one can compute the probability that the independent events \( x_1, \ldots, x_N \) will be all "heads"; or that the first two are heads and the rest tails; or that some other preassigned proportion of tosses be heads. These computations are a matter of arithmetic. As the result, one finds (one version of) the law of large numbers: if the "probability of heads" equals \( p \) then the "probability that the proportion of heads in a long sequence of independent throws will converge to \( p \)" equals 1. Remember that the probabilities in question (the numbers \( p \) and 1 in our last sentence) are subjective degrees of belief, defined as guides in choosing between actions.\(^\text{10}\)

We shall now sketch out the proof that any "preconceived" (or "a priori") degrees of belief "should" be modified in the light of experience, in a certain well-defined way, which is fundamentally the one at the basis of sampling statistics. (Note the "should"! We are talking of norms!).

Suppose that my degree of belief in drawing a black ball from an urn, on condition that the world is in state \( W \), is \( \pi \). This quantity is also called the "likelihood of the state \( W \) on the basis of an observed drawing of a black ball." If the urn contains only black and red balls, then clearly the likelihood of \( W \) on the basis of an observed drawing of a red ball is \( 1 - \pi \). Similarly, let us denote by \( \bar{\pi} \) and \( 1 - \bar{\pi} \), respectively, the likelihoods of the alternative state, \( \overline{W} \), on the basis of an observed drawing of a black or of a red ball. Let \( x_i \) be the \( i \)-th independent observation, and let "\( x_i = b \)" and "\( x_i = r \)" mean that this observation is the drawing of a black or a red ball, respectively. We can then rewrite our definitions as follows:

\[
\pi = p(x_i = b \mid W) \; ; \; 1 - \pi = p(x_i = r \mid W)
\]

\[
\bar{\pi} = p(x_i = b \mid \overline{W}) \; ; \; 1 - \bar{\pi} = p(x_i = r \mid \overline{W}).
\]
Consider in particular, the first observation, \( x_1 \), and write the degree of belief that the world is in the state \( W \) and that the first drawing will be "black," thus:

\[
p(W; \text{ and } x_1 = b).
\]

On the other hand, denote separately the degrees of belief that the world is in state \( W \) (regardless of observations) and that the first drawing is black, by

\[
p^0(W) \text{ and } p(x_1 = b),
\]

respectively. If we now recall our definition of conditional probabilities, we have

\[
p(W; \text{ and } x_1 = b) = p(x_1 = b|W) \cdot p^0(W)
\]

and also

\[
p(W; \text{ and } x_1 = b) = p(W|x_1 = b) \cdot p(x_1 = b).
\]

The quantities \( p^*(W) \) and \( p(W|x_1 = b) \) are called, respectively, the \( a \) \( priori \) and the \( a \) \( posteriori \) degree of belief in \( W \); while \( p(x_1 = b|W) \) is, as we recall, the likelihood \( \pi \). Comparing the last two equations, we obtain

\[
p(W|x_1 = b) \cdot p(x_1 = b) = \pi \cdot p^0(W).
\]

And similarly (replacing \( W \) by \( \bar{W} \) and therefore \( \pi \) by \( \bar{\pi} \))

\[
p(\bar{W}|x_1 = b) \cdot p(x_1 = b) = \bar{\pi} \cdot p^0(\bar{W}).
\]

Dividing the last equation into the preceding one, one obtains a simple form of the celebrated "Bayes Theorem":
\[
\frac{p(W|x_1 = b)}{p(\bar{W}|x_1 = b)} = \frac{p^0(W) \cdot \pi}{p^0(\bar{W}) \cdot \bar{\pi}}
\]

(1.7)

In words: the *a posteriori* degree of belief in a certain state of the world, given a certain observation, is proportional to the product of the *a priori* degree of belief in that state of the world times the likelihood of that state of the world given the observation. In this way, the observation leads to modifying the *a priori* degree of belief. The likelihoods, \(\pi\) and \(\bar{\pi}\), operate as "correction factors." If the observation \(x_i\) were "red" instead of "black," the correction factors would be \(1 - \pi\) and \(1 - \bar{\pi}\), respectively.

Suppose now the process is repeated: we take \(N\) independent observations \(x_1, x_2, \ldots, x_N\), each time using the previously corrected degree of belief, correcting it further in the light of the new observation. Suppose we drew a black ball \(n\) times, (and therefore a red ball \(N-n\) times). Applying Bayes Theorem,

\[
\frac{p(\bar{W}|x_1, \ldots, x_n)}{p(W|x_1, \ldots, x_n)} = \frac{p^0(\bar{W}) \cdot \pi^n (1 - \bar{\pi})^{N-n}}{p^0(W) \cdot \pi^n (1 - \pi)^{N-n}}
\]

(1.8)

But, by the law of large numbers, as stated above, if the true state of the world is \(W\), then the degree of belief is 1 that the proportion of black drawings, \(n/N\) will converge to \(\pi\). Therefore, the number of black drawings, \(n\), will converge to \(\pi N\); and the ratio between the *a posteriori* degrees of belief into \(\bar{W}\) and \(W\), as given in our last equation, will approach

\[
\frac{p^0(\bar{W}) \left(\bar{\pi}^n (1 - \bar{\pi})^{1-\pi}\right)^N}{p^0(W) \left(\pi^n (1 - \pi)^{1-\pi}\right)}.
\]

We see that as the number \(N\) of observations increases, the role of the *a priori* degrees of belief into the two alternative states of the world, \(p^0(\bar{W})\) and \(p^0(W)\), diminishes, overshadowed by the results of observations. Furthermore simple calculus shows that the expression

\[
\bar{\pi}^n (1 - \bar{\pi})^{1-\pi},
\]
considered as a function of a variable \( \pi \), has its maximum when \( \pi = \pi \). Hence, in (1.6), the quantity that is raised to the \( N \)-th power, is a ratio of a smaller positive number to a larger one; and its \( N \)-th power approaches 0 as \( N \) increases. Therefore, of the two \textit{a posteriori} degrees of belief that form the fraction on the left side of (1.6), and that must add up to 1, the numerator converges to 0 and the denominator to 1. That is, the degree of belief that will be assigned to the true state of the world \( W \) will approach 1. Hence, the investigator who will repeatedly apply Bayes Theorem and choose his decisions on the basis of degrees of belief thus computed will in effect believe that he would obtain a better result than if he would not do so. Perhaps this is what writers on probability mean when they use the expression "practically certain" or "certain for all practical purposes,"\textsuperscript{11} - an expression we shall also have the opportunity for using in the next two sections.
We shall concern ourselves now with the uncertainties encountered by the social scientist when he tries to predict the behavior of people. Such prediction is rarely exact. It is usually "probabilistic" or statistical, even when the sample used is very large and even when the prediction is made, not about individuals, but about large aggregates of people. After discussing the probabilistic character of descriptive social science, I shall give examples illustrating an important methodological problem that has recently occupied statisticians as well as social scientists: that of identifiability of structural characteristics. This will throw some light on the following more general fact: whether statistical data can yield the desired prediction depends not only on the size of the sample and on the goodness of statistical formulae, but also on the nature and validity of the assumptions which the investigator had to make before processing a given set of data or, preferably, even before collecting them. These assumptions (sometimes called "a model") are, of course, based on formerly acquired knowledge. In the case of a social scientist, this is often the knowledge of "plausible" or of "meaningful" relations.

NORMS VERSUS HABITS.

We discussed in the first lecture some norms "recommended" to decision makers who face uncertainty. Those norms or behavior postulates were similar to the rules of logic or geometry. It was not asserted that such norms were fully obeyed by all or even a sizable proportion of men or women, in our own or any other civilization, just as logicians and mathematicians do not assert that all or the majority of their countrymen or of members of any other society are immune to errors of logic or arithmetic. It is merely recommended that those errors be avoided. Recommended norms and actual habits are not the same thing.

As a matter of empirical psychology it may be interesting to find in what manner a given individual deviates from such norms: how often is he apt to fall victim to a particular sophism, or to have trouble with his sums, or—nearer to our field—to be inconsistent in his preferences. As a matter of social science in general, we
may be interested in the ethical and social conditions which affect the frequency of deviations from norms of reasonable thinking, counting and choosing. This knowledge of conditions affecting people's behavior is, first, a matter of scientific curiosity. But it has also its practical side. If we know what makes people more or less illogical, or mathematically inept or poor decision makers, we may also find how best to enable them to learn the "recommended" type of behavior—how, for example, they can get the habit of "stopping to think." The normative and the descriptive analysis complete each other.

In the previous section, we did not specify the source of the uncertainty that faces the decision maker. It may be uncertainty about nature. It may be uncertainty about the actions of other men. In our initial example, the U.N. commander had to puzzle out whether his Chinese adversary is or is not "aggressive" (in a well defined sense of producing certain observable results under specified circumstances). This is uncertainty about people. If our commander estimates (or acts as if he had estimated) that the odds for the Chinese command having "aggressive" designs are such and such, he has done a bit of descriptive social science. It may happen that, in my best judgment, my adversary behaves as he "should" behave according to the norms of reasonableness which I recommend to myself. Any application of the Theory of Games (in its present form) is based on the assumption that this symmetry of behavior norms is an actual fact. However, this need not always be the case, nor need it be a useful approximation to reality. It may be more useful to actually study my adversary, as I would study weather or soil or any other uncertain natural phenomenon—provided such study of other people's behavior is feasible and not too costly. A commander—or, for that matter, a diplomat, or a labor union representative—will combine the theory of games ("What would I reasonably do if I were in the other fellow's position?") with as good an intelligence system as his resources can afford, whether by sending spies or by employing anthropologists. And, of course, what was said for a problem arising in fighting an opponent is also true for problems arising in forming an alliance, in building and operating within a social organization—in short in making decisions whose outcome will depend on actions of my fellow men about whom my knowledge is uncertain.
PREDICTION

Uncertain knowledge is not ignorance. Nineteenth-century social scientists were fascinated by "iron laws of nature," when they took eighteenth-century physics for their ideal. It was fashionable, a hundred years ago, to speak of the iron law of demand and supply or (with Karl Marx) to claim predictability for society's future. We have learned today—possibly reflecting some trends in physical sciences—to be quite happy when we can make a prediction only with some (specified) probability. For the "reasonable" man of our last lecture, a decision maker who weighs utilities with probabilities and maximizes the weighted sum, this kind of knowledge is certainly not useless! At the same time, such knowledge is usually the best that we can ever hope to have, in the field of social sciences, where uncertainties of our physical environment are topped by the diversity and capriciousness of human nature.

Briefly, empirical social science consists of statements about probability distributions. As a trivial but useful example, suppose a social statistician is hired to find what makes people smoke much or little or not at all. Or he wants to find what makes people communists. Let $x$ denote the number of cigarettes smoked by a certain man on a certain day ($x$ may be 0, 1, 2, ..., a many-valued variable); or let $x$ denote a certain man's being or not being a communist (here $x = 0$ or 1: a two-valued variable). What is the probability that $x$ will have a certain value? The tendency to believe in smoking, or in communism, will depend on certain individual characteristics among which one will expect to find the age, sex, education, occupation of the subject; but also his income, the characteristics of his parents and siblings ... but possibly also his past income and occupation ... but then also his whole past history and that of his ancestors, and, in fact, his whole genetic and cultural endowment. This is a long list of factors. Call them $z^{(1)}, z^{(2)}, \ldots, z^{(N)}$, with $N$ very large. We could make the desired prediction with certainty, for a subject whose $z^{(1)}, \ldots, z^{(N)}$ are known, if we also knew which combinations of the values of those variables make $x = 0$, which combinations make $x = 1$, etc. That is, the social scientist can predict $x$ from $z^{(1)}, \ldots, z^{(N)}$ if there exists a function

$$x = \phi (z^{(1)}, \ldots, z^{(N)})$$ (2.1)
(say) and if he has been able to find this function on the basis of his previous observations on \( x, z^{(1)}, \ldots, z^{(n)} \) for each of a sufficient number and variety of persons. However, a complete list of factors that may influence the man's smoking habits or his political beliefs is hopelessly long! The social scientist picks out a smaller number of factors -- say \( z^{(1)}, \ldots, z^{(h)} \) -- and replaces the assumption (2.1) of an "exact" relationship by a "probabilistic" assumption, as follows: The variable \( x \) is a random variable (also called a chance variable or a stochastic variable); and the probability with which it takes any given value is a function of the variables \( z^{(1)}, \ldots, z^{(n)} \) only. In other words, there exists a "conditional probability of \( x \) given \( z^{(1)}, \ldots, z^{(n)} \),"

\[
\pi(x|z^{(1)}, \ldots, z^{(n)}),
\]  

(2.2)

say. With the assumption thus changed, the predictor's task is to find, not the function \( \phi \) in (2.1) but the function \( \pi \) in (2.2), from the values of \( x, z^{(1)}, \ldots, z^{(n)} \) that he had observed on a sufficient number and variety of individuals. He can then state, for some new individual, with given \( z^{(1)}, \ldots, z^{(n)} \) but an unknown \( x \), the probability distribution of \( x \): "the probability that Tom Smith, elevator boy at the State Department, middle-aged and unmarried, is a communist is 0.001%"; "the probability that Mrs. Brown, fashion editor, divorced, smokes 0, 1, \ldots, 30 cigarettes per day is, respectively, 1%, 5%, \ldots, 0%." 

A very special, but convenient and familiar, form of this assumption will serve as illustration. One assumes that there exist \( n+2 \) numbers \( \mu, \sigma, \alpha_1, \ldots, \alpha_n \) with the following property: the difference (called "disturbance")

\[
x - \sum_{i=1}^{n} \alpha_i z^{(i)}
\]  

(2.3)

is independent of the \( z^{(i)} \) and has normal distribution with expectation (mean) \( \mu \) and standard deviation \( \sigma \). In this case, the conditional distribution (2.2) of \( x \), given \( z^{(1)}, \ldots, z^{(n)} \), is also normal. Its expectation is

\[
\mu + \sum_{i=1}^{n} \alpha_i z_i
\]
and is called the "explained part of x." Its standard deviation, σ, is identical with that of the disturbance (2.3). The numbers α, can be regarded as "weights" (not necessarily all positive) attached to the n explanatory variables; each of those variables contributes an additive component \( \alpha z^{(i)} \) of x. If one could find, on the basis of past observations on the x and the \( z^{(i)} \) of a sufficient number and variety of individuals, then, for any new individual, with known \( z^{(i)} \) but unknown \( x \), one would be able to make a "probabilistic" prediction such as, for example: the odds are 19:1 that this man's x lies in the interval

\[
\mu + \sum \alpha_i z^{(i)} \pm 2\sigma
\]  

(2.4)

The looks of this interval are familiar enough from textbooks and routines of research workers in social science. I am afraid its understanding is often confused by complications irrelevant from our point of view. One complication is that the constants \( \mu, \sigma, \) and the \( \alpha \) cannot be determined exactly from a finite number of observations. One can only determine their "estimates". These are suitably chosen functions of a finite number of observed values of the variables \( x, z^{(1)}, \ldots, z^{(n)} \). Since these variables are [or at least \( x \) is -- according to (2.2)] random, the estimates are also random variables (they are "subject to sampling errors") so that the interval (2.4), computed from those estimates, it itself "wobbly." But this is beside the point. We are not interested in sampling theory today. We can assume today that the investigator has collected a very large sample indeed; and that he has used appropriate formulas to compute estimates that are practically certain to approach the estimated constants as the sample increases (such estimates are called "consistent"). Our point is that, even using an infinite sample and consistent estimates, he will not be able to predict x from a finite number (n) of explanatory variables except in a "probabilistic" fashion typified by the statement that "x will fall within the interval (2.4) with probability 0.05."

Another concept that we wish to disregard here is that of "errors of observation," a term originating in observatories and physical laboratories. There the number N of possible explanatory factors is supposed to be small enough to be manageable; and the fact that the exact, non-probabilistic assumption (2.1) is not satisfied is (or was till the advent of modern statistical physics) blamed upon the frailty of the observer's eye and of his other instruments. In social science, human frailty is shared between the observer
and his object. After repeatedly asking a blindfolded man which of two weights is heavier, the psychologist may have to record a frequency distribution of contradictory statements: "The first weight is heavier," "It is lighter," or "No difference." The psychologist will have to formulate his conclusions and predictions in a probabilistic manner even if he is sure that his own errors of observation (confusing the weights or misunderstanding the subject’s responses) are negligible. A "random disturbance" originating in the observed person not in the observer will always be present. Furthermore, even if it should be possible to design laboratory conditions so as to make both the errors of observation and the random disturbance of the observed phenomenon negligible, this cannot be achieved with data collected outside of laboratory walls -- like the data on the smoking or the communist persuasion of persons. The variables, immense in number (N-n), which the investigator of such a problem has decided to leave out of account, are supposed by him to show their joint effect in the random disturbance (2.3). For example, each of the subject's ancestors has contributed something to his make-up, but each only a little: just like each of the innumerable causes that determine an honest card deal contributes a little. Such joint effects of a large number of separately insignificant causes are empirical phenomena that have been idealized in the mathematical concept of a random variable and a probability. An error of observation and a random disturbance, as defined, may both have their origin in those numerous, separately insignificant causes. However, if you prefer to ascribe the randomness of errors and disturbances to the free will of individuals, you may do so. This will not affect the choice of investigation methods.

The selection of the n explanatory variables, and the assumption that the difference (2.3) is independent of z, and has a normal probability distribution [or, more generally, that x has some probability distribution (2.2) for any fixed value of the n variables selected as explanatory ones] may be false. If it is, it will lead to probabilistic prediction statements that are false, even if they are based on consistent estimates from very large samples. This suggests, of course, that one should test predictions like (2.4) against new facts; and then revise the assumption (2.2) if necessary. For example, one may change the list of explanatory variables or replace the linearity and normality assumptions by other ones. Thus the selection of assumptions need not be entirely based on pre-scientific knowledge but can be helped by progressive experience.
AGGREGATION

Large parts of social science deal, not with single individuals but with averages or aggregates. It has been occasionally remarked that random variations of behavior from one individual to another (or, for the same individual, from one day to another) tend to "cancel out" when one deals with masses of man. Tsere-telli, a political leader of the pro-Communist period of the Russian revolution, even went so far as to say (in predicting the failure of Communists to win the masses!), "Persons can err, masses never"; and social scientists are apt to quote Boyle's law of gases, which is exact enough, yet can be derived from the random behavior of billions of small particles. Do social scientists fool themselves when they appeal to this analogy, thus trying to reduce the prediction interval to a single point and to evade the inconveniences of probabilistic thinking?

As an example, let me refer to the economists' attempt at explaining the aggregate investment, i.e., the total money outlay on plant and equipment made in a year by all the firms of the country.

This outlay is, of course, the sum of the outlays of individual firms on their plant and equipment. One might try to explain the decision of a manager or a management board of an individual firm by those variables that are likely to influence the level of anticipated profits, thus making the expansion of the firm's activities in a given year a promising or a bad proposition. Thus, one would think of explanatory variables like the following: the profits that are already being made at the time of decision; the current demand for the firm's product; the interest rate at which money can be borrowed to expand the firm's plant and machinery; the cost of new machinery and building. Clearly, many other variables will be needed to describe the particular circumstances in which a given firm finds itself, and which will affect the manager's anticipations of the future of his particular market, and his decision to build a new wing of the plant or, on the contrary, to sell parts of his machinery for scrap without replacing them by new ones. If the particular circumstances do not vary too strongly from one firm to another, one might be able to explain the annual change in the aggregate plant and equipment of the American industry (so called rate of aggregate investment) as a function of those general variables: the total profits of the past year; the total demand for all products (represented by total national income); the average interest rate, the average cost of machinery and building, etc. To
do this, one would be able to utilize the published historical statistics, which gives for each year and even each quarter or month, the aggregate outlays on plant and equipment, the total profits, the interest rate, prices, etc. In this hope, various investigators have looked for a statistical relationship between the aggregate annual investment of the American industry and the several aggregate variables that we have mentioned. In particular, attempts have been made to construct prediction intervals like (2.4), where \( \mu \) and \( \sigma \) are the mean and standard deviation of the estimate of annual aggregate investment; while the \( z_i \) are the various mentioned explanatory variables, also of aggregative nature: total profits of the preceding year, total national income, average interest rate, etc. However, the standard error \( \sigma \) of the estimate, and hence the prediction interval, proved to be very large: or, what is the same thing, the "unexplained residual," the "disturbance" (2.3) — with \( x \) interpreted as the annual rate of aggregate investment — has a very large standard deviation. The meaning of the prediction interval being "very large" is simply this: it is so large as to be useless for predictions.

For the social scientists, an interesting possible explanation of the strong random variability of the unexplained residual of some aggregative variables, is the phenomenon variously called "leadership," "imitation," "contagion," or "Zeitgeist" or "fashion." The random disturbances in the response (e.g., in the decision to expand plant) of any pair of individual members of a human group to a given set of circumstances may be positively correlated, both disturbances being partly determined by unspecified common cause. If this is the case, the response of the corresponding aggregative variable to a given set of circumstances will be subject to more unpredictable random fluctuations than if that correlation were absent.

A mathematical illustration will help. Suppose \( p \) variables \( v_1, \ldots, v_p \) have a joint normal distribution with a common expectation \( \mathbb{E} v_j = 0 \) and a common variance \( \mathbb{E} v_j^2 = \sigma^2 \). Let them also have a common correlation coefficient for any pair \( j \neq k \), \( \mathbb{E} v_j v_k / \sigma^2 = \rho > 0 \). Thus

\[
\mathbb{E} v_j v_k = \begin{cases} 
\sigma^2 & \text{when } j = k \\
\sigma^2 \rho & \text{otherwise.}
\end{cases}
\]

What is the distribution of \( \bar{v} \), the average of these variables? \( \bar{v} \) has normal distribution. Its expectation = 0. And its variance
$$E(\bar{v})^2 = E\left(\frac{v_1 + \ldots + v_p}{p}\right)^2 = \frac{1}{p^2} E \sum_j \sum_k v_j v_k$$

$$= \frac{1}{p^2} \left[p\sigma^2 + p(p-1)\sigma^2 \rho\right] = \sigma^2 \left[\frac{1 - \rho}{p}\right] + \rho .$$

$$E(\bar{v})^2 \longrightarrow \sigma^2 \rho \quad \text{as} \quad p \longrightarrow \infty . \quad (2.5)$$

Thus, under our assumptions, the variance of the average tends to zero if and only if the correlation $\rho = 0$. Applying this to our problem of averaging the individual responses (e.g., the investment decisions of a number of firms), we can interpret $v_j$ as follows:

$$v_j = x_j - \sum_{i=1}^n \alpha_i \ z_i^{(j)} - \mu \quad (j = 1, \ldots , p) \quad (2.6)$$

where the $\alpha_i$ and $\mu$ have the same meaning as in (2.3), (2.4), and the subscript $j$ indicates the particular firm; its response (investment) $x_j$ is explained by the variables $z_1^{(j)}, \ldots , z_n^{(j)}$, except for a random disturbance $\mu + v_j$, which has mean $\mu$ and variance $\sigma^2$. The variable $x_j$ has then also variance $\sigma^2$, the same for all firms. If, for all pairs of firms, the random disturbances are not correlated ($\rho = 0$) then indeed the average of the responses of all firms tends to have zero variance as the number of firms increases. But this is not so if there is, pairwise, a positive correlation between the individual disturbances. In the extreme case, when all firms act in unison ($\rho = 1$), the average has the same variance as any individual response: an obvious result. This mathematical illustration might be made more general – for example we might admit positive correlation coefficients of different sizes for different pairs of firms. But our result suffices, to show how "fashion" endangers predictability – a warning against the abuse of a misunderstood "law of averages." If the decision-makers, however numerous, follow a few leaders, then the personal circumstances, the ulcers or divorce proceedings of the leaders become important factors in predicting the average or aggregate of the decisions.
IDENTIFICATION

Suppose two random characteristics $X_1$ and $X_2$ are being measured on each person of a sample. For example, $X_1$ may be a person's consumption of cigarettes, and $X_2$ his consumption of cereals; or $X_1$ may be the person's answer to the question "Do you prefer blondes?" and $X_2$ his answer to the question "Do you feel comfortable with strangers?" Suppose that a parameter $Z$ is known to exist which varies from person to person but is fixed for each person; and which influences $X_1$ and $X_2$ in a "probabilistic" fashion. That is, the probability that a person's $X_1$ and $X_2$ will take a given pair of values (or fall within given intervals) depends on the value of the person's $Z$. Thus the joint distribution of the numbers of cigarettes and bowls of cereals consumed by a person during a week may depend on the person's income, $Z$. In the case of the two questions - about blondes and about strangers - $Z$ may stand for some characteristic, such as "ethnocentricity", of which the answers to the two questions are symptoms. We shall denote the joint probability distributions of $X_1$, $X_2$, conditional upon $Z$, by

$$\pi(X_1, X_2 | Z), \quad (2.7)$$

a function of three variables. It is quite analogous to (2.2), with the vector $(z^{(1)}, \ldots, z^{(n)})$ replaced by a single $Z$, and the number $x$ replaced by the vector $(X_1, X_2)$. [In a more general case, both the random variable to the left and the parameter to the right of the bar in (2.7) will be vectors.]

Suppose one has succeeded in estimating the function $\pi$ from a sample. This knowledge could then be used in two opposite ways, to answer questions about a person not in the sample:

Question 1: given this person's $Z$, predict his $X$'s.

Question 2: given his $X$'s, make statements about his $Z$.

Why either of these applications of our knowledge of the conditional probability function $\pi$ should ever arise in practice, will belong into our third lecture (on policies). For the time being, we shall regard both questions as legitimate expressions of scientific curiosity.

If the sampled persons had lent themselves to the measurement, not only of their $X$, but also of their $Z$, the problem of
estimating the function $\pi$ from the sample, and of subsequently answering Questions 1 or 2 about a person not in the sample, would be a familiar one. For example, the main use of computing a regression equation from a sample is to answer Question 1, in the manner of (2.4); using a sample to answer Question 2 is also a problem well known to statisticians, a discrimination problem.\textsuperscript{15} However, a new twist is introduced by making $Z$ a non-observable (at least not observable on the sampled individuals), a "latent" characteristic. It is natural to expect that this loss of information must be made up by some additional information if the problem is to remain solvable.

The following simple hypothesis (which is itself a bit of "additional information") will be used for the case of "cigarettes and cereals": $X_1$ and $X_2$ are proportional to $Z$, subject to random percentage deviations, denoted by $U_1$ and $U_2$, respectively. That is, there exist two constants, $\Lambda_1$ and $\Lambda_2$ such that

$$X_i = \Lambda_i \cdot Z (1 + U_i), \ i = 1, 2.$$  

Remember that $X_1, X_2$ have been measured on persons of the sample, but $Z$ has not. (Collectors of family budget data are, in fact, able to ascertain single consumption items but have difficulty in finding the incomes!) It will be convenient to use logarithms. Write $\log Z = z$, $\log X_i = x_i$, $\log \Lambda_i = \lambda_i$, $\log (1 + U_i) = u_i$, $i = 1, 2$. Then

$$x_i = \lambda_i + z + u_i, \ i = 1, 2. \quad (2.8)$$

We shall further assume that random deviations cancel out in a very large sample, in the sense that the expectations of $u_1, u_2$ vanish:

$$Eu_i = 0 \ (i = 1, 2); \quad (2.9)$$

assume also that $u_1, u_2$ are jointly normally distributed. We shall write $Eu_i u_j = \sigma_{ij} (1, j = 1, 2)$. Our hypothesis involves thus 5 unknown parameters $\lambda_1, \lambda_2, \sigma_{11}, \sigma_{22}, \sigma_{12}$. Can we determine them and use them to make a probabilistic prediction of $x_1$ and $x_2$ for any given $z$?\textsuperscript{16}

The answer depends on what information is in our possession before we collect the data. The main purpose of the present
example is to show how the character of this "a priori information" makes the problem determinate or indeterminate. With insufficient a priori information, the problem is indeterminate, however large the sample.

Assume, for example — in addition to the assumptions already made [(2.8), (2.9) and normality of the distribution of \( u_1, u_2 \)] — that we know the frequency distribution of \( z \)\(^{17} \): viz., normal with known mean and variance \((\zeta \text{ and } \omega, \text{ say})\), and independent of the disturbances \( u_1 \text{ and } u_2 \); hence, \( E u_1 = 0 = E u_2 \). As before, we shall neglect sampling errors (by assuming our sample very large), so that the sample averages \( \bar{x}_1, \bar{x}_2 \) are almost equal to the corresponding expectations; then by (2.8), (2.9), approximately

\[
\bar{x}_i = E x_i = \lambda_i + \zeta;
\]

hence

\[
\lambda_i = \bar{x}_i - \zeta \quad (i = 1, 2),
\]

thus determining the two unknowns \( \lambda_1, \lambda_2 \). Moreover, the sample moments (sample averages of squares and products) computed from the observed values of \( x_i \) in the (very large) sample, serve to determine \( \sigma_{11}, \sigma_{22}, \sigma_{12} \), the parameters of the distribution of \( u_1, u_2 \). For, by (2.8), (2.9),

\[
x_i - \bar{x}_i = z - \zeta + u_i;
\]

therefore

\[
E (x_i - \bar{x}_i) (x_k - \bar{x}_k) = E (z - \zeta)^2 + E u_i u_k + (z - \zeta) (E u_i + E u_k)
\]

\[
= \omega + \sigma_{ik}; \quad i, k = 1, 2.
\]

Hence, in a very sample, approximately

\[
\begin{align*}
\text{sample average of } (x_1 - \bar{x}_1)^2 &= \omega + \sigma_{11} \\
\text{sample average of } (x_2 - \bar{x}_2)^2 &= \omega + \sigma_{22} \\
\text{sample average of } (x_1 - \bar{x}_1) (x_2 - \bar{x}_2) &= \omega + \sigma_{12},
\end{align*}
\]

where \( \omega \) and the sample moments are known. The parameters
\(\sigma_{11}, \sigma_{22}, \sigma_{12}\) thus estimated serve to determine (using the tables for bivariate normal distribution), the probability with which \(u_1\) will fall into some given interval, and \(u_2\) will fall, at the same time, into some given interval.

This would answer our Question 1. As to Question 2, I shall not go into details here. Suffice it to say that, once we know the conditional distribution of \(x_1, x_2\) given \(z\), we can also estimate the parameter \(z\) characterizing an individual, from values \(x_1, x_2\) observed (once or in repeated observations) on an individual.

But suppose our a priori knowledge is less complete than in the example studied so far. For example, suppose we know the mean \((\zeta)\) of \(z\) (i.e., the geometric mean of average incomes of the U.S. population) but do not know its variance \(\omega\). We see from (2.10), (2.11) that we can then determine \(\lambda_1, \lambda_2\) but not \(\sigma_{11}, \sigma_{22}, \sigma_{12}\). Consequently we can predict \(x_i\) for a given \(z\) only in the sense of providing the conditional expectation of \(x_i (= \lambda_i + z)\) but not of stating an interval into which the \(x_i\) should fall with preassigned probability. Nor shall we be able to estimate an individual's \(z\) from an observation on his \(x_1, x_2\).

Or suppose we have the knowledge of both \(\mu\) and \(\omega\), but have also a larger list of unknowns, because we are convinced that the assumption (2.8) is too special and should be replaced by a more general one:

\[
x_i = \lambda_i + \beta_i \cdot z + u_i^{10} \quad i = 1, 2,
\]

(2.12)

with \(\beta_1, \beta_2\) unknown. Using this equation, we can again express — similarly to (2.10), (2.11) — the two sample means of \(x_1\) and \(x_2\) and their three sample moments in terms of the \(7\) unknowns. We shall be \(2\) equations short. The problem is indeterminate.

Consider, on the other hand, the case when we have fewer unknown constants than we have equations to determine them. Suppose again that we know both \(\zeta\) and \(\omega\) and let our hypothesis again be expressed by (2.8), (2.9). But, suppose, in addition, we know from other sources that \(\sigma_{12} = 0\), i.e., there can be, at any fixed income, no correlation between eating of cereals and smoking. We have then only four constants to determine \((\lambda_1, \lambda_2, \sigma_{11}, \sigma_{22})\), yet can derive from the observations the same five equations (2.10), (2.11) as in our first example, in which \(\sigma_{12}\) was unknown. If the hypothesis now used is correct, i.e., if (2.8), (2.9), as well as the assumed frequency distribution of \(z\) and the assertion \(\sigma_{12} = 0\), are all valid, then one of those equations derived from observations
is redundant for the purpose of determining unknown parameters. But then it can be used to test the hypothesis. Indeed, the assertion \( \sigma_{12} = 0 \) makes the last of the equations (2.11) into a relation between two known quantities, viz., between the observed cross-moment of \( x_1 \) and \( x_2 \) and the known variance \( \omega \) of \( z \). If this relation is not, in fact, satisfied, our hypothesis is wrong: e.g., (2.8) may be false, or \( z \) or \( u_1 \), \( u_2 \) are not distributed normally, or \( \omega \neq \text{variance of } z \), or \( \sigma_{12} \neq 0 \), etc.\(^\text{20}\)

To sum up our examples: We started with a case in which all unknown constants could be determined from an equal number of (independent) equations obtained from observations — all constants were "identifiable." By modifying the assumptions used or the kind of data available we obtained a case in which some of the constants, and another in which all of the constants were (however large the number of observations!) not determinable, "non-identifiable." Finally, we have had an example of "over-identification": observations yielded more (independent) equations than there are unknown constants; if these equations are inconsistent,\(^\text{21}\) the hypothesis used must be rejected.

The term identifiability was suggested by T. C. Koopmans and the relevant mathematics were studied by the staff of the Cowles Commission for Research in Economics in considerable detail (see, e.g., the Commission's Monographs No. 10 and 14) because of the importance of the identification problem in economics.\(^\text{22}\) However, the problem seems to be present in other social sciences as well, e.g., in the Lazarsfeld theory of latent structures and in Thurstone's factor analysis.) In fact, my examples were chosen with the very purpose of providing a link between problems encountered in economics and those more familiar to a sociologist or psychologist, but, alas, less familiar to me. Let me, then, venture again to interpret \( x_1 \) and \( x_2 \) as answers to two questions asked of a person possessing an unknown degree \( z \) of some measurable characteristic. \( Z \) is a parameter, changing from person to person. It is not a random variable.\(^\text{23}\) The variables \( x_1 \) and \( x_2 \) can each take values 1 (for "yes") or 0 (for "no"). Assuming that a very large number of persons have been questioned, we want to determine the conditional joint probability

\[ \pi = \pi(x_1, x_2 | z). \quad (2.13) \]

This is the same as (2.2), with \( n = 1 \) and with \( x \) interpreted as a vector consisting of two components. Since \( x_1, x_2 \) can have only
two values each, it is inappropriate to state an assumption about the distribution of a random "unexplained residual," a continuous quantity such as (2.3), or the \( u_i \) in (2.8). Instead, we specify directly the function (2.13) of \( z \), i.e., assume that the respective probabilities of the four possible alternatives [1] \( x_1 = x_2 = 0 \); 2) \( x_1 = x_2 = 1 \); 3) \( x_1 = 0, x_2 = 1 \); 4) \( x_1 = 1, x_2 = 0 \) are determined by \( z \) in a specified fashion (Lazarfeld's "trace lines"). Suppose, for example that the function \( \pi \) can be tabulated as follows (each cell corresponding to the probability of one combination of answers)

<table>
<thead>
<tr>
<th>( x_1 )</th>
<th>( x_2 = 0 )</th>
<th>( \alpha_{00} + \beta_{00} z )</th>
<th>( \alpha_{01} + \beta_{01} z )</th>
<th>( \alpha_{10} + \beta_{10} z )</th>
<th>( \alpha_{11} + \beta_{11} z )</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>( \alpha_{00} + \beta_{00} z )</td>
<td>( \alpha_{01} + \beta_{01} z )</td>
<td>( \alpha_{10} + \beta_{10} z )</td>
<td>( \alpha_{11} + \beta_{11} z )</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>( \alpha_{00} + \beta_{00} z )</td>
<td>( \alpha_{01} + \beta_{01} z )</td>
<td>( \alpha_{10} + \beta_{10} z )</td>
<td>( \alpha_{11} + \beta_{11} z )</td>
<td></td>
</tr>
</tbody>
</table>

(This presupposes of course that \( z \) has an upper and a lower limit). The \( \alpha \)'s and \( \beta \)'s are unknown. Since the four probabilities must add up to 1 for any \( z \), the \( \alpha \)'s must add to 1, and the \( \beta \)'s to 0. We have thus 6 (not 8) unknown constants. Can we determine them from the sample? Call \( p_{ij} \) the proportion of people, in a very large sample, whose pair of answers is: \( x_i = i, x_j = j \). Denote by \( \zeta \) the mean of \( z \) in the sample. Then

\[
p_{ij} = \alpha_{ij} + \beta_{ij} \zeta \quad i, j = 0, 1.
\]

\( \zeta \) can be chosen as a unit of measurement of the \( z \). The four proportions \( p_{ij} \) are all observable, but since they must add up to 1, only three of the four equations (2.15) are independent. This is not enough to determine the 6 unknowns. They are not identifiable.

But suppose we know with certainty the following property of our questionnaire: for any given person, the probability of answering the first question by yes is independent of whether he answers the second question by a yes or a no (though, of course, both probabilities vary from person to person, depending on \( z \)). Obviously, not every questionnaire satisfies this condition. This is a case of a "pure test," in Lazarsfeld's terminology. The probability that a person answers with "yes" the question "Is your name John?" depends on how the same person answers the question "Are you a
woman?" If the probability of a person's being a man is 0.5, and the probability of a person's being called John is 0.1 (the average between the probability 0.2 of a man's being called John and the probability 0.0 of a woman's being called John), then the probability of a person's both being a man and being called John is not $0.5 \times 0.1$ (as it would be if these two events were independent) but $0.5 \times 0.2$.

Or consider two questions that are almost identical in content. Suppose the probability for the first question to be answered affirmatively by any person with $z$ degrees of "ethnocentricty" is $q$; the corresponding probability for the second question will then be near $q$; and the probability that a person with $z$ degrees of ethnocentricty will answer both questions in the affirmative will be, not near $q^2$ but near $q$.

It may also be that the degree of dependence between the answers depends itself on $z$. An extreme ethnocentric may strongly associate dark hair with fast talk (and perhaps dislike both) while a man on the other end of the ethnocentricty scale will not have this ready image in his mind: thus a test may be "pure" for some persons, not "pure" for others.

Suppose, however, the investigator is sure to have formulated his questions so that, for a fixed $z$, the answers are completely independent; he is sure that all traces of possible associations, logical or otherwise, between the questions, have been weeded out. This a priori knowledge permits him to specify the function $\pi$ more narrowly. For example, in the case (2.14) the assumption of independence will be expressed by the condition

$$\frac{\alpha_{00} + \beta_{00} z}{\alpha_{10} + \beta_{10} z} = \frac{\alpha_{01} + \beta_{01} z}{\alpha_{11} + \beta_{11} z}. \quad (2.16)$$

Multiplying out and transferring all terms to one side, we obtain a quadratic expression in $z$ which must vanish for any $z$. Hence each of its three terms vanishes, thus providing us with the three missing equations. Our 6 constants have now become identifiable.

In this example, the assumption of independence between answers (in the sense defined) has proved powerful enough to make an otherwise unidentifiable set of unknowns identifiable. The result can be used to estimate the latent characteristic $z$ of a person outside the sample (the discrimination problem mentioned above): e.g., if his answer is "no" to both questions, the maximum
likelihood estimate of his $z$ is obtained by maximizing with respect to $z$ the probability $\pi (0,0|z)$. We therefore put $\alpha_{00} + \beta_{00} z = 1$. One can also estimate an appropriate confidence interval for $z$. If all the unknown constants $\alpha_{00}, \beta_{00}$, etc. had been unidentifiable it would have been impossible to estimate $z$. We can say that non-identifiability would make the confidence interval for $z$ infinite. It can be conjectured that if $z$ is identifiable both with and without the assumption of independence, the appropriate confidence interval for $z$ is shorter when this assumption is made than when it is not made.

Since the assumption of independence between questions can have such powerful implications, the investigator, in using or designing the questionnaire, will have to be very critical. As in many other cases, the soundness of statistical results will much depend on the soundness of pre-statistical assumptions which prove inaccessible to statistical test.
Section III. PROBABILITY AND POLICY

In the first lecture we discussed the following rule of conduct: choose that decision which makes the "moral expectation" (= the mathematical expectation of utility) as large as possible. This presupposes consistency in the decision-maker's system of preferences. A rational policy maker must "know what he wants." In addition (at least in the approaches of Bayes and Ramsey), he must "know what he believes" - his system of subjective probabilities, too, must be consistent. To require these consistencies is on the same plane as to require that the policy maker do not make errors of logic and arithmetic: that he avoid contradictions. Logic is self-control, and this includes logic of decision. In Ramsey's words, one should be able to "stop to think it out," rather than to act on the temporarily uppermost desire and belief. This is an ideal, a skill which good decision-makers possess in a higher degree than poor ones. A bad decision-maker tosses in bed and decides in a half-dream. A good decision-maker dresses up a payoff matrix (though not necessarily with pen on paper). I suspect this skill is not entirely inborn and can be acquired by training: a fruitful field for applied psychology.

We also found in our first lecture that the rule of maximizing moral expectation leads to the proposition that degrees of belief (which we needed to know in order to compute moral expectations) are, in the limit, equal to relative frequencies obtained from samples. Accordingly, we devoted our second lecture to showing how certain probability distributions that interest a social scientist are obtained from observations. The probability distributions that interest the social scientist must be usable for prediction of the behavior of people. In the concluding part of the second lecture, we found that certain probability distributions needed for such prediction may be inaccessible to estimation, no matter how large the sample. Certain parameters are "non-identifiable." Yet these parameters are often just the ones the social scientist is particularly curious about: he feels they constitute his "theory," as distinct from "mere description." He is interested in "structures," "latent parameters," and not in mere "empirical regularities" and must be disappointed whenever he finds that the goal cannot be attained with the available kind of data.

What are those "latent" properties, those "theories" behind the manifest data? Why are they interesting? And what makes them so
elusive? My tentative answer will be: that "theories" are sought for the sake of decision-making; and that the difficulty in getting at them is due to the difficulty of performing experiments.

To take up one of the examples of our second lecture: Why is a market research organization not interested in merely estimating from its data the bivariate distribution of the consumption of cigarettes and the consumption of cereals? That would permit them to predict the one from the other. But this is not interesting. Instead, there is lurking a theory that both smoking and the expenditure on cereals depend on the third variable, income, which happens to be difficult to ascertain and is thus "latent" rather than "manifest." We have shown that under certain hypotheses it is possible, and under other hypotheses impossible, to obtain from data on cereals and cigarettes consumed by a sample of people, a relationship that will predict cereals consumption from income and that is presumably more "theoretical," more of a "structure," than the relationship between the consumption of cereals and that of cigarettes. The relationship involves two kinds of knowledge: the "pre-statistical" hypotheses just mentioned and the numerical parameters estimated from the sample. E.g., the hypothesis may be: cigarettes and cereals are consumed in proportion to incomes, apart from a random percentage deviation for each of the two items, these deviations being distributed normally, with zero correlation. The numerical parameters to be estimated may have been, in this case, the unknown (geometric) mean of the ratios of each of the two expenditure items to the individual's income, and the variances of each of the two deviations. We have seen that if we know the correlation (e.g., zero) between the two deviations, and know the income distribution (though not the individual incomes) in the population from which the sample was taken, the set of parameters is identifiable. But some or all of them become non-identifiable if our a priori knowledge is less complete — for example, if we do not have grounds to assume that the average relation between income and smoking or cereal-eating is one of proportionality, or that the two deviations from proportionality are non-correlated (i.e., that a man's smoking beyond his means is compatible with over-eating as well as with under-eating cereals, compared with the average breakfast of his income group). In these cases, knowledge of the "latent" relation between income and the consumption items becomes inaccessible, with the type of data we have assumed in our example.

Why hanker after this inaccessible knowledge? Why not be content — to continue with our example — with predicting cereal consumption from cigarette consumption or vice versa, which is perfectly possible on the basis of our data? Presumably because the
market research organization does not expect the latter kind of prediction to be of much use in future practical situations. It does find it useful to predict consumption from income. Why? Because it visualizes the following situation: given a new market, with customers' income distribution known and different from that of the population previously studied, predict the demand for cereals or cigarettes. If the decision to face new markets would never have to be taken, research about the influence of income (or other such factors) on consumption would be unnecessary: it would suffice to know past consumption.

Similarly, if in line with the last example of our previous lecture – the manifest data are answers to a questionnaire and the latent parameter deemed to underly those answers is the "degree of ethnocentricity," the reason why we are interested in measuring the latter is, presumably, its potential usefulness. It is not useful to predict that a "yes" answer to a certain question entails, with a certain probability, a "no" answer to another question (with population the same as the sampled one). I presume that what is really wanted is to predict the action of some new individual or group. It is assumed a priori that, for example, the discourteous treatment of immigrants by an official is determined by his degree of ethnocentricity which also influences his answers to questionnaires and which (if "identifiable") is revealed by those answers. One wants to use those answers of the aspirant to an office to predict the behavior of the future official. A convenient way to state the relation between the observable answers to the questionnaire and the virtually observable action of the official is via the "degree of ethnocentricity" which influences both. It is assumed to be "behind" those observables, just like the genetic make-up of an individual is "behind" observable data on his and his ancestors' hereditary features, even though genes might never become observable through the strongest ultra microscope. The study both of genotype and of "ethnocentricities" would not be called for if predictions of the effects of changed genotype and changed ethnocentricity (or their changed distribution within a population), as in problems of animal breeding or the administration of policies towards ethnical minorities, were never to be made.

Permit me to use an example from economics: a severely simplified variant of a "Keynesian" model designed to discuss fiscal policy as an instrument for maintaining employment and stable prices.

Consider the dollar value of all goods and services produced during a year (not counting those used for repair or renewal of the
existing stock. This dollar amount can be called the (net) national product. It is composed of the dollar amounts — called incomes — paid out, in the course of production, to workers, factory-owners, capital-lenders. Thus net national product is identical with national income. Denote this quantity by \( y \). Consider now another quantity: the dollar value of all goods and services demanded during a year (again not counting repair and renewal). This national demand is composed of three parts: 1) the demand of consumers for food, shelter, etc. — denote it by \( c \); 2) the demand of businessmen for machines and other goods to increase their plants and inventories (a matter I had occasion to mention in the previous lecture) — denote it by \( b \); 3) the demand of the government for the services of its employees and for public buildings, armament goods, etc. — we shall denote it (for reasons that will appear presently) by a Greek letter, \( \rho \) (for Roosevelt). In general the national demand \( c + b + \rho \) is not identical with the national product (income) \( y \). People who decide about production (and hence about the incomes disbursed) may or may not be able to quickly adjust production to demand. However in our context we need not discuss this adjustment process, its motives and form, except in a footnote later.\(^{25}\) We shall assume that production is adjusted to demand instantaneously, so that always

\[
y = c + b + \rho. \tag{3.1}
\]

Of these quantities, only \( \rho \) is directly controlled by the policymaker, the government. What determines the rest, namely \( b \) and \( c \)?

As to \( b \), let us assume that businessmen feel encouraged to expand their plant when the current national income, \( y \), is high; at least they feel so "on the whole," i.e., apart from some random deviation. To fix the ideas, let us use a linear approximation:

\[
b = \beta y + v_b, \quad \beta > 0; \tag{3.2}
\]

where \( \beta \) is a parameter characterizing businessmen's behavior and \( v_b \) is a normally distributed random variable; denote its mean and variance by the Greek letters \( \mu_b \) and \( \sigma_{vb} \) respectively.

As to \( c \), let us assume that people's consumption depends, apart from a random deviation \( v_c \), only on the amount of income that remains in their hands after payment of taxes, and on nothing else. Using again a linear approximation,
\[ c = \gamma (y - \tau) + v_c, \quad \gamma > 0; \] (3.3)

where \(\gamma\) is a parameter descriptive of consumers' behavior and \(\tau\) denotes the tax collected. The mean and variance of \(v_c\) (assumed normal) will be denoted by \(\mu_c\) and \(\sigma_{cc}\), respectively. Since the random deviations in the behavior of consumers may or may not be correlated with those in the behavior of businessmen we need also a symbol \(\sigma_{bc}\) to denote their covariance.

Let us now marshall all our symbols:

- government-controlled parameters: \(\rho, \tau\);
- non-controlled parameters: \(\beta, \gamma, \mu_b, \mu_c, \sigma_{bb}, \sigma_{cc}, \sigma_{bc}\);
- other variables: \(b, c, x\).

Finally, we assume for simplicity that the government determines directly the tax collected, \(\tau\) (and not, as is actually the case, the tax rates only).

Let us now marshall all our quantities:

- government-controlled parameters: \(\rho, \tau\);
- non-controlled parameters: \(\beta, \gamma, \mu_b, \mu_c, \sigma_{bb}, \sigma_{cc}, \sigma_{bc}\);
- dependent variables: \(b, c, y\);
- random deviations: \(v_b, v_c\).

We have three dependent variables and a system of three equations. We can solve for \(b, c, y\). For example, equation (3.4) below gives the solution for \(y\). Since the system involves random deviations \(v_b, v_c\), the dependent variables \(b, c, y\) are also random: they fluctuate with \(v_b, v_c\). If the deviations \(v_b, v_c\) were observable, one could determine, for every pair \(v_b, v_c\), the values that the dependent variables \(b, c, y\) would take, provided one knows the parameters \(\rho, \tau, \beta, \gamma\). Actually \(v_b, v_c\) are not observable but their probability distribution is fully determined by the parameters \(\mu_b, \mu_c, \sigma_{bb}, \sigma_{cc}, \sigma_{bc}\). If one knows these parameters, one can find the probability distribution of \(b, c, y\) for any given set of values of the controlled parameters \(\rho, \tau\) and of \(\beta, \gamma\).

We are not concerned here with whether the model is economically sound. It is too crude. We merely use it as an example for our methodological discussion. The model's three equations, together with the probability distribution of the random deviations.
$v_b, v_c$, purport to "explain" the observed distribution of the three dependent variables $b, c, y$. The model expresses a hypothesis about the behavior of consumers and businessmen, summarized by the non-controlled parameters — the Greek letters excluding $\rho$ and $\tau$. The policy problem is for the government to choose the best values of the controlled parameters $\rho, \tau$, — the values that will maximize the government's "expected utility," a term defined in our first lecture. If the government were concerned simply with the national income $y$ (a dollar amount, as you remember) it would choose $\rho$ and $\tau$ so as to maximize the expected value (the mean) of $y$ — call it $\mu_y$, one of the parameters of the distribution of $y$. How is the distribution of $y$ determined by $\rho$ and $\tau$? The random variable $y$ is related to $\rho, \tau$ in a linear fashion, as is seen by solving our three equations:

$$y = \lambda (\rho - \gamma \tau) + \lambda (v_b + v_c), \tag{3.4}$$

where $\lambda = 1/1 - \beta - \gamma$. Since $v_b + v_c$ is distributed normally, so is $y$. The mean of $y$ is

$$\mu_y = \lambda (\rho - \gamma \tau) + \lambda (\mu_b + \mu_c); \tag{3.5}$$

this is how $\mu_y$ is affected by the government's choice of public expenditure $\rho$ and tax revenue $\tau$.

Note in passing that the policy-maker might be concerned, not only with a high mean national income but also with a high predictability of national income, as expressed by its variance, $\sigma_{yy}$. As the late Senator Arthur Vandenberg said in criticizing the New Deal: "We don't want to live on a flying trapeze." It is worthwhile to remark, as an exercise in the logic of these matters, that in our particular model the policy-maker is unable to affect $\sigma_{yy}$. The variance of $y$ is not affected by adding to $y$ a constant; it is, by (3.4), simply proportional to the variance of $(v_b + v_c)$ and is equal to

$$\sigma_{yy} = \lambda^2 (\sigma_{bb} + \sigma_{cc} + 2\sigma_{bc}). \tag{3.6}$$

Thus, the knowledge of the last three of the non-controlled parameters of our list above, is irrelevant for policy purposes if the policy goal depends on $\mu_y$ and/or $\sigma_{yy}$. If our model were non-linear, the result might be different, and the knowledge of $\sigma_{bb}, \sigma_{cc}$,
probability important to the policy-maker concerned with high and/or reasonably predictable national income. In any case, this knowledge is necessary if one wants to "predict", in the sense of our Lecture II, the random variable, income: i.e., if one wants to estimate the distribution of y. This is, in our example, described by $\mu_y$ and $\sigma_y$, and to know $\sigma_y$ one has to know $\sigma_{bb}$, $\sigma_{cc}$, $\sigma_{bc}$. But such prediction may be of no concern to a practical policy-maker.

Continuing with a government bent on maximizing the expected value $\mu_y$ of national money income, return to our equation (3.5). We shall assume $\lambda > 0$, that is $\beta + \gamma < 1$. This assumption is not based on systematic statistical studies but on a general estimate of plausible behavior of consumers and businessmen and also on the observed "stability" of a system, a consideration which we cannot discuss here in any detail. If, then, $\lambda = \beta + \gamma - 1 < 0$, the expected money incomes $\mu_y$ as determined in (3.5) is increased by raising $\rho$ and cutting down $\tau$. If there is an upper limit on government expenditure - say, $\rho_{\text{max}}$ - and a lower limit on tax revenue - say, $\tau_{\text{min}}$ - then money income is highest with $\rho = \rho_{\text{max}}$ and $\tau = \tau_{\text{min}}$ (possibly zero). Now knowledge about the behavior parameters $\beta, \gamma$ is necessary to find these optimal values of the government-controlled parameters, except the validity of the assumption $\beta + \gamma - 1 < 0$.

This result is due to the fact that, under our assumptions, - as summarized in (3.5) - the expected money income $\mu_y$ changes monotonically (has no turning points) in response to changes in $\rho$ and $\tau$. Therefore $y$ can achieve a maximum only at some boundary values of $\rho$ and $\tau$. If we introduced, for realism's sake, an upper limit on the deficit $\rho - \tau$, the same general result would remain true: no knowledge of the non-controlled parameters would be necessary to determine the policy $(\rho, \tau)$ that maximizes expected income $(\mu_y)$.

Let us now change our example so as to make it more up-to-date. To think of a government that tries to maximize the expected national income measured in dollars (the money income $y$) was possible in times of depression when prices were relatively stable, and a rise in money income was about equivalent to a rise in physical production and employment. In present inflationary times one has to think not only of the money income $y$ but also of the price level, call it $p$. The physical production is $Y = y/p$; and the government is concerned with both $Y$ and $p$. This is how we modify the previously assumed utility function of the policy maker. We have also to modify the model (3.1), (3.2), (3.3), by adding statements about what determines the dependent variables we have just added to our list, viz., $Y$ and $p$. 
One simple hypothesis is to assume that physical production \( Y \) cannot rise above a maximum, \( \eta \) (the "full employment output"); and that the price level \( p \) is constant (\( p = \pi \), say), as long as this maximum output is not reached. That is, our two additional dependent variables, \( Y \) and \( p \), are determined as follows, in terms of our old dependent variable \( y \) and the parameters \( \pi \) and \( \eta \):

\[
\begin{align*}
Y &= y/\pi, \quad p = \pi \quad \text{when } y \leq \eta \\
Y &= \eta, \quad p = y/\eta \quad \text{when } y > \eta \pi.
\end{align*}
\] (3.9)

Roughly, the first line corresponds to an unemployment situation and the second line to a full employment situation. In the former, money income \( y \) changes because of changes in physical output \( Y \); in the latter, \( y \) changes because of changes in price level \( p \). If we retain, in addition, our old equations (3.1), (3.2), (3.3) and therefore also their implication (3.4) we see from (3.9) that the new dependent variables \( Y \) and \( p \) are random variables whose distribution depends on the same parameters as does the distribution of \( y \), and, in addition, on the parameters \( \eta \) (maximum output) and \( \pi \) (constant price level during depression).

We can assume that the policy maker regards high physical output as desirable, and a strong rise in prices over their depression level as undesirable — e.g., because of the injustice that such a rise would inflict on certain people. Thus his utility function — \( Y(Y, p) \), say — is increasing in \( Y \) and decreasing in \( p \), with \( p \geq \pi \). As a simple example, we may have, as his utility

\[
u = Y(Y, p) = Y - Ap, \quad A > 0, \quad p \geq \pi.\] (3.10)

The reader will notice that we use italicized capitals to indicate a new class of properties — not the parameters (controlled or non-controlled) of the model but the characteristics of the decision-maker's valuations, his "tastes." In particular, \( A \) is the number of units of real income (e.g., billions of dollars) with the purchasing power of the year (1940) that he thinks it worthwhile to sacrifice in order to avoid the rise of price level by one point. In assigning any pair of values to the controlled parameters \( \rho \) and \( \tau \), he will affect the distribution of \( Y \) and \( p \), and hence the expected value of \( u \), that is, his "moral expectation," in the sense of our Lecture 1. Our assumptions (3.9), (3.10) will help as an
Illustration. Under these assumptions, the utility $u$ is the following function of money income $y$:

$$
u = \begin{cases} 
\frac{y}{\pi} - A\pi & \text{when } y \leq \eta\pi \\
\eta - \frac{Ay}{\eta} & \text{when } y \geq \eta\pi.
\end{cases}$$  \hspace{1cm} (3.11)

This function is represented on Figure III.1 by a broken line consisting of two straight line segments: utility $u$ rises with money income till the latter reaches $\eta\pi$; then $u$ falls. Each bell-shaped curve (of which only two are drawn) indicates one of the possible normal distribution density functions of $y$, with a fixed variance [which, by (3.6), does not depend on policies] and a varying mean [which, by (3.5) does depend on $\rho$ and $\tau$]. (The two drawn curves have respective means $\mu_y$ and $\mu'_y$.) Thus the policy $(\rho, \tau)$ determines the position of the bell-shaped curve, i.e., gives the probability that money income $y$ will fall into any given (small) interval. If this probability is multiplied by the utility that corresponds to

Figure III.1
the value of \( y \) in this (small) interval and that is determined by (3.11), and if all such products are added, one obtains the expected (mean) utility that corresponds to a given policy \((\rho, \tau)\). Using our diagram, multiply in your mind each ordinate of the bell-shaped curve by each corresponding ordinate of the broken line, and add all such products. Their sum is the expected utility. Clearly, as the bell-shaped curve moves too far to the right or to the left, the higher utilities (i.e., the ordinates of the broken line taken in the proximity of the full-employment point, \( \eta \pi \)) are multiplied with the lower probabilities (i.e., with the ordinates at the left or right "tail" of the bell-shaped curve). Hence there exists some intermediate position of the bell-shaped curve — and hence some value of \( \mu_y \) — that results in the highest expected utility.

To this optimal value of \( \mu_y \) will correspond, by (3.5), an optimal value of \( \rho - \gamma \tau \), given the non-controlled parameters.

Hence the knowledge of those parameters \((\beta, \gamma, \mu_b, \mu_c)\) is needed to find those pairs of value of the policy-parameters \( \rho \) and \( \tau \), that will maximize expected utility, given the "taste-parameter" \( A \) of the policy-maker. Of course, the knowledge of the feasible limits such as \( \tau_{\text{min}}, \rho_{\text{max}} \) will also have to be used, to sift out the non-feasible combinations of taxes and expenditures. But, unlike in our previous example, this knowledge will not suffice. It becomes necessary to estimate the parameters that characterize the behavior of consumers and businessmen.

We see thus that the need for the knowledge of certain parameters, and therefore for their identifiability (in the sense of Lecture 2), depends, in general, on the whole model in which these parameters occur; and on the utility function of the decision-maker.

Does not this land us in a rather crook pragmatism? What about theory for its own sake, a theory that is not used to give advice, to choose, to act? A pragmatist will say that theory provides us with solutions which are potentially useful for a large class of decisions. It is welcome because we cannot foresee which particular decisions we shall have to take. Our decisions may or may not be such as to leave certain properties of the system unchanged. Hence, the more we know about its properties the better. If we merely want to know how long it takes to boil an egg, the best is to boil one or two without going into the chemistry of protein molecules. The need for chemistry is due to our want to do other and new things!

The word "value" is more respectable than "utility," yet means essentially the same. If the satisfaction of scientific
curiosity is recognized by our culture as one of the major values, the search for facts as well as theories is justified even if no predictive uses can be seen for the results. Yet, an additional scrutiny in terms of other social values remains important. Especially in social science there has been a temptation to use statistical tools for laborious rediscovery of the trivial, or the recording of the useless. One should not brush away as "utilitarian" or "pragmatic" the reminder that we ought to help in the making of socially important decisions.27
4.

1. The lecture was given on 8 December 1950.

2. Such a weighted average of utilities has been called "mathematical expectation of utility" or simply "expected utility." The founders of the theory of probabilities when speculating about gamblers' choices, called this weighted average the "moral expectation" (= the mathematical expectation of "moral" as distinct from "physical," or "monetary" wealth). The term "military worth" (of a military situation or objective), encountered in military writings is presumably equivalent to "utility." It is likely that the often heard term, "calculated risk" of a military enterprise, if the term has any meaning at all, means the negative of "expected military worth," in the sense of an average of the utilities of the alternative outcomes of the enterprise, weighted with their respective probabilities.


6. Recent studies by Savage and by Herstein and Milnor (to be published) have simplified or generalized the earlier formulations of von Neumann and Morgenstern, and of J. Marschak.

8. David Hilbert, *Die Grundlagen der Geometry*, Leipzig. (This is more complete, for the problem that interests us here, than the English edition, Chicago 1910).

9. Note that here and in what follows we are taking a more general view of a "state of the world" than in our previous examples. A state of the world may now include more than one possibility; and each possibility can, in turn, be subdivided. Thus "W" is a state of the world, and its subdivision, the alternatives "W and V" and "W and V" are also states of the world.

10. However the arithmetical operations leading to the statement are independent of this interpretation of it. They are the same as those used by James Bernoulli (Daniel's kin) who possibly did not think of probabilities as subjective.

11. A rigorous analysis on these lines is contained in an unpublished study by Roy Radner on "Consistent Decision Functions."

12. In our example, this assumption can be used only as an approximation since normal distribution applies to a continuous (and therefore infinite-valued) variable whereas x(cigarettes, communism) is discrete. Such approximations are convenient. For example, in Lloyd Warner's studies, the variable x is social status measured on a 12-point scale on the basis of neighbors' opinions; the $z^{(1)}$, $z^{(2)} ...$ are, respectively, income, housing conditions (again a discrete variable) etc.; and "weights" $\alpha_i$ are estimated, as well as $\mu$ and $\sigma$, by the method of "multiple regression." With two-valued variables, this appraisal degenerates into familiar "tests of dual hypotheses" (chi-squared, analysis of variance).

13. The problem may be posed somewhat differently. Given the joint distribution of the random disturbances $v$ of the responses of individual firms (i.e., given $\mu$, $\sigma$ and $p$), what is the distribution of the disturbance of the sum (not of the average)? One may, for example, compare the constant ratio $\frac{\sigma}{\mu}$ (the "coefficient of variation" of $v$) and the ratio between the following two constants: the standard deviation of the sum of random disturbances and the expectation of this sum. This ratio converges to $\frac{\sigma}{\mu} \sqrt{p}$. (A more difficult problem is to estimate the coefficient of variation of the sum: it is the expectation of a random ratio, not the ratio of two constants).


15. See, for example, A. M. Mood, *Introduction to the Theory of Statistics*, pp. 299-301. New York 1950. To quote Mood's example, "Anthropologists...make measurements x on skulls, of known age z, then
estimate the age $z_0$ of a skull of unknown age with measurements $x'$.

16. The random variable $x_1 - z(= u + \lambda_1)$ corresponds thus to the expression (2.3), with $n = 1$ and all the $\alpha_i$ known. It is normally distributed (with mean $\mu = \lambda_1$), jointly with another random variable, $x_2 - z$.

17. For example, although we cannot ascertain the incomes of families who tell us their consumption of cigarettes and cereals, we may know from some other source the distribution of those incomes; e.g., if our sample is large and we have reasons to regard it as representing the whole U.S. population, then the income distribution is known from income tax statistics.

18. Remembering that $\lambda_1, x_1, z$ stand for the logarithms of $\Lambda_1, X_1, Z$, this means that the unknown proportionality factor $\Lambda_1$ in (2.7), can be estimated as the ratio of the geometric mean of $X_1$ (observed in a large sample) to the (known) geometric mean of $Z$ -- a result that is not unexpected, in view of the multiplicative character of the random deviation $(1 + U_1)$.

19. Implied that the consumption item $X_1$ is (apart from random disturbances) proportional, not to the individual's income $Z$, but to some unknown power of it, $Z^\alpha$.

20. If the sample is not very large, the test procedure is more delicate, since the non-fulfillment of equations such as (2.11) may be due to sampling errors. The question asked then is whether the equation is "significantly" unfulfilled. Further, redundant information (such as $\sigma_{12} = 0$ in our case) may be used, not to test the hypothesis, but to increase the precision of the estimate of the parameters.

21. Rather, "significantly" inconsistent, unless the sample is infinite. See previous footnote.


23. This characterization of $Z$ will simplify the exposition that follows presently, and also that of "policies" to be discussed in lecture III.


25. See footnote 26 below.
26. Suffice it to say that equation (3.1), stating that production (measured in dollars), \( y \), is adjusted instantaneously to demand (measured in dollars), \( b + c + \rho \), is merely an approximation of some equation that would state how this adjustment proceeds in time. For example, if one thinks that producers are stimulated by the excess of demand over supply one might try to replace (3.1) by the differential equation

\[
\frac{dy}{dt} = \alpha (b + c + \rho - y),
\]

where \( t \) is time and \( \alpha \), a positive constant, is a behavior parameter measuring the speed of the producers' reaction to a unit discrepancy between demand and production. Using (3.2), (3.3), the equation (3.7) becomes

\[
\frac{dy}{dt} = \alpha [y(\beta + \gamma - 1) + \rho - \gamma \tau + v_b + v_c];
\]

if \( v_b \) and \( v_c \) were constant, (3.8) would imply that \( y \) grows exponentially ("explodes") through time if \( \beta + \gamma - 1 > 0 \), and declines exponentially ("fades out") towards the "equilibrium value" given in (3.4), if \( \beta + \gamma - 1 < 0 \). Because of random fluctuations of \( v_b \) and \( v_c \), these trends will be distorted somewhat but it remains true that the inequality \( \beta + \gamma - 1 > 0 \) implies an "explosive" rise in income (measured in dollars) such as was historically observed only in times of a few exceptional hyper-inflations. At least when one has to decide between small upward or downward variations of \( \rho \) and \( \tau \), the behavior of consumers and businessmen can be assumed to be such as to preclude explosive hyper-inflation, and hence to preclude that \( \beta + \gamma > 1 \), i.e., to preclude \( \lambda \) negative.